Sensor Privacy as a Spectrum: Quantifying Privacy in Edge and Multimodal Systems through Games

Jainta Paul
University of Utah
u1471999@utah.edu

Miles Bovero
University of Utah
u1048888@utah.edu

Swapnil Saha
STMicroelectronics Inc.
swapnilsayan.saha@st.com

Mahesh Chowdhary
STMicroelectronics Inc.
mahesh.chowdhary@st.com

Pratik Soni
University of Utah
psoni@cs.utah.edu

Luis Garcia
University of Utah
la.garcia@utah.edu

Abstract

Edge intelligence is often assumed to improve privacy because raw sensor streams remain local and only constrained outputs (e.g., binary events, compressed embeddings, or coarse labels) are exposed. We show this assumption is misleading: even minimal on-device outputs can retain structured semantics that enable adversaries to infer sensitive behaviors under realistic context and access regimes. We introduce a game-based framework that separates two sources of leakage: *statistical leakage*, bounded by the information capacity of the observable channel, and *algorithmic leakage*, unlocked when adversaries exploit temporal coherence, multi-channel structure, or auxiliary context within that bound. Across embedded, smartphone, and multimodal sensing datasets, we find that a quantized motion interface can preserve 70–75% of behavioral structure (via information- and divergence-based scores) and enable ∼65% activity inference accuracy (about 4× random guessing) once temporal continuity is restored. Exposing richer continuous channels further increases in-domain leakage but becomes strongly placement- and device-specific, degrading transfer across sensors and datasets. Finally, we show that representation learning can induce cross-modal bridges that erode sensing-layer constraints, making seemingly low-sensitivity IMU signals more predictive of private semantic attributes associated with hidden high-fidelity modalities. Together, these results show that privacy in edge AI is shaped by an interaction between physical constraints and observable-interface design, motivating co-designed sensing, model, and evaluation choices that quantify and limit interface-induced leakage rather than assuming locality implies privacy.

Keywords

edge sensing, privacy leakage, multimodal inference, information flow

1 Introduction

The rapid deployment of on-device and edge machine learning across wearables, smart homes, and industrial IoT has shifted sensing pipelines from cloud upload to local inference, fueling an intuition that "keeping data local" preserves privacy [10, 34, 37]. In practice, these systems rarely export raw sensor streams; instead they expose constrained interfaces such as binary events, quantized embeddings, or compressed features. Because such interfaces appear to reveal only low-dimensional, task-specific summaries rather than full-fidelity measurements, they are often marketed (and widely perceived) as privacy-preserving by design.

We argue that this assumption is misleading. Human (*un-*) interpretability is a poor proxy for privacy: outputs that are opaque to people can still be highly informative to a model equipped with population priors, temporal continuity, and auxiliary data. Popular protections such as federated learning, trusted execution environments, and differentially private training primarily address training-time or input-handling risks, but they do not directly quantify inference-time leakage through the *exported interface* [8, 33, 53]. A system can therefore "keep data local" while still enabling off-device adversaries to infer sensitive attributes (activities, identity, environment) from constrained outputs and context. This gap also surfaces in emerging IoT "privacy/security labels" framed as nutrition labels [11, 12]: they communicate static safeguards and coarse device properties, but not *interface-conditioned leakage*—how much an adversary can still infer given what the device exports and what context is observable. Instead, privacy claims should be tied to (i) the *access tier* (output granularity and sensor visibility) and (ii) the *available context* (temporal continuity and cross-modal priors), so privacy is reported as measured leakage under explicit assumptions rather than "local = private." Importantly, our claim is not that edge processing is inherently privacy-violating. Rather, locality alone is insufficient as a privacy argument unless the exported interface and the adversary's observable context are also considered. We therefore study privacy as an interface-conditioned property: what can be inferred depends on what the system releases, how much temporal or semantic structure is preserved, and what auxiliary context is available.

We introduce a *game-based* privacy framework that separates what an observable interface can reveal in principle from what realistic models can extract in practice. Specifically, we distinguish *statistical leakage*, determined by sensing physics, placement, and quantization (i.e., the information capacity of the observable channel), from *algorithmic leakage*, which arises when adversaries exploit residual structure using learning and auxiliary knowledge. Concretely, we define privacy games that vary both access (from on-device discrete outputs to single-modality and multi-modality sensor visibility, and onward to representation-level cross-modal alignment) and context (from shuffled, context-free windows to

Proceedings on Privacy Enhancing Technologies 2026(4), 341–357

https://doi.org/10.56553/popets-2026-0124

contiguous sessions and auxiliary priors). We quantify leakage with task metrics (Accuracy/Macro-F1), uncertainty-reduction metrics (MI/NMI), guessing-risk metrics based on Bayes vulnerability, and distributional metrics (NRKL), and relate these to capacity-style ceilings derived from entropy and temporal/statistical structure. Operationally, we ask: *given what the device exports and what context an adversary has, how much better than chance can the adversary infer a private target?* We formalize this as an adversarial advantage $Adv(A)$ in §3.

Across four IMU motion datasets (UTAH-STM-HAR, MotionSense [23, 24], UCI-HAR [32], and UTD-MHAD [5]), we observe a consistent gradient. Coarse binary interfaces impose a statistical ceiling, with 7-class activity inference near 30–35% top-1 in context-free settings and consistent with MI/Fano-style limits. Restoring temporal continuity then enables substantial algorithmic leakage even without raw private modalities. Exposing richer continuous channels (accelerometer and/or gyroscope) further increases within-dataset inference, while transfer experiments reveal when leakage is placement- or device-specific. Finally, we connect sensing-layer leakage to representation-layer threats by interpreting contrastive "align-then-transfer" pipelines as adversarial capabilities. We show that cross-modal linking can make low-sensitivity IMU representations more predictive of private semantic attributes associated with hidden high-fidelity modalities, even without reconstructing those hidden sensor streams.

