Beyond Motion Primitives: Behavioral Activity Recognition from Head-Mounted IMU

Anonymous CVPR submission

Paper ID *****

Abstract

AR smart glasses need continuous behavioral context to offer proactive assistance, yet their most practical always-on sensor, the head-mounted Inertial Measurement Unit (IMU), detects only motion primitives such as walking or standing. We push beyond motion primitives to behavioral-level recognition, defining five categories that balance AR application need with sensor observability. To this end, we construct a 160K-sample Ego4D dataset with a four-tier quality assurance framework spanning 8 activity scenarios, and propose HiT-HAR, a 703K-parameter hierarchical model that outperforms prior head-mounted IMU models on five-class action and eight-class scenario recognition. We further map the observability frontier of head-mounted IMU through per-class separability analysis, identifying which behavioral categories are reliably observable (Locomotion), which benefit from temporal context (Object Transfer, Task Operation), and where scenario-dependent signal overlap poses remaining challenges. Our results indicate that architectural choices exploiting temporal context and scenario structure outperform simply scaling model size.

1. Introduction

AR assistance that goes beyond passive display requires the system to understand the user's behavioral context: not just whether they are moving, but what they are functionally doing. Head-mounted inertial measurement units (IMUs) offer a privacy-preserving, always-on window into user behavior, and as AR glasses move toward consumer deployment, the embedded IMU becomes the most practical sensor for continuously sensing behavioral context without the battery cost of always-on cameras.

Yet most IMU-based activity recognition targets high-momentum motion primitives such as walking, running, and sitting [3, 12, 15]. These categories are well-separated in accelerometer space but tell an AR assistant little about *what* the user is functionally doing. Knowing that a user is walking says nothing about whether they are searching for a tool or transferring materials; an AR system needs exactly this behavioral distinction to decide what assistance to offer.

Consider a user performing a mechanical repair. Over 30 seconds, they pick up a wrench (Object Transfer), tighten a bolt (Task Operation), pause to inspect their work (Stationary), then walk to get more parts (Locomotion). An AR assistant that only detects "walking" vs. "stationary" cannot distinguish these functional states.

We define five behavioral categories that balance what AR applications need against what the IMU can observe, then systematically map the observability boundaries of head-mounted IMU using Ego4D [6] data.

Our contributions are threefold.

1. **Annotated dataset with quality framework.** A 160K-sample annotated dataset from Ego4D head-mounted IMU spanning 8 activity scenarios, with 27K gold labels from 12 annotators and a four-tier quality framework (Sec. 3).
2. **HiT-HAR (Hierarchical Temporal Human Activity Recognition).** A 703K-parameter architecture combining multi-dilation CNN-GRU local encoding with Transformer temporal aggregation and a scenario-informed gated action head, outperforming prior head-mounted IMU architectures including IMU2CLIP [12] on five-class action F1 (Table 2).
3. **Observability frontier analysis.** A systematic mapping of which behavioral categories are reliably observable from head-mounted IMU, which benefit from temporal context, and where scenario-dependent signal overlap poses remaining challenges (Sec. 6).

2. Related Work

Activity taxonomy and the behavioral level. HAR research organizes movements into hierarchical granularity levels [2, 11, 13]. Motion primitives (walking, standing) are well-studied and reliably detectable from inertial sensors, while fine-grained manipulation (hammering, pouring) typically requires vision or multi-sensor fusion [16]. Multi-

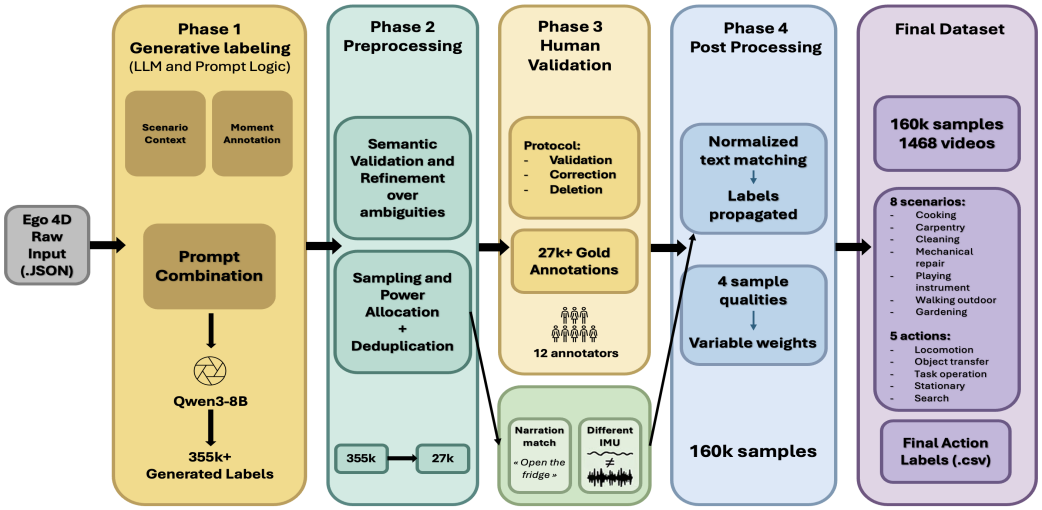


Figure 1. End-to-end data pipeline: LLM-generated labels are preprocessed, verified by 12 human annotators producing 27K gold annotations, then propagated to near-duplicate narrations yielding ∼160K labeled samples with four quality tiers.

level annotation schemes such as OPPORTUNITY++ [17] formalize this hierarchy for sensor datasets, defining postures, gestures, and high-level activities as separate label layers. Between these extremes lies a *behavioral level* that captures functional intent: not how the user moves, but what they are trying to accomplish. However, prior taxonomies are designed primarily around signal separability or annotation convenience rather than downstream application need. Designing taxonomies at this level requires balancing application relevance against sensor observability, a trade-off we formalize through dual-criteria taxonomy design (Sec. 3.1).

