Multimodal A2: Sensor Fusion with Attention

Rajeev Atla

1. Introduction

Human activity recognition underpins many wellness applications, from rehabilitation and fall detection to everyday fitness tracking. Inertial measurement units are a natural starting point for motion understanding, but they struggle during low-motion or stationary states such as sleep or quiet rest. Physiological signals like heart rate supply complementary context that can stabilize predictions in those edge cases. Fusing physiological and kinematic data can therefore lift robustness and overall accuracy, especially when sensors are noisy or degraded.

This report studies multimodal fusion on the PAMAP2 Physical Activity Monitoring dataset [1], [2], which aligns heart rate with IMU streams from the accelerometer, gyroscope, and magnetometer across 18 activities. The task is challenging for three reasons: different sampling rates between modalities, missing or unreliable measurements that include IMU drift and noisy heart rate, and temporal offsets such as delayed heart rate changes after a subject sits down. Fusion models also risk overfitting to the denser modality, which hurts validation performance and transfer when inputs go missing.

2. Approach

We compare three fusion strategies for activity recognition: early, late, and hybrid fusion.

Early fusion: Signals are temporally aligned and concatenated at the input level, then processed by a shared encoder. This captures joint correlations but can overfit to the most informative or highest-rate stream and absorb its noise.

Late fusion: Each modality is encoded by a separate transformer, and predictions are combined near the output. This preserves modularity and can improve calibration, but it may miss fine-grained cross-modal interactions that arise earlier in the pipeline.

Hybrid fusion: Separate encoders are linked with cross-attention so that features exchange information at multiple depths. Each encoder continues to learn modality-specific structure while cross-attention adjusts the relative influence of signals in context.

3. Results

3.1. Fusion Comparison

Figure 1 compares the three strategies on accuracy, F1, calibration, and latency. Early fusion achieves the highest accuracy at TODO and an F1 of TODO by tightly coupling heart rate and IMU features. In practice, it operates as an IMU-centric model because the motion stream dominates in sampling rate and dimensionality. This yields strong results when every sensor is present and synchronized.

Hybrid fusion delivers a different tradeoff. It gives up roughly one percentage point of accuracy while improving calibration and resilience to degraded inputs. Cross-attention regularizes the IMU representation and tempers overconfidence. Late fusion remains the most modular and often well-calibrated, but it captures fewer synergistic patterns and therefore trails in peak accuracy.

3.2. Missing Modality Robustness

TODO: figure studies performance as sensors are removed. Accuracy falls from TODO with all inputs to TODO with a single stream. Among single-modality baselines, IMU-hand reaches TODO and IMU-ankle reaches TODO, both ahead of IMU-chest at 0.539 and heart rate at TODO. These results confirm that kinematics carry most discriminative power and that heart rate alone is a weak activity cue.

3.3. Attention

TODO: figure averages cross-modal attention over test windows. The matrix shows strong self-attention for IMU-hand and heart rate, and notable couplings between IMU-hand and chest in both directions. Heart rate attends to IMU-hand as a global modulator, while motion features attend back to heart rate to refine interpretation during transitions. The pattern suggests a hierarchy: dense intra-IMU interactions capture coherence of movement, and heart rate shapes global context. This explains why hybrid fusion improves calibration and stability despite a slight dip in peak accuracy.

3.4. Uncertainty Calibration

TODO: figure reports calibration results. We observe that temperature scaling reduces expected calibration error and that hybrid and late fusion retain low ECE in noisy or low-motion segments, while early fusion tends to overpredict confidence in those same regions.

3.5. Ablation Studies

We vary the number of attention heads to study specialization versus stability. A single head captures broad relationships and converges smoothly, but it misses fine temporal or cross-modal structure. Eight heads increase representational power and improve accuracy by about 0.4%, but they also introduce noisy attention patterns that slightly harm calibration and raise complexity. Four heads strike the best balance, offering diverse yet stable interactions with efficient inference and dependable uncertainty estimates. The single-head variant converged faster but underperformed on fusion quality.

TODO: figure compares multimodal fusion to single-sensor and naive concatenation baselines. IMU-hand and IMU-ankle lead among individual streams, while IMU-chest and heart rate lag. The fusion model improves on any single stream by 25–30%. The left panel shows that performance grows almost linearly with the number of available sensors, reinforcing the value of properly aligned fusion.

4. Discussion

Hybrid fusion provides the most balanced profile across accuracy, calibration, and robustness. Early fusion wins on clean, full-sensor accuracy at TODO: ADD ACCURACY, which reflects a strong bias toward the IMU stream in settings with reliable motion cues. Hybrid fusion introduces cross-attention that adapts modality weightings to context and uncertainty. It does not raise peak accuracy in ideal conditions, but it holds confidence in check and sustains performance when inputs are noisy or incomplete. Late fusion remains attractive for modular systems and calibration, yet it learns fewer synergistic features and adapts less to partial failure.

Failure modes differ in shape. Early fusion relies on dense concatenation, so missing IMU inputs destabilize its shared embedding. Hybrid fusion mitigates this through correlated signals and learned redundancy, such as using motion statistics to anticipate heart rate changes. After temperature scaling, hybrid and late models maintain low expected calibration error below 0.03, while early fusion remains overconfident in low-motion segments. Attention maps support this story with clear bidirectional coupling between heart rate and motion streams and strong intra-IMU structure that encodes coherent movement.

5. Conclusion

Multimodal fusion improves more than accuracy. It builds sensing systems that behave reliably under noise and partial failure. Integrating physiological and kinematic data yields predictions that are sensitive to context and transparent about uncertainty, instead of overconfident outputs dominated by a single stream. Early fusion is a strong choice in stable settings where one modality reigns. Late fusion simplifies system integration and keeps calibration in good shape. Hybrid fusion adds cross-attention that preserves high accuracy while offering real-world robustness, making it the most practical default for deployment.

6. References

[1] A. Reiss and D. Stricker, “Introducing a New Benchmarked Dataset for Activity Monitoring,” in *2012 16th International Symposium on Wearable Computers*, 2012, pp. 108–109. doi: 10.1109/ISWC.2012.13.

[2] A. Reiss, “PAMAP2 Physical Activity Monitoring.” 2012.