Contributions. Our contributions are as follows:

- **A capacity-vs.-exploitation lens for inference privacy.** We model leakage as the combination of (i) *statistical capacity* induced by sensing physics, placement, and quantization, and (ii) *algorithmic exploitation* enabled by temporal continuity, auxiliary priors, and cross-signal structure. The framework yields interface-conditioned privacy games with comparable metrics: Accuracy/Macro-F1, MI/NMI, Bayes vulnerability, and NRKL.
- **A multi-dataset measurement of privacy under realistic interfaces.** Across four motion datasets, we show that binary event streams impose an information-theoretic ceiling, with 7-class activity inference near 30–35% top-1 in context-free settings. Restoring *temporal continuity* alone lifts inference to ∼55–60%, while exposing *raw sensor channels* (accelerometer and/or gyroscope) raises recovery to ∼70–99% in-domain. These gains also vary by *placement* (e.g., wrist vs. thigh), reflecting underlying kinematics and distinguishability. These measurements are grounded in motion-sensing datasets and are intended to demonstrate interface-conditioned leakage under controlled access regimes, rather than a universal claim about all edge systems.
- **Actionable guidance for "privacy by interface."** We translate these findings into reporting and design practices: capacity-aware summaries that report uncertainty reduction, one-shot guessing risk, and task performance together, and concrete interface levers such as quantization level, sampling rate, channel breadth, temporal continuity, and placement.

2 Preliminaries

This section establishes the signal-level conventions and metric definitions used in our privacy games. We first position our interface-conditioned view relative to motion-signal privacy, inference attacks under fixed observable channels, and quantitative information-flow metrics. We then define sensor streams, latent physical phenomena, temporal/cross-sensor structure, and the leakage metrics used in §4–§5. The adversary model and privacy games are formalized in §3.

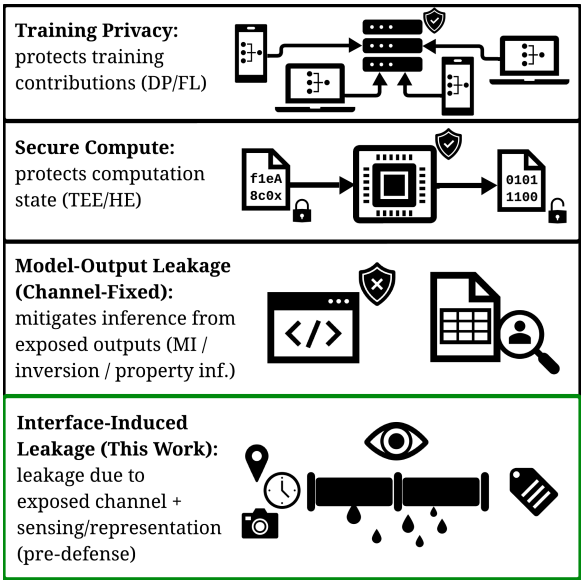

Figure 1: Positioning of privacy protections in edge sensing pipelines. Most prior defenses assume a fixed released interface (training privacy, secure compute, or output perturbation). We study *interface-induced leakage*: what is revealed by the exported channel and the adversary's observable context.

2.1 Motivation and Positioning

Motion privacy and cross-modal inference. Motion data are increasingly recognized as privacy-sensitive. Prior work shows that accelerometer streams can reveal activities, location, identity, health and body features, and typed input [18]. Recent attacks further show that rhythm features persist even at very low sampling rates [52]. Richer motion representations create similar risks: VR motion traces can support user identification [48], and skeleton motion can leak personally identifiable information through linkage attacks [4]. Complementary multimodal learning work shows that IMU streams can be embedded into shared spaces for retrieval and inference [13, 21, 26], while IMU foundation models highlight strong temporal structure within motion signals [14, 49, 50]. Together, these works establish that motion-derived signals leak sensitive structure. Our focus is how that leakage changes as the exported interface is coarsened, enriched, temporally preserved, or aligned across modalities.

Inference attacks and privacy-game formulations. Leakage can persist even when raw data are not directly exposed. Feature

sharing in heterogeneous and federated settings can create implicit leakage channels [6, 25, 45]. ML privacy work likewise shows that inference remains possible under black-box or label-only access [9, 17, 33, 44, 46]. Prior privacy-game work also shows how adversarial background knowledge and temporal correlation can reduce uncertainty even when released data are obfuscated [36]. We build on this game-based view of inference privacy, but instantiate the game around sensing interfaces rather than assuming a fixed released dataset, trace, or model output.

Quantitative information flow. QIF provides a useful lens for interpreting leakage as either uncertainty reduction or improved guessing success [38]. We use this perspective operationally: NMI captures aggregate reduction in target uncertainty, while Bayes vulnerability captures one-shot guessing risk. We define these metrics in §2.5.

Gap: defenses assume the interface; we quantify the interface. Existing defenses are important, but they often take the sensing configuration, released output, and adversary context as fixed [15, 19, 28, 31, 35, 43, 47, 51]. Our contribution is complementary. Rather than proposing another defense or attack on a fixed channel, we characterize *interface-induced leakage* by varying what is exposed, what temporal or auxiliary context remains available, and whether representation alignment is present.

2.2 Signals, Windows, and On-Device Outputs

A sensor modality $s \in \mathcal{S}$ produces a discrete-time stream $x_s(t) \in \mathbb{R}^{d_s}$ sampled at rate f_s. We write a window of length W starting at t as $x_s[t : t+W]$. When sensor placement matters, we index by location $\ell \in \mathcal{L}$ and write $x_{s,\ell}(t)$.