Head-mounted IMU recognition. EgoCHARM [15] introduces a hierarchical architecture pairing a per-window encoder with a sequence-level aggregator, classifying 3 motion primitives and 9 activity scenarios from a single egocentric IMU (∼85K params). IMU2CLIP [12] aligns IMU embeddings with CLIP for zero-shot classification and has become a standard baseline in head-mounted IMU recognition, adopted by EgoCHARM, PRIMUS [4], and COMODO [3]. COMODO distills video supervision into an IMU encoder on Ego4D [6]. MopFormer [19] applies a Transformer encoder to wearable-sensor motion primitives, showing that self-attention over temporal windows outperforms purely recurrent models, though it targets body-worn sensors and coarse motion categories rather than head-mounted behavioral recognition. Haresamudram *et al*. [7] showed that language supervision underperforms standard training for sensor HAR, motivating task-specific architectural choices over pre-training at scale. However, none of these targets the behavioral level — the functional intent behind motion patterns.

Architectural techniques for IMU HAR. Per-channel recalibration via Squeeze-and-Excitation (SE) attention [8] and multi-dilation convolutions [19] capture motion patterns at multiple time scales; our Window-Level Encoder adopts both. Gated Multimodal Networks [1] learn multiplicative gates to blend signals from different sources. We repurpose gated fusion for blending local per-window and contextual sequence-level representations within a single-sensor pipeline, guided by the observation that class separability varies by scenario.

Datasets and annotation for IMU HAR. Existing head-mounted IMU benchmarks derive labels from Ego4D scenario metadata [6] or motion-primitive ontologies [15, 17], but none provide fine-grained behavioral action labels from egocentric IMU. IMUGPT [9] demonstrated that LLMs can generate plausible action labels at scale, yet without human verification the resulting labels are noisy and inconsistent. We build on this paradigm with an LLM-human backfeed loop: Qwen3-8B generates initial labels over 355K narrations, then human annotators produce 27K gold annotations with a four-tier quality framework (Sec. 3.2).

3. Dataset and Quality Framework

We construct an annotated dataset from the Ego4D egocentric video corpus [6], pairing head-mounted 6-axis IMU recordings (accelerometer and gyroscope, 50 Hz) with behavioral-level action labels. The dataset spans 8 activity scenarios across 1,468 videos, totaling ∼160K labeled samples with a four-tier quality framework.

3.1. Five-Class Behavioral Taxonomy

Following the HAR hierarchy [2, 11], we target a *behavioral level* between motion primitives (walking, standing) and fine-grained manipulation (hammering, pouring). The taxonomy balances two criteria: *application relevance*, where each class corresponds to a distinct AR assistance action enabling a concrete system response, and *observability hypothesis*, where each class has a plausible head-motion signature distinguishing it from others in at least some activity scenarios.

We derive five categories by clustering Ego4D narration verbs according to semantic similarity, filtering for application relevance, and assessing the observability hypothesis against head-motion intuition (Table 1). Object Transfer (picking up, carrying) triggers spatial guidance and produces intermittent head turns toward targets. Task Operation (tightening, cutting) triggers step-by-step prompts and involves a steadier, task-focused gaze. Stationary (standing idle, observing) signals an opportunity for ambient information display and produces low energy across all channels. Locomotion (walking, climbing stairs) triggers navigation assistance and is characterized by periodic gait acceleration. Search (scanning a shelf, looking around) corresponds to a common trigger for proactive AR assistance. We hypothesize that it involves distinctive head rotation, yet some instances instead exhibit static gaze, producing signals with limited head movement. We include Search despite its limited observability because of its high application value and validate this observability boundaries in Sec. 6.

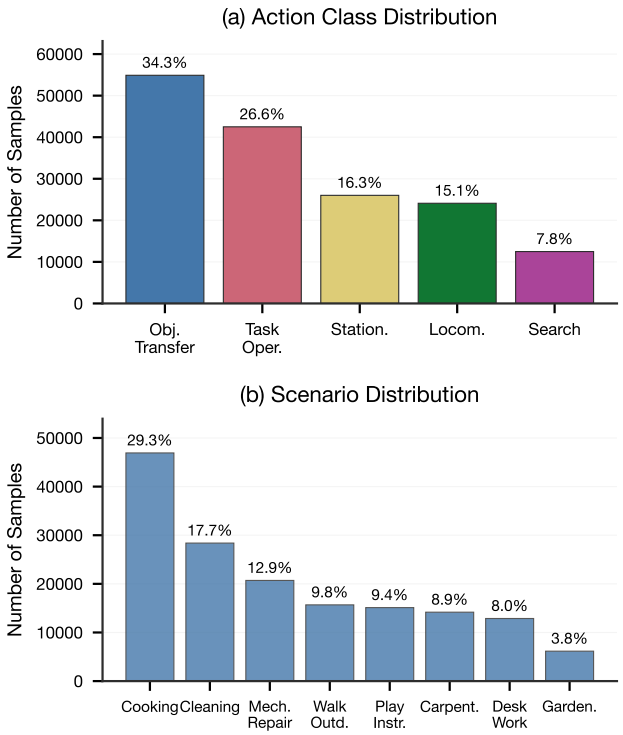


Figure 2. Dataset distribution over the full 160K-sample dataset. (a) Five action classes. (b) Eight Ego4D activity scenarios.

Table 1. Five-class behavioral taxonomy with per-class distribution (full 160K-sample dataset).