We view all observed streams as being driven by a shared latent physical phenomenon $\phi(t)$ (e.g., walking, turning), which induces modality- and placement-dependent projections:

$$x_{s,\ell}(t) \;=\; \mathcal{H}_{s,\ell}\big(\phi(t)\big) \;+\; \eta_s(t),$$

where $\mathcal{H}_{s,\ell}$ captures sensing/placement transfer and η_s denotes noise [1, 27]. We use this formulation only to emphasize that different modalities may be statistically linked through the same underlying dynamics.

On-device outputs. Many edge pipelines export only a constrained interface (e.g., a discrete event stream) rather than raw signals. We abstract an on-device detector as emitting a binary output

$$y(t) \triangleq \mathbf{1}\big\{g\big(x_s[t : t+W]\big) > \tau\big\},$$

where $g(\cdot)$ is a windowed statistic and τ a threshold, covering common embedded rule engines and lightweight classifiers [29, 42]. This abstraction aligns with our privacy games in §3, where the adversary's observable channel may include only $y(t)$.

2.3 Coupling and Temporal Structure

Leakage in multimodal sensing is amplified by two broad sources of structure: (i) *cross-sensor coupling* (shared dependence induced by $\phi(t)$) and (ii) *temporal coherence* (dependence across time).

Cross-sensor coupling. Coupling arises when multiple sensors capture the same underlying dynamics (e.g., accelerometer and gyroscope during gait) or when one sensing channel carries consequences of another (e.g., motion → audio). When we need a measurable notion of coupling, we use standard correlation and coherence tools: lagged cross-correlation $R_{ij}(\tau)$ and magnitude-squared coherence

$$\gamma_{ij}^2(f) \;=\; \frac{|S_{ij}(f)|^2}{S_{ii}(f)\,S_{jj}(f)},$$

where $S_{ij}(f)$ denotes the cross power spectral density (PSD) and $S_{ii}(f)$ the auto PSD [29, 42].

Temporal coherence (What Sequences Encode). Even when the exported interface is highly compressed (e.g., binary output labels), sequences can retain strong structure. For continuous signals, temporal dependence is summarized by autocorrelation $R_{ii}(\tau)$ and the PSD $S_{ii}(f) = \mathcal{F}\{R_{ii}(\tau)\}$ (Wiener–Khinchin). For discrete on-device outputs, we characterize temporal structure using simple sequence statistics such as (i) run-length distributions and (ii) transition rates/edge rates. Intuitively, periodic activities (walk/jog) induce repeatable rhythms, while static activities (sit/stand/lie) yield long dwells and sparse transitions. These are exactly the properties that become exploitable once the adversary observes contiguous sessions rather than shuffled windows.

Operationalizing coherence. In our experiments, we separate *context-free* settings (windows shuffled to break long-range dependence) from *context-aware* settings (contiguous sessions that preserve temporal coherence). This separation helps distinguish gains associated with temporal structure from those due only to marginal signal statistics.

2.4 Statistical Ceilings from Information

Let X denote an observable channel available to the adversary (e.g., a binary output, a single IMU modality, or fused IMU), and let Y denote the target label with C classes. The achievable prediction performance is constrained by the information that X carries about Y. We therefore report mutual information (MI) and normalized mutual information (NMI = $I(X;Y)/H(Y)$) as classifier-agnostic indicators of how much label uncertainty is revealed by the observable channel [3, 7, 16]. In §5.1, we use these quantities to separate *capacity* (what the interface can reveal in principle) from *exploitation* (how much an adversary can extract under different context/access regimes). Because entropy-based quantities do not fully capture one-shot guessing risk, we complement MI/NMI with Bayes vulnerability in the metric suite below.

2.5 Metrics Reported in Our Privacy Games

We quantify leakage using complementary metric families that appear throughout §4 and §5.

Task metrics. For classification targets, we report Accuracy and Macro-F1, which are standard in activity recognition [2, 20].

Information metrics. To compare leakage across datasets with different class counts and imbalance, we report MI and *normalized* MI (NMI), which scales MI by label entropy $H(Y)$ [7]. We also report NRKL, a normalized reverse-KL score that maps higher values to better alignment between predicted and true label distributions, following standard divergence-based privacy–utility analyses [3, 22].

Guessing-risk metrics. To complement NMI, we report Bayes vulnerability as a guessing-risk metric. For target Y and observable interface O, vulnerability measures the adversary's expected

Symbol	Meaning
$\phi(t)$	Latent physical phenomenon at time t (e.g., activity/gesture).
$\mathcal{S}, s$	Set of sensor modalities; a modality (e.g., accel, gyro).
$\mathcal{L}, \ell$	Set of placement locations; a location (e.g., wrist, pocket).
C	Environment-dependent constants (e.g., device/setting constants).
Π	Prior over latent phenomena/activities.
$\mathcal{E}$	Environment tuple: $\mathcal{E} = (\mathcal{S}, \mathcal{L}, C, \Pi)$.
$x_{s,\ell}(t)$	Sensor stream from modality s at location ℓ.
f_s	Sampling rate of modality s.
W	Window length.
$x_s[t : t+W]$	Windowed segment of stream $x_s(\cdot)$.
$\mathbf{s}(t)$	Multimodal sensor observation at time t (stack of all modalities/placements).
$\mathcal{G}_{\mathcal{E}}(\cdot)$	Generative process mapping $\phi(t)$ to $\mathbf{s}(t)$ under $\mathcal{E}$.
$\mathbf{x}(t)$	Full sensor state at time t (all modalities).
$\mathbf{x}_c(t)$	Compromised/adversary-visible sensor state at time t.
$\mathbf{x}_p(t)$	Private (hidden) sensor state at time t.
$\mathcal{S}_c$	Set of compromised modalities; $\mathcal{S}_p = \mathcal{S} \setminus \mathcal{S}_c$.
$y(t)$	Exposed on-device output released at inference time (discrete or continuous).
M	On-device model producing $y(t)$ from sensor inputs.
$\mathcal{S}_m$	Subset of modalities consumed by M (model input set).
$\mathrm{tgt}(\cdot)$	Private target function to be inferred by the adversary.
A, $\mathcal{A}$	Adversary; adversary capability/model class.
Aux	Auxiliary knowledge available to the adversary.
$Perf(\cdot)$	Leakage/performance metric used in the privacy game (Acc/F1/NMI/NRKL/etc.).
$BaseAdv$	Baseline performance under $Perf(\cdot)$ (e.g., random or majority-class).
$Adv(A)$	Adversarial advantage: $Adv(A) \triangleq Perf(A) - BaseAdv$.
ε	Privacy tolerance threshold; tgt-privacy holds if $Adv(A) \le \varepsilon$.
X	Generic observable channel used for analysis.
Y	Target label/attribute with C classes.
O	Observable interface/channel available to the adversary.
$I(X;Y)$	Mutual information between observable X and target Y.
NMI	Normalized mutual information: $I(X;Y)/H(Y)$.
$V(Y)$	Prior Bayes vulnerability: best one-shot guessing probability before observing O.
$V(Y \mid O)$	Posterior Bayes vulnerability: expected best one-shot guessing probability after observing O.
Vul	Reported Bayes vulnerability / one-shot guessing risk.
NRKL	Normalized reverse-KL score (higher is better alignment).
ρ	Duty cycle of a binary stream (fraction of time $y(t) = 1$).
λ	Edge/transition rate of a binary stream (transitions per unit time).