Class	Example narration	Count	%
Object Transfer	"picks up the wrench"	54,897	34.3
Task Operation	"tightens the bolt"	42,499	26.6
Stationary	"stands idle"	26,012	16.3
Locomotion	"walks to the shelf"	24,104	15.1
Search	"scans the shelf for a pan"	12,487	7.8

The dataset spans 8 Ego4D scenarios (Cooking, Carpentry, Cleaning, Desk Work, Mechanical Repair, Playing Instrument, Walking Indoors, Walking Outdoors). Fig. 3 illustrates how these behavioral classes map to head-mounted IMU signals in a representative scenario.

3.2. LLM-Human Backfeed Annotation

Our annotation follows an *LLM-human backfeed loop* [18]: Qwen3-8B, an 8-billion-parameter reasoning model, classifies 355K Ego4D narrations into five categories given the narration text, activity scenario, and taxonomy definitions, producing per-sample reasoning chains. Following the LLM-assisted paradigm also explored in IMUGPT [9], 12 human annotators then verify labels across two rounds using the narration, scenario context, and LLM reasoning as evaluation inputs, producing 27,355 gold annotations.

Sampling strategy: To oversample rare classes relative to proportional sampling, we use square-root frequency resampling (α=0.5), which increases the annotation budget for minority classes such as Search. Narration strings repeat across the corpus: many distinct (video, time) rows share the same text. To avoid redundant annotation, we deduplicate narrations that share both normalized text (lowercased, articles and trailing punctuation removed) and LLM-assigned label; narrations with the same text but different labels remain separate.

Verified label propagation: Verified labels are propagated to near duplicates (for their IMU diversity), to yield 160K total samples across 1,468 videos (5.8$\times$ expansion). The LLM achieves 92.7% agreement with gold labels.

Key finding: Of 470 multi-label conflicts, 85.9% stem from taxonomy boundary ambiguity rather than LLM errors. The dominant conflict pairs involve Object Transfer/Task Operation and Search/Stationary, consistent with

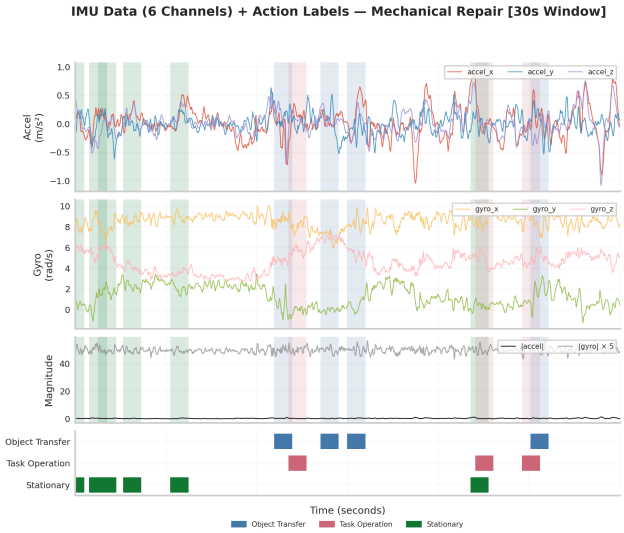


Figure 3. Head-mounted IMU signals (30-second window, Mechanical Repair) with 5-class action labels. Frequent transitions between Object Transfer, Task Operation, and Stationary illustrate behavioral diversity within a single scenario.

the observability hypothesis that these class pairs share similar head-motion signatures in certain scenarios.

3.3. Quality Tier System

Not all gold labels are equally reliable. Following work on learning under temporal label noise [14], we assign four quality tiers based on annotator behavior signals: Tier 1 (high confidence, 30.9%) requires a Gold verdict with no secondary choice and an unambiguous verb (weight 1.0); Tier 2 (moderate, 35.0%) allows a secondary choice or ambiguous verb (weight 0.8); Tier 3 (corrected, 9.6%) captures cases where annotators corrected the LLM label (weight 0.5–0.7); and Tier 4 (excluded, 24.5%) covers skipped or deleted samples (weight 0.0). These confidence weights modulate the focal loss [10] during training (Sec. 4.5).

Label sparsity: Action labels cover 17.4% of total video time after propagation, reflecting the inherent sparsity of narration-based annotation in Ego4D. This sparsity, combined with the narration-to-IMU semantic gap (visual narrations describe hand-level events like "picks up the cup" while the IMU captures only head motion), motivates a model that exploits temporal context and multi-task scenario supervision.

4. Method

The challenges above — overlapping class distributions in raw IMU space, sparse labels, and the narration-to-IMU semantic gap — motivate three design choices: multi-dilation convolutions to capture motion at multiple time scales, a Transformer-based sequence aggregator to exploit 30 seconds of context for resolving single-window ambiguity, and scenario-informed gating to leverage scenario-dependent separability.

4.1. Architecture Overview

HiT-HAR is a lightweight hierarchical model for joint scenario classification and behavioral-level action recognition from head-mounted IMU (Fig. 4). Following the hierarchical encoder-aggregator design introduced by EgoCHARM [15], a *Window-Level Encoder* (WLE) maps each 1-second IMU window to a 128-dimensional embedding. A *Window Aggregation Transformer* (WAT) aggregates 30 such embeddings, capturing 30 seconds of temporal context. Two prediction heads operate on the aggregated representation: a scenario head and a gated action head.

4.2. Window-Level Encoder

The WLE processes each 1-second window of 8-channel input (6-axis IMU plus acceleration and gyroscope norms) at 50 Hz. A stem convolution projects the input to 64 channels, followed by three multi-dilation CNN blocks with dilations $\{1, 2, 4\}$ and channel recalibration via Squeeze-and-Excitation attention [8]. A bidirectional GRU with attention pooling yields the 128-dimensional per-window embedding $\mathbf{e}_t$.

4.3. Window Aggregation Transformer

The WAT aggregates $S{=}30$ window embeddings into sequence-level and per-window contextual representations using a single-layer Transformer encoder with a learnable CLS token, learnable positional embeddings, 4 attention heads, and feed-forward dimension 512. The CLS output $\mathbf{h}_{\text{cls}}$ serves as the sequence-level embedding; per-window outputs $\mathbf{h}_t$ provide contextual representations spanning the full 30-second window. The 30-second context spans approximately six typical action transitions in Ego4D narrations.

4.4. Gated Action Head

Two heads produce predictions from the WAT outputs. The *scenario head* applies a linear classifier to $\mathbf{h}_{\text{cls}}$, predicting one of eight scenarios. The *gated action head* fuses local (WLE) and contextual (WAT) signals for per-window action prediction:

$$\mathbf{a}_{\text{loc}} = W_{\text{loc}}\,\mathbf{e}_t, \quad \mathbf{a}_{\text{ctx}} = W_{\text{ctx}}\,\mathbf{h}_t, \tag{1}$$

$$g = \sigma\big(\text{MLP}([\mathbf{e}_t;\, \mathbf{h}_t])\big), \tag{2}$$

$$\mathbf{a}_t = (1-g)\,\mathbf{a}_{\text{loc}} + g\,\mathbf{a}_{\text{ctx}}, \tag{3}$$

where $g \in [0, 1]$ is a learned scalar gate that adaptively blends local motion evidence with longer-range context.

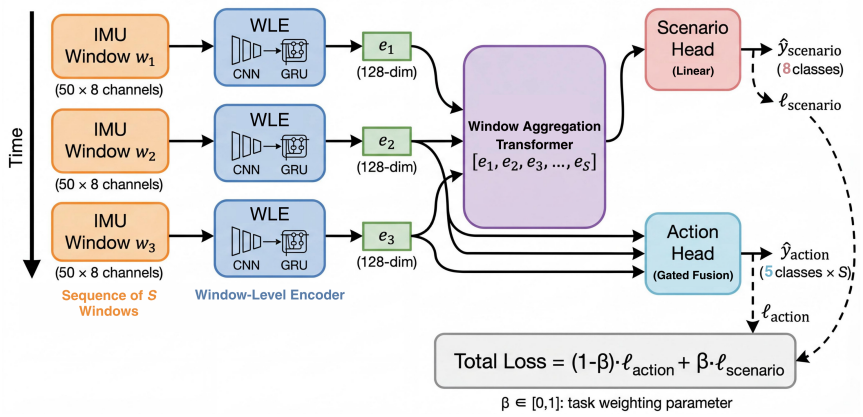


Figure 4. HiT-HAR architecture. The Window-Level Encoder (WLE) processes each 1-second IMU window independently via CNN and GRU, producing 128-dimensional embeddings. The Window Aggregation Transformer (WAT) aggregates a sequence of window embeddings. A scenario head predicts from the CLS token, while a gated action head fuses local and contextual predictions. Total: 703K parameters.

The scenario classification head supervises $\mathbf{h}_{\text{cls}}$, which participates in self-attention with all $\mathbf{h}_t$ tokens, encouraging $\mathbf{h}_t$ to encode scenario-discriminative structure that the gate can exploit.

4.5. Multi-Task Training

The total loss combines scenario and action objectives:

$$\mathcal{L} = \beta \cdot \mathcal{L}_{\text{scenario}} + (1 - \beta) \cdot \mathcal{L}_{\text{action}} \tag{4}$$

where $\beta = 0.3$ balances scenario and action objectives (Sec. 5.3). Both losses use focal loss [10] (γ=2.0) with inverse-frequency class weights. We train with AdamW, cosine scheduling, gradient clipping (max norm 1.0), EMA (decay 0.999), and label smoothing (ϵ=0.05).

5. Experiments

5.1. Experimental Setup

We train and evaluate on our Ego4D-derived dataset (Sec. 3), splitting 1,468 videos by UID into train (111K samples), validation (24K), and test (25K) partitions. Our primary metric is five-class macro-F1. We train for 40 epochs with AdamW (lr = 10^{-4}, weight decay 5×10^{-4}), batch size 128, cosine scheduling with 3-epoch warmup, and early stopping on validation action F1 (patience 15). IMU signals are globally z-score normalized using training-set statistics and per-video centered to reduce device drift. During training, we apply random augmentation: jittering (σ=0.02), uniform scaling (0.9–1.1), temporal masking, and small 3D rotation ($\pm15°$). We compare against four baseline architectures that span the design space explored in recent head-mounted IMU work [15]. IMU2CLIP [12] uses the stacked CNN-GRU encoder designed for CLIP-aligned IMU pretraining. CNN-LSTM-GRU uses LSTM-GRU sequence modeling without Transformer aggregation. CNN-MLP replaces the temporal encoder and aggregator with MLPs. MLP-MLP uses multi-layer perceptrons throughout. All baselines are trained on the same data with identical focal loss and class weights. We also evaluate Mantis [5], a general time-series foundation model (8M params), with frozen features + SVM and adapter-head fine-tuning on action classification only (no scenario head).

Table 2. Main results on 5-class behavioral-level action recognition. F1 is macro-averaged; Acc is micro-averaged. Mantis has no scenario head (—).