Table 1: Notation used throughout the preliminaries and privacy framework.

confidence in its best one-shot guess after observing O[38]. Thus, while NMI captures aggregate uncertainty reduction, vulnerability captures how easily the adversary can make a confident single prediction.

Simple temporal descriptors (binary outputs). For discrete interfaces (e.g., binary output labels), we additionally summarize sequence structure with duty cycle and edge rates. These descriptors capture how often the output is active and how often it changes [29, 42].

Interpreting leakage values. These metrics are complementary rather than interchangeable. Accuracy and Macro-F1 measure practical inference success; NMI measures aggregate uncertainty reduction; Vulnerability measures one-shot guessing risk; and NRKL measures distributional alignment. Thus, near-random task performance indicates limited exploitable leakage, while high Accuracy/F1 or high Vulnerability indicates that the interface supports reliable or confident inference.

3 Formal Privacy Framework

We now define an empirical privacy game for measuring interface-conditioned leakage. Figure 2 summarizes the sensing-to-inference pipeline. A latent phenomenon induces multimodal sensor streams, an on-device model consumes a subset of those streams, and the adversary observes only compromised streams, exposed outputs, and auxiliary context. Our framework specifies (i) the observable interface, (ii) the target attribute, (iii) the adversary's auxiliary knowledge and capability class, and (iv) the metric used to measure leakage. We use this framework as an empirical evaluation protocol, not as a universal privacy guarantee; measured risk is always relative to the specified interface, context, target, and adversary class.

3.1 System and Environment

We model the physical world as generating a latent phenomenon $\phi(t)$ (e.g., walking, sitting, turning) drawn from a prior distribution Π. This latent physical state induces multimodal sensor observations through a generative physical process

$$\mathbf{s}(t) = \mathcal{G}_{\mathcal{E}}(\phi(t)),$$

where the environment $\mathcal{E}$ specifies the sensing configuration. We use $\mathcal{X}$ to denote the space of realized sensor states, and write $\mathbf{x}(t) \in \mathcal{X}$ for the full sensor state available in the privacy game. Figure 2 summarizes this sensing-to-inference pipeline and highlights attacker-accessible components.

We define the environment as

$$\mathcal{E} = (\mathcal{S},\ \mathcal{L},\ C,\ \Pi),$$

where $\mathcal{S}$ is the set of available sensor modalities (e.g., accelerometer, gyroscope, microphone), $\mathcal{L}$ denotes their physical placement locations (e.g., wrist, belt, pocket), C captures environment-dependent constants (e.g., room acoustics, lighting), and Π is the prior over latent activities or behaviors. Each sensor $s \in \mathcal{S}$ produces a stream $x_s(t)$, and stacking all modalities/placements yields the observation $\mathbf{s}(t)$.

An on-device machine learning model M consumes only a subset $\mathcal{S}_m \subseteq \mathcal{S}$ of modalities over a sliding window to produce an exposed inference output

$$y(t) = M\big(\mathbf{x}_{\mathcal{S}_m}[t - \Delta,\ t]\big)\,.$$

This output is typically accessible to the adversary and forms one component of the observable interface in the privacy game that follows.

3.2 Target Privacy Definition

We define a target function

$$\mathrm{tgt} : \mathcal{X} \to \mathbb{R}^k,$$

which maps the full sensor state (a realization) $\mathbf{x}(t) \in \mathcal{X}$ to a sensitive physical or semantic attribute (e.g., fine-grained gesture, audio trace, environment class). The target need not be a raw sensor stream. In our experiments, it is typically a semantic attribute such as an activity or gesture class.

Figure 3 illustrates the adversary's task. Given only the compromised sensor modalities $\mathbf{x}_c(t)$, the model's exposed output $y(t)$,