Model	Act. F1	Act. Acc	Scen. F1	Scen. Acc	Params
MLP-MLP	0.339	0.436	0.528	0.578	1.03M
Mantis-8M frozen [5]	0.301	0.321	—	—	8M
IMU2CLIP [12]	0.385	0.435	0.559	0.590	4.0M
Mantis-8M fine-tuned	0.370	0.467	—	—	8M
CNN-MLP	0.367	0.451	0.523	0.554	1.07M
CNN-LSTM-GRU	0.378	0.455	0.508	0.551	1.19M
HiT-HAR (ours)	**0.457**	**0.490**	**0.569**	**0.601**	**703K**

5.2. Main Results

Table 2 shows that HiT-HAR achieves the highest action F1 (0.457) and action accuracy (0.490) among all models, outperforming CNN-LSTM-GRU by 0.069 F1 and IMU2CLIP by 0.094 F1 while using 5.7× fewer parameters than IMU2CLIP. Mantis-8M, despite having 11× more parameters, reaches only 0.370 action F1 with adapter fine-tuning, suggesting that Mantis's general-purpose time-series pre-training transfers poorly to behavioral-level IMU recognition. The gap between macro-F1 and accuracy reflects class imbalance: accuracy is dominated by frequent classes, while F1 gives equal weight to each class includ-

ing the rare Search category. Per-class analysis reveals a clear observability hierarchy: Locomotion (F1 = 0.596) is reliably detected through gait periodicity, Object Transfer (0.519) and Task Operation (0.510) are partially separable with temporal context, Stationary (0.386) is moderate, and and Search (0.273), the most application-relevant class for proactive AR, marks the boundary where complementary sensors such as eye tracking would yield the greatest benefit. This per-class hierarchy directly mirrors the raw-IMU separability analysis in Sec. 6.

5.3. Ablation Study

Multi-task weighting (β sweep). We select β=0.3 as the balanced operating point: action F1 (0.457) is only 0.009 below the action-only optimum (β=0), while scenario F1 improves from 0.121 to 0.569. For AR assistance, scenario context has practical value in enabling context-aware guidance.

Architecture scaling. Expanding HiT-HAR to 1.1M parameters (3-layer WAT, wider hidden dimensions) yields no improvement in action F1 (0.457 in both configurations). Language-based label alignment (SBERT, CLIP, Qwen text encoders) also provides negligible gains (+0.004 at best), in line with Haresamudram *et al*. [7]. These results confirm that exploiting temporal context and scenario structure matters more than scaling model size.

6. Observability Frontier Analysis

Beyond model performance, we ask: *what can head-mounted IMU physically distinguish?* We map the observability frontier through two complementary analyses, characterizing the boundaries of what head-mounted sensing can resolve and where complementary modalities would be needed.

6.1. Per-Class-Pair IMU Separability

We measure pairwise class separability directly in the 8-dimensional raw IMU feature space (6-axis plus 2 norms) using two complementary metrics (Fig. 5). The *Bhattacharyya distance* quantifies distributional overlap under a Gaussian assumption (higher values indicate better separability), while the nonparametric *MMD two-sample test* assesses whether two class distributions are statistically distinguishable without distributional assumptions (permutation-based p-values, n=1,000, Bonferroni-corrected).

Both metrics reveal the same structure. Locomotion separates cleanly from all four other classes (Bhattacharyya 3.3–5.0, MMD^2 0.016–0.048, all $p < 0.05$), and Object Transfer separates from Search ($p = 0.020$), yielding 5 of 10 class pairs statistically distinguishable from head motion alone. The hardest pairs remain Object Transfer vs. Task Operation (MMD^2 $= 0.006$, $p = 0.94$) and Stationary vs. Search (MMD^2 $= 0.001$, $p = 1.00$), confirming that these class pairs have substantial overlap in per-window IMU features and that temporal aggregation, as in HiT-HAR, is needed to exploit sequential context.

However, separability varies by scenario: in scenarios with frequent object displacement (e.g., Cooking, Carpentry), the head-turning pattern during Object Transfer becomes more distinctive, while Task Operation involves a steadier gaze. Similarly, Search and Stationary overlap because most Search instances involve static visual scanning, but Search during outdoor walking produces more head rotation and is better separable. These scenario-dependent patterns motivate the multi-task design of HiT-HAR, where scenario context aids per-window action classification.

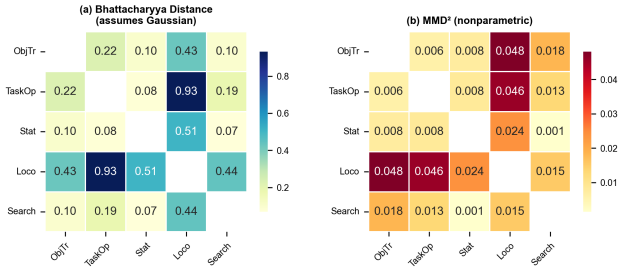


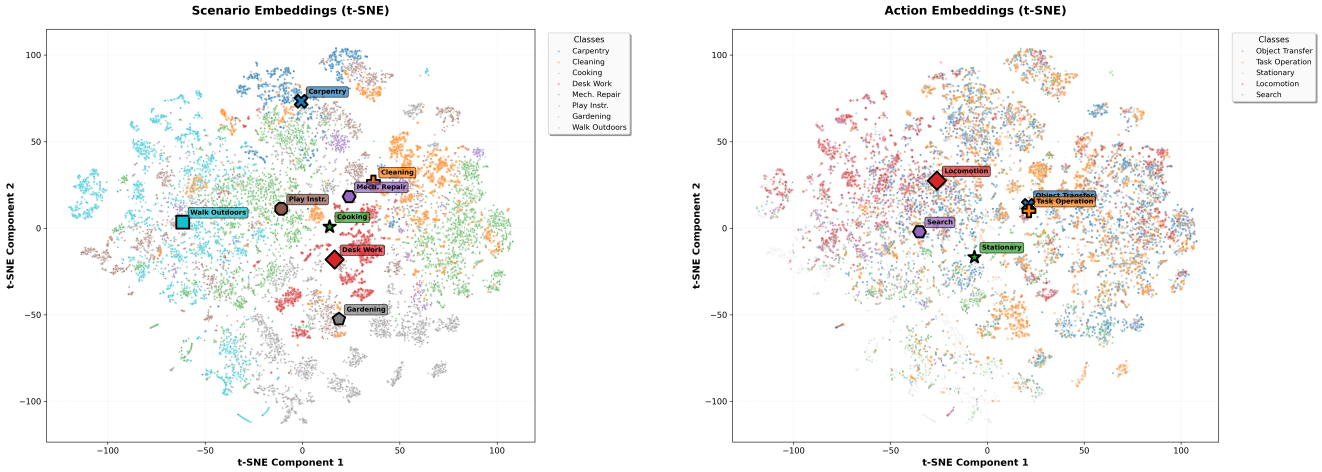
Figure 5. Pairwise class separability in 8-dim IMU feature space. (a) Bhattacharyya distance: Locomotion stands out (3.3–5.0). (b) MMD^2 (n=1,000 permutations, Bonferroni-corrected): five pairs are significant ($p < 0.05$).