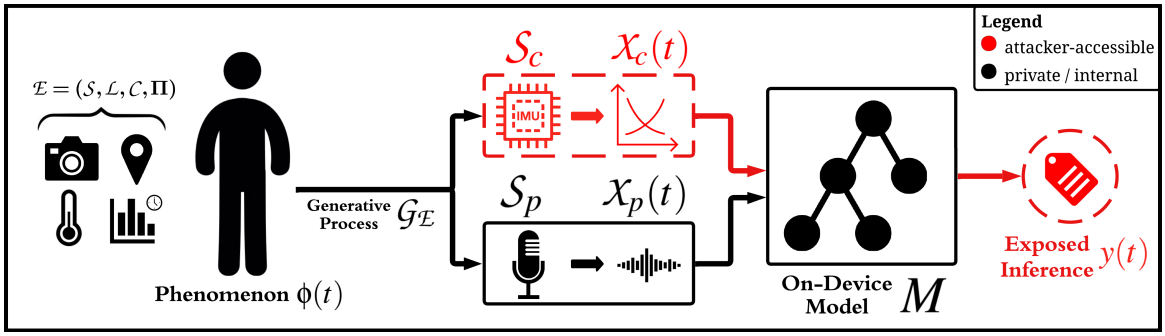

Figure 2: Sensing-to-inference pipeline and attacker-observable channel. A latent phenomenon $\phi(t)$ generates multimodal observations under environment $\mathcal{E} = (\mathcal{S}, \mathcal{L}, C, \Pi)$. The full sensor state $\mathbf{x}(t)$ is partitioned into adversary-visible components $\mathbf{x}_c(t)$ and private components $\mathbf{x}_p(t)$; an on-device model M consumes $\mathcal{S}_m \subseteq \mathcal{S}$ and releases adversary-visible output $y(t)$.

and any auxiliary knowledge, the adversary produces a prediction $\hat{z}(t)$ attempting to infer the hidden target value $\mathrm{tgt}(\mathbf{x}(t))$.

Adversarial Advantage. Let $\mathrm{Perf}(\cdot)$ be a task-appropriate metric and BaseAdv a trivial baseline (e.g., random guessing). We define the adversary's advantage as

$$\mathrm{Adv}(A) \triangleq \mathrm{Perf}\big(\hat{z}_A(t^\star),\ \mathrm{tgt}(\mathbf{x}(t^\star))\big) - \mathrm{BaseAdv}.$$

Under a specified interface, adversary class, and metric, we say the system satisfies tgt-privacy (under capability class $\mathcal{A}$) if for all $A \in \mathcal{A}$,

$$\mathrm{Adv}(A) \le \varepsilon,$$

for a tolerance ε fixed by the evaluation.

3.3 Adversarial Privacy Game

We model privacy leakage as an interactive game between a sensing system and an adversary A who observes only partial information and attempts to infer a private target.

Notation. Let $\mathcal{S}$ be the set of device sensors and $\mathcal{X}$ the space of their realizations. We partition sensors into compromised (adversary-visible) $\mathcal{S}_c \subseteq \mathcal{S}$ and private $\mathcal{S}_p = \mathcal{S} \setminus \mathcal{S}_c$. At time t, the full sensor state is $\mathbf{x}(t) = (\mathbf{x}_c(t), \mathbf{x}_p(t)) \in \mathcal{X}$. An on-device model M (using some subset $S_m \subseteq \mathcal{S}$) outputs $y(t)$. The private target is a function $\mathrm{tgt} : \mathcal{X} \to \mathbb{R}^k$.

Setup. The adversary's information is summarized as

$$Obs \ = \ \big(\mathbf{x}_c(\cdot),\ y(\cdot),\ Aux\big),$$

where $\mathbf{x}_c(\cdot)$ are the compromised streams, $y(\cdot)$ the exposed model outputs, and Aux optional auxiliary knowledge (e.g., public datasets $\mathcal{D}$, placement hints $\mathcal{L}_c$, environment constants C_c).

Game execution. Algorithm 1 summarizes how we instantiate attackers across all privacy games: the observable interface Obs determines the attacker input representation $\Phi(\cdot)$, the capability class $\mathcal{A}$ constrains the model family, and leakage is scored via $\mathrm{Adv}(A)$ under a fixed BaseAdv and tolerance ε. We next describe how different access regimes define Obs and how we operationalize Φ, $\mathcal{A}$, and Perf in experiments.

3.4 Quantifying Adversarial Leakage

We organize leakage evidence into three tiers that guide both theory and measurement.

Algorithm 1: Instantiating an adversary in the privacy game

Input: $Obs = (\mathbf{x}_c(\cdot), y(\cdot), Aux)$; target $\mathrm{tgt}(\cdot)$; metric Perf; baseline BaseAdv; tolerance ε.
Output: Adversary A and advantage $\mathrm{Adv}(A)$.

```
/* (1) Choose representation of observations   */
z ← Φ(Obs)
/* (2) Train                                   */
Fit A ∈ 𝒜 to predict tgt(x) from z using available data
/* (3) Predict at challenge time               */
ẑ(t*) ← A(Φ(x_c(t*)), y(t*), Aux))
/* (4) Score leakage and compute advantage     */
score ← Perf(ẑ(t*), tgt(x(t*)))
Adv(A) ≜ score − BaseAdv
/* (5) Break condition                         */
if Adv(A) > ε then
    return A, Adv(A)          // breaks tgt-privacy
return A, Adv(A)
```

(I) Impossibility baseline. If tgt fundamentally depends on private inputs (i.e., $\mathrm{tgt}(\mathbf{x})$ is independent of $(\mathbf{x}_c, y)$ under the data-generating process), then no function of $(\mathbf{x}_c, y, Aux)$ should beat BaseAdv + ε. This provides an upper bound on attainable inference.

(II) Achievability. When $\mathrm{tgt}(\mathbf{x}) = \mathrm{tgt}(\mathbf{x}_c)$ (or there exists a deterministic mapping from $(\mathbf{x}_c, y, Aux)$ to tgt), perfect inference is achievable in principle, indicating a fundamental privacy risk irrespective of algorithmic choices.

(III) Relaxed achievability. In realistic settings tgt may be only partially recoverable from $(\mathbf{x}_c, y, Aux)$. We quantify *non-trivial* leakage using: (a) task metrics (e.g., accuracy, F1, AUC), (b) reconstruction error for continuous targets (e.g., MSE, normalized ℓ_2), and (c) information-theoretic dependence (e.g., mutual information $I(\hat{z}; \mathrm{tgt})$). Leakage is present when these exceed BaseAdv by more than ε. Statistical hints may exist below the impossibility frontier, and the question is whether they rise above a practically meaningful threshold. In Section 5, we instantiate this relaxed-achievability view by varying interface properties such as granularity, temporal continuity, sampling rate, and cross-modal alignment while keeping the target and evaluation protocol fixed within each comparison.