6.2. Learned Embedding Structure

Fig. 6 visualizes the learned embeddings via t-SNE. Scenario embeddings form recognizable clusters (Walk Outdoors, Carpentry, Desk Work are well separated), explaining the strong scenario F1 of 0.569. Action embeddings are far more entangled: Object Transfer, Task Operation, Stationary, and Search overlap extensively, with only Locomotion showing partial separation, corroborating the MMD analysis above.

6.3. Temporal Transition Structure

Beyond per-window separability, behavioral states form structured transition sequences that vary by scenario. Fig. 7 shows the global transition matrix alongside Desk Work as a contrasting scenario. Object Transfer and Task Operation form a dominant off-diagonal cycle (0.24/0.43), reflecting the natural fetch-use-fetch workflow in manipulation-heavy scenarios. Desk Work shows a fundamentally different pattern: Stationary self-transition jumps from 0.33 to 0.55, and Locomotion transitions primarily to Stationary (0.37) rather than to manipulation classes. These scenario-dependent



(a) Scenario embeddings

(b) Action embeddings

Figure 6. t-SNE of HiT-HAR learned embeddings (β=0.3). (a) Scenario embeddings form distinct clusters. (b) Action embeddings show regions of class overlap.

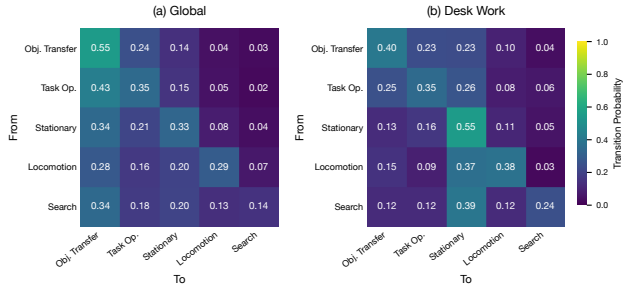


Figure 7. Behavioral state transition probabilities. (a) Global: Object Transfer and Task Operation form a dominant cycle. (b) Desk Work: Stationary dominates, reflecting a structurally different behavioral dynamic.

temporal patterns further justify multi-task scenario supervision in HiT-HAR.

6.4. Summary

The separability analysis, embedding structure, and transition patterns collectively delineate an observability frontier: Locomotion is fully observable through gait periodicity, Object Transfer and Task Operation become separable with temporal context and scenario supervision, and Search shows the strongest scenario dependence with better separability in outdoor active-scanning scenarios.

7. Discussion and Conclusion

We push head-mounted IMU beyond motion primitives by defining five behavioral categories that balance application need with sensor capability, then systematically mapping the observability boundaries of this modality. HiT-HAR, a 703K-parameter hierarchical model, outperforms established IMU-HAR baselines on action recognition (Table 2), while ablations confirm that model scaling provides negligible returns.

The observability frontier analysis confirms a clear hierarchy from reliably observable (Locomotion) through context-dependent (Object Transfer, Task Operation), with scenario-dependent signal overlap as the remaining challenge for classes like Search. Architectural choices that exploit temporal context and scenario structure prove more effective than simply scaling model capacity for behavioral-level recognition from head-mounted IMU.

Limitations. We use a single head-mounted IMU; wrist or body sensors could provide complementary signals. Action labels cover 17.4% of total video time; the remaining unlabeled time could support self-supervised pretraining. Our labels are derived from narration timestamps that mark when an action is mentioned rather than its full duration, so a quick manipulation and a sustained one both occupy the same 1-second window. We evaluate offline on recorded data; real-time inference latency and on-device deployment remain to be validated.

Future work. The observability gaps identified here point toward targeted sensor fusion — eye tracking to resolve Search from Stationary, wrist IMU for separating manipulation classes, and audio for disambiguating Task Operation subtypes. Self-supervised pretraining on the unlabeled video time, alongside systematic exploration of temporal window lengths and alternative architectures, can probe the current frontier further. The structured transition patterns in

Sec. 6.3 also suggest that lightweight next-state forecasting models could enable proactive AR assistance, leveraging the non-uniform, scenario-dependent transitions as a training signal. Because our annotations are paired with Ego4D video, the same behavioral taxonomy can serve as supervision for cross-modal tasks.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. Gated multimodal networks. *Neural Computing and Applications*, 32:10209–10228, 2020. 2

[2] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3):33:1–33:33, 2014. 1, 3

[3] Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora D. Salim. COMODO: Cross-modal video-to-IMU distillation for efficient egocentric human activity recognition. 2025. 1, 2

[4] Arnav M. Das, Chi Ian Tang, Fahim Kawsar, and Mohammad Malekzadeh. PRIMUS: Pretraining IMU encoders with multimodal self-supervision. In *ICASSP*, 2025. 2

[5] Vasilii Feofanov, Songkang Wen, Marius Alonso, Romain Ilbert, Hongbo Guo, Malik Tiomoko, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. Mantis: Lightweight calibrated foundation model for user-friendly time series classification. 2025. 5