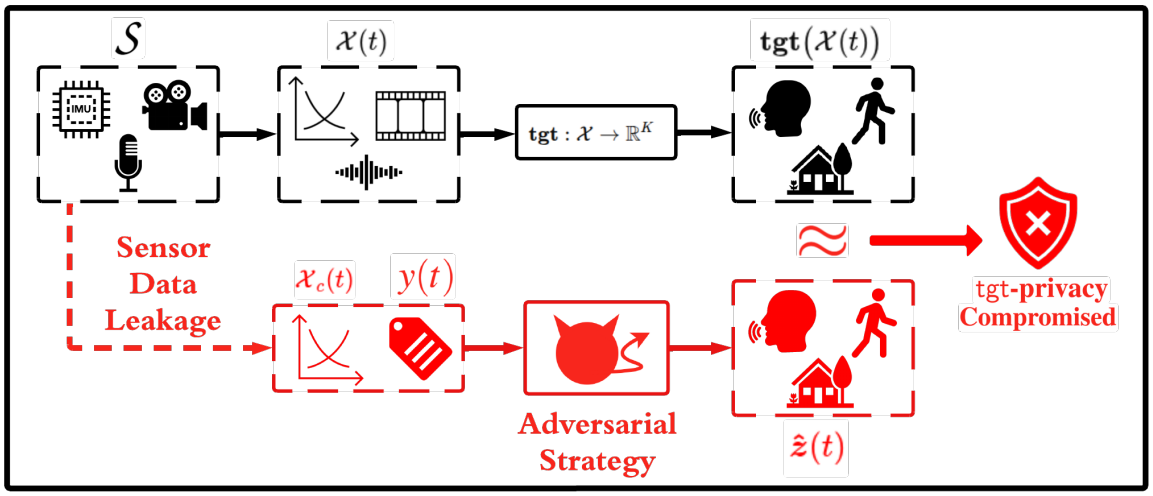

Figure 3: The adversary observes only the compromised sensor streams $\mathbf{x}_c(t)$, the model's exposed output $y(t)$, and any auxiliary knowledge. The adversary attempts to predict the private target value $\mathrm{tgt}(\mathbf{x}(t))$ and succeeds only if its performance exceeds a baseline (e.g., random guessing) by more than a small allowable margin ε. A system is tgt-private when no such adversary can achieve a non-negligible advantage.

We now instantiate these privacy framework formalisms in Section 4 with concrete signals, models, and datasets to evaluate leakage empirically.

4 Privacy Game Experimental Setup

Having established a formal framework for reasoning about privacy leakage in the previous section, we now turn to its empirical instantiation. The goal of this section is to ground the abstract notions of adversarial access, target functions, and leakage metrics in concrete experimental settings. We do this by defining the physical systems under study, specifying the adversarial games that capture realistic threat models, and explaining how we measure privacy degradation in practice. Before detailing the datasets and games, Table 2 summarizes how the evaluation probes different properties of the observable interface. This table serves as a roadmap for the results in Section 5.

4.1 Sensing Scenarios

To instantiate concrete environments of the defined privacy games in Section 3, we evaluate adversarial inference across four motion-sensing scenarios that vary in the latent phenomenon $\phi(t)$, sensor placements $\mathcal{L}$, and available modalities $\mathcal{S}$. Our aim is not to benchmark activity-recognition pipelines, but to span representative human–motion regimes whose structure (e.g., periodicity, limb asymmetry, placement-dependent distortion, and cross-limb coherence) directly shapes privacy leakage. Table 3 summarizes these scenarios.

Existing human-activity datasets lack two properties required to faithfully instantiate our privacy game: (1) *on-device inference outputs* reflecting a realistic embedded ML pipeline rather than offline labels, and (2) *synchronized multi-placement IMU* for the same subject, which is necessary to analyze how leakage propagates across sensor locations. To supply these elements, we collected a new dataset, **UTAH-STM-HAR**, using the SensorTile.box PRO platform [40]. We leverage the Machine Learning Core [39] to develop the on-sensor lightweight training and inference module that emits a *binary motion signal* $y(t)$ (e.g., "shake" vs. "no shake" [41]) each timestep.[1] We adopt this binary signal as the representative on-sensor inference throughout our evaluation because it mirrors the kind of aggressively compressed outputs used in commercial wearables and IoT devices, thereby providing a concrete instantiation of the exposed model output in our privacy game. Using this setup, eleven participants performed seven everyday activities (walking, jogging, upstairs, downstairs, sitting, standing, lying) while wearing five IMUs (left/right wrists, left/right ankles, front pocket). Each session contains raw 6-axis IMU signals, the binary on-device output $y(t)$, and anthropometric metadata.

To ensure that our findings are not artifacts of SensorTile hardware or its binary algorithm, we additionally evaluate on three widely-used benchmarks: **UCI-HAR** (waist-mounted smartphone; six activities) [32], **MotionSense** (front-pocket smartphone; six activities) [24], and **UTD-MHAD** (wrist + thigh IMUs with RGB video; 27 gestures) [5]. These datasets cover alternative sensing geometries, motion regimes, and activity taxonomies, enabling cross-domain and cross-modal evaluation of whether the privacy leakage we observe persists (or fails to generalize) across devices and environments.

On-device binary inference signal. All runs use a compact on-device pipeline configured for the LSM6DSOX: raw accel/gyro at 50 Hz, windowed features over 1.0–1.2 s (32–40 samples), and a shallow decision tree (a few levels) trained on features such as variance, peak-to-peak amplitude, zero-crossing counts, and energy. The resulting configuration deterministically generates a *binary inference*

[1] Full configuration, thresholding logic, and firmware parameters appear in Appendix B.

Interface factor	Formal object	Evaluation knob	Corresponding experiment
Granularity / channel breadth	Observable channel O_g or compromised stream set S_c: $O_{G1} = \{y(t)\}$, $O_{G2} = \{x_m(t), y(t)\}$, $O_{G3} = \{x^{acc}(t), x^{gyro}(t), y(t)\}$	Increase exposed detail from discrete output to one continuous modality to fused IMU: $G1 \rightarrow G2A/G2B \rightarrow G3$	UTAH-STM-HAR G1–G3 comparisons in §5.1 and Table 5; same/cross-placement transfer in Table 8 and App. Table 13; UTD-MHAD placement comparison in Table 10.
Temporal structure / interface resolution	Window sequence $\{O_g[t : t + W]\}_{t=1}^{T}$, temporal dependence across windows, sampling rate f_s, and window size W	Break or preserve temporal continuity; vary segmentation and sampling resolution: shuffled/context-free vs. contiguous/context-aware sessions, window-size sweep, and sampling-rate sweep	UTAH-STM-HAR context ablations in §5.1 and Table 5; sampling-rate sensitivity in Table 6; non-contextual window-size comparison in App. Table 12.
Adversary model	Adversary class $\mathcal{A}_g$	Compare simpler statistical or tree-based models with learned temporal models while keeping the observable interface fixed	UTAH-STM-HAR model/adversary robustness comparison in Table 7; checks whether the qualitative G1–G3 leakage ordering is stable across representative adversary classes.
Cross-modal alignment	Pretrained embedding $f(x_c)$ and auxiliary paired data $\mathcal{D}_{pair} = \{(x_i^{IMU}, x_i^{priv})\}$	Correctly aligned pairs vs. shuffled/mismatched pairs; reduce paired-data fraction	UTD-MHAD G4 ablations in §5.5 and Table 11: no-align IMU, aligned pairs, mismatched pairs, and reduced-pairing settings.

Table 2: Interface factors evaluated in our privacy games. Each row maps an observable-interface property to the corresponding evaluation knob and experiment. Within each comparison, we keep the target, data splits, feature pipeline, and adversary-selection protocol fixed where possible.

Dataset	Phenomenon $\phi(t)$	Placement(s)	Modalities	Target $\mathrm{tgt}(\mathbf{x})$
UTAH-STM-HAR	Daily human motion	Wrists, ankles, pocket	3-axis Accel, 3-axis Gyro	Activity class (7)
UCI-HAR	Body motion	Waist (smartphone)	3-axis Accel, 3-axis Gyro	Activity class (6)
MotionSense	Mobile motion	Front pocket (phone)	3-axis Accel, 3-axis Gyro	Activity class (6)
UTD-MHAD	Full-body gestures	Wrist, thigh (IMUs) + Video	Accel, Gyro, RGB video	Activity / gesture label (27)

Table 3: Summary of datasets and corresponding sensing scenarios. Each dataset observes a physical phenomenon $\phi(t)$ through specific sensor placements and modalities, defining concrete instantiations of the privacy-game variables $(\mathbf{x}_c, \mathbf{x}_p, y, \mathrm{tgt})$.

signal $yt) \in \{0, 1\}$ (event/no-event). We apply the same feature set and tree structure across all datasets to ensure consistency of the exposed interface. This fixed construction ensures that changes in leakage across G1–G3 reflect changes in the observable interface and temporal context rather than changes in the event detector itself.

4.2 Game Definitions and Access Models

We instantiate the privacy game of Section 3 by specifying what an adversary may observe at inference time, how this visibility maps onto the system variables (S_c, y, tgt), and what computational tools the adversary may apply. Algorithm 2 operationalizes Algorithm 1 by fixing Obs_g, Φ_g, A_g, and the selection rule per game. Each *privacy game* corresponds to a distinct visibility regime motivated by realistic leakage channels, e.g., device APIs often expose only coarse binary outputs, compromised apps may access a single IMU stream, firmware-level access exposes multiple modalities, and cross-modal latent spaces become available when pretrained representations can be queried or reverse-engineered. These regimes isolate orthogonal leakage axes (i.e., statistical regularities, intra-signal dynamics, cross-signal coherence, and representation alignment), allowing us to characterize how privacy degrades as adversarial visibility expands. Table 4 summarizes the set of privacy games we describe in detail below.

G1: Discrete-Output Adversary. The adversary observes only the device's exposed output $y(t)$, a discrete finite-state signal derived from inaccessible raw modalities. Formally, $Obs = \{y(t)\}$ and the adversary must infer tgt without access to any continuous sensor streams. Leakage arises solely from temporal structure or statistical bias in the device output.

G2: Single-Modality Adversary. Here the adversary observes one continuous sensor stream (e.g., accelerometer *or* gyroscope) in addition to the exposed output $y(t)$. Formally, $Obs = \{\mathbf{x}_m(t),\ y(t)\}$, where $\mathbf{x}_m(t)$ denotes the time series from a single compromised modality $m \in \mathcal{S}_c$ (and $\mathcal{S}_c$ contains exactly one modality in this game). The adversary may exploit temporal or spectral patterns within that single signal, but cannot leverage cross-channel relationships.

G3: Multi-Modal Adversary. The adversary obtains multiple raw sensor modalities simultaneously, $S_c = \{\mathbf{x}^{\mathrm{acc}}(t), \mathbf{x}^{\mathrm{gyro}}(t), y(t)\}$, enabling inference based on cross-signal coherence, joint dynamics, and multi-channel representations. This regime approximates an attacker with full access to embedded IMU data.