[6] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 1, 2

[7] Harish Haresamudram, Apoorva Beedu, Mashfiqui Rabbi, Sankalita Saha, Irfan Essa, and Thomas Ploetz. Limitations in employing natural language supervision for sensor-based human activity recognition — and ways to overcome them. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 273–281, 2025. 2, 6

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 2, 4

[9] Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Ploetz. IMUGPT 2.0: Language-based cross modality transfer for sensor-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3), 2024. 2, 3

[10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017. 4, 5

[11] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006. 1, 3

[12] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. IMU2CLIP: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13246–13253, 2023. 1, 2, 5

[13] Md Golam Morshed, Tangina Sultana, Aftab Alam, and Young-Koo Lee. Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors*, 23(4): 2182, 2023. 1

[14] Sujay Nagaraj, Walter Gerych, Sana Tonekaboni, Anna Goldenberg, Berk Ustun, and Thomas Hartvigsen. TENOR: Learning under temporal label noise. In *ICLR*, 2025. 4

[15] Akhil Padmanabha, Saravanan Govindarajan, Hwanmun Kim, Sergio Ortiz, Rahul Rajan, Doruk Senkal, and Sneha Kadetotad. EgoCHARM: Resource-efficient hierarchical activity recognition using an egocentric IMU sensor. 2025. 1, 2, 4, 5

[16] Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54 (3):2259–2322, 2021. 1

[17] Daniel Roggen, Kilian Forster, Anna Calatroni, and Gerhard Tröster. Opportunity++: A multimodal dataset for video- and wearable, object and ambient sensors-based human activity recognition. *Frontiers in Computer Science*, 3:792065, 2022. 2

[18] Shao-Heng Wang, Hao-Cheng Lo, Ji-Lung Hsieh, and Shih-Chieh Dai. Human-LLM collaborative annotation through effective verification of LLM labels. In *CHI*, 2024. 3

[19] Hao Zhang, Zhan Zhuang, Xuehao Wang, Xiaodong Yang, and Yu Zhang. MoPFormer: Motion-primitive transformer for wearable-sensor activity recognition. In *NeurIPS*, 2025. 2

Appendix

A. Additional IMU Signal Visualizations

Fig. 8 and Fig. 9 show 30-second head-mounted IMU windows from two additional Ego4D scenarios, complementing the Mechanical Repair example in the main paper (Fig. 3).

Figure 8. Head-mounted IMU signals (30-second window, Cooking) with 5-class action labels. Frequent Object Transfer and Stationary transitions reflect the pick-prepare-pause workflow.

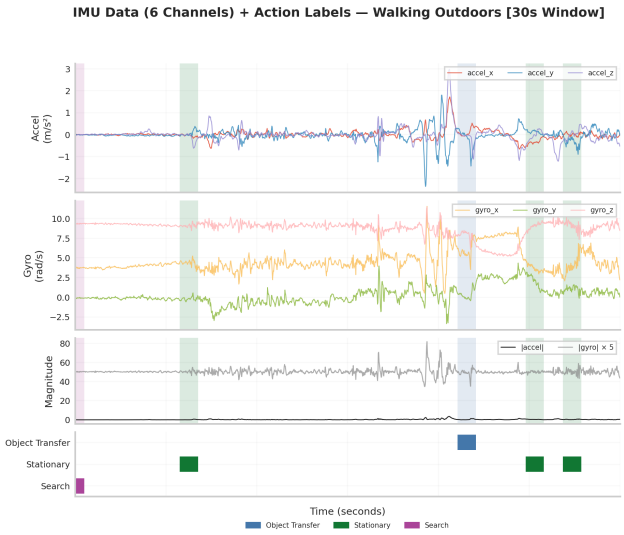


Figure 9. Head-mounted IMU signals (30-second window, Walking Outdoors) with 5-class action labels. Locomotion produces distinctive periodic gait patterns in the accelerometer channels.

B. Data Quality Analysis

Fig. 10 details the annotation quality framework from Sec. 3.3, showing per-class tier distributions, label coverage before and after propagation, and propagation consistency. Fig. 11 breaks down the 470 multi-label conflicts by source, confirming that the dominant challenge is taxonomy boundary ambiguity rather than LLM errors.

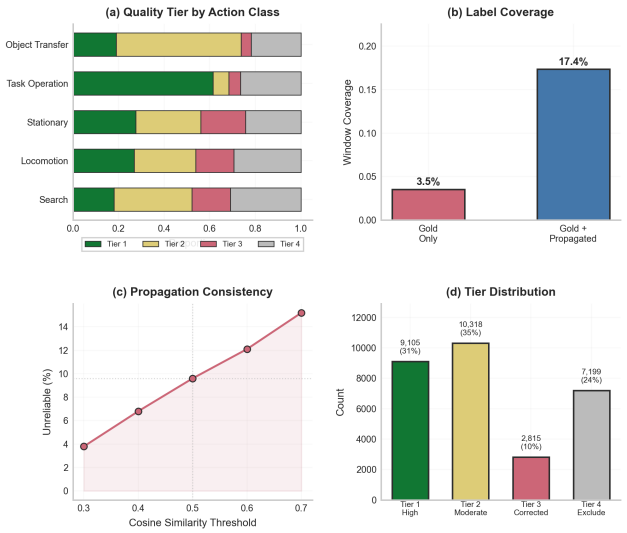


Figure 10. Annotation quality analysis. (a) Quality tier distribution by action class. (b) Label coverage: gold-only (3.5%) vs. gold + propagated (17.4%). (c) Propagation consistency vs. cosine similarity threshold. (d) Overall tier distribution (27K gold annotations).