G4: Cross-Modal Representation Adversary. In this setting, the adversary does not directly observe additional sensors, but instead operates in a pretrained latent space that aligns heterogeneous modalities (e.g., video–IMU via a multimodal encoder). The adversary receives $S_c = \{f(\mathbf{x}(t)), y(t)\}$ where f maps raw signals into a shared embedding space; leakage emerges from representational alignment rather than raw signals themselves. This captures threats where an attacker leverages public pretrained models or paired data to infer protected attributes across sensing families.

Game	Observable interface Obs_g	Example	Exposure Mechanism
G1	Discrete outputs	Binary motion or FSM signal	Statistical regularities
G2	Single continuous modality	Accelerometer or Gyroscope	Intra-signal dynamics
G3	Multi-continuous modalities	Accelerometer and Gyroscope	Cross-signal correlation
G4	Latent cross-modal space	Video–IMU	Representation alignment

Table 4: Privacy games defined by the adversary's observable interface Obs_g and primary exposure mechanism. Each game represents a distinct interface/access regime used to evaluate interface-conditioned leakage.

Algorithm 2: Empirical instantiation of privacy games (Sections 3–4)

Input: Dataset $\mathcal{D}$ with sensor streams $\mathbf{x}(t)$ and targets Y; game $g \in \{\text{G1}, \text{G2}, \text{G3}, \text{G4}\}$ defining observable interface Obs_g; candidate adversary families $\{\mathcal{A}_{g,1}, \ldots, \mathcal{A}_{g,m}\}$; metrics Perf and baseline BaseAdv; selection tolerance δ (e.g., 1% NMI).

Output: Selected adversary $A_g^\star$ and leakage scores (Acc/F1/NMI/Vul/NRKL).

`/* (1) Construct observable interface and features */`

Derive Obs_g from $\mathcal{D}$ `// e.g., y(t) for G1; one modality for G2; accel+gyro for G3; embeddings for G4`

Choose context regime (context-free shuffling vs. context-aware sequences)

$\Phi_g \leftarrow$ feature map applied to Obs_g

Build $\Phi_g(Obs_g)$ using a fixed train/val/test split

`/* (2) Train candidate adversaries */`

for $i \leftarrow 1$ **to** m **do**

  Train $A_{g,i} \in \mathcal{A}_{g,i}$ to predict Y from $\Phi_g(Obs_g)$ (train set)

  Tune hyperparameters on validation set

`/* (3) Evaluate leakage */`

for $i \leftarrow 1$ **to** m **do**

  Compute test metrics: $\text{Acc}(A_{g,i}), \text{F1}(A_{g,i}), \text{NMI}(A_{g,i}), \text{Vul}(A_{g,i}), \text{NRKL}(A_{g,i})$

  $\text{Adv}(A_{g,i}) \leftarrow \text{NMI}(A_{g,i}) - \text{BaseAdv}_{\text{NMI}}$

`/* (4) Select least-complex near-optimal attacker */`

Let $i_{\max} \leftarrow \arg\max_i \text{NMI}(A_{g,i})$

Select $A_g^\star \leftarrow \arg\min_{A_{g,i}} \text{Complexity}(A_{g,i})$ s.t. $\text{NMI}(A_{g,i}) \geq \text{NMI}(A_{g,i_{\max}}) - \delta$

return $A_g^\star$ *and its leakage scores*

4.3 Adversarial Learning and Leakage Metrics

To operationalize the adversary in each privacy game, we evaluate how well an attack model can approximate the private target $\text{tgt}(\mathbf{x})$ from the observable variables $(\mathbf{x}_c(t), y)$ defined in Section 3. Because adversarial success can manifest as classification accuracy, regression quality, or distributional alignment, we quantify leakage using a combination of task-level and information-theoretic measures that map directly onto the relaxed-achievability tier (Section 3). This unified quantification protocol allows empirical comparisons across sensing regimes, datasets, and visibility levels, independent of the particular architecture used by the adversary. In the comparisons below, we keep the target, data split, feature pipeline, and adversary-selection rule fixed wherever possible, while varying the observable interface or context under study. This provides a controlled way to examine how changes in interface granularity, temporal structure, and cross-modal alignment affect measured leakage.

Task-Level Metrics. When tgt corresponds to a discrete activity label, we report standard classification metrics such as accuracy and F1-score. These capture how often the adversary recovers the correct class but are sensitive to the number of classes and to label imbalance, and thus cannot be compared directly across datasets or sensing scenarios.

Information-Theoretic Metrics. To obtain dataset-agnostic measures of leakage, we compute mutual information (MI) and Kullback–Leibler divergence (KL) between the true label distribution and the adversary's predicted probabilities. Because raw MI increases with the number of classes, we report *Normalized Mutual Information (NMI)*, defined as MI scaled by the entropy of the label distribution. NMI lies in $[0, 1]$ and captures the fraction of label uncertainty revealed by the adversary. Similarly, KL divergence is asymmetric and unbounded, making it difficult to interpret across datasets. We therefore report *Normalized Reverse KL (NRKL)*, which rescales KL by its theoretical maximum ($\log C$ for C classes) and inverts the scale so that higher values indicate better alignment between prediction and truth. NRKL provides a calibrated measure of how sharply and confidently the adversary matches the underlying distribution. We also compute Bayes vulnerability from the adversary's predicted posterior distribution, reporting the expected confidence of the adversary's best one-shot guess. This complements NMI by capturing guessing risk rather than aggregate uncertainty reduction.

Adversary Model Families. We instantiate adversaries using model families whose inductive biases match the visibility regime of each privacy game. Decision trees capture discrete thresholding in **G1**; shallow CNNs and temporal models capture local periodic structure in **G2**; lightweight cross-channel fusion architectures capture multi-stream coupling in **G3**; and linear probes over pretrained cross-modal embeddings instantiate **G4**. To avoid attributing leakage to excess model capacity, we select the *least-complex model whose NMI is within 1% of the best-performing model* per game and dataset. This ensures that observed leakage arises from increased adversarial