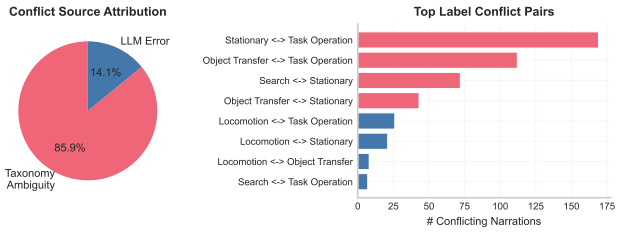


Figure 11. Label conflict source attribution. Left: 85.9% of multi-label conflicts arise from taxonomy boundary ambiguity, not LLM errors. Right: top conflict pairs ranked by frequency.

C. Per-Window Feature Ceiling

Fig. 12 compares the per-window feature ceiling (KNN-5 with GroupKFold on 42 statistical IMU features) against HiT-HAR across three taxonomy granularities. The deep model exceeds the per-window ceiling in all configurations,

confirming that temporal aggregation captures patterns beyond what single-window statistical features can represent.

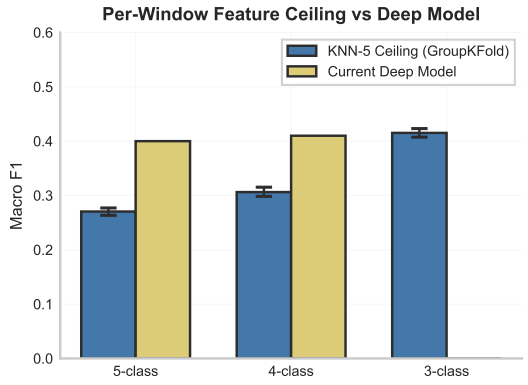


Figure 12. Per-window feature ceiling (KNN-5, GroupKFold) vs. deep model macro F1 across 5-class, 4-class, and 3-class taxonomies. The deep model exceeds the per-window ceiling, demonstrating the value of temporal aggregation.

D. Task-Weighting Sensitivity Analysis

Fig. 13 shows the full β sweep across five values. At β=0, the model is action-only and scenario F1 is near chance (0.12); introducing even moderate scenario supervision (β=0.3) raises scenario F1 to 0.57 while preserving action performance. Action F1 remains stable for $\beta \in [0.0, 0.7]$ but collapses at β=1.0, where the loss is scenario-dominated and the action head receives no direct gradient signal. This confirms that the optimal operating point lies at low β values where the scenario auxiliary task regularizes without degrading the primary action objective.

Fig. 14 reframes the same sweep as a Pareto frontier in the action–scenario F1 plane. HiT-HAR at $\beta \in \{0.3, 0.5, 0.7\}$ Pareto-dominates IMU2CLIP on both axes, demonstrating that hierarchical multi-task learning achieves a strictly better trade-off than the strongest single-task baseline.

E. Model Efficiency

Fig. 15 plots combined F1 (mean of action and scenario macro F1) against model size for all evaluated architectures. HiT-HAR (β=0.3) achieves the highest combined F1 (0.51), while IMU2CLIP requires 4× more parameters to reach a lower combined score (0.47). The lightweight baselines (MLP-MLP, CNN-MLP, CNN-LSTM-GRU) cluster at ∼1M parameters but trail by 6–8 points in combined F1, confirming that HiT-HAR's gains stem from architectural design rather than parameter scaling.

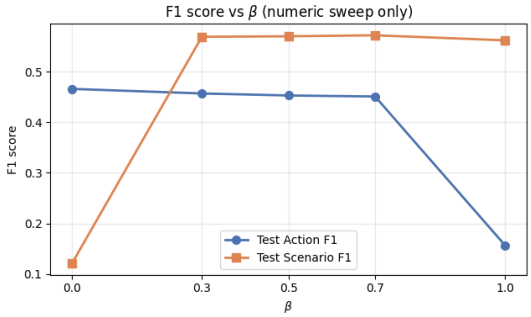


Figure 13. Test macro F1 vs. task-weighting coefficient β. Scenario F1 saturates at $\beta \geq 0.3$; action F1 remains stable until β=1.0, where loss is entirely scenario-driven and action performance collapses.

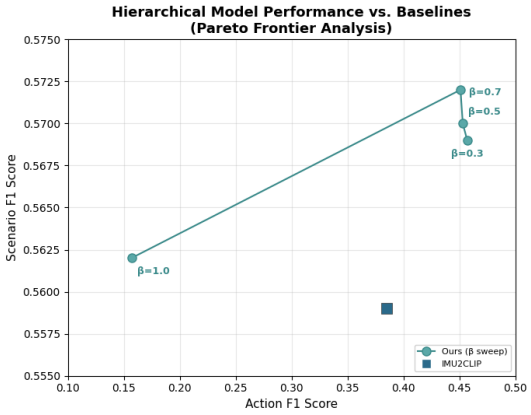


Figure 14. Pareto frontier in the action–scenario F1 plane. HiT-HAR ($\beta \in \{0.3, 0.5, 0.7\}$) strictly dominates IMU2CLIP on both tasks simultaneously.

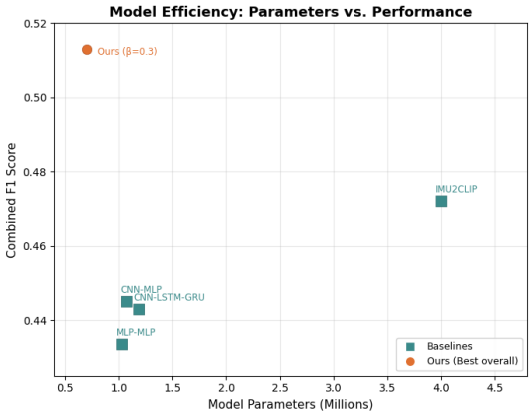


Figure 15. Model efficiency: combined F1 vs. parameter count (millions). HiT-HAR achieves the best performance at the smallest model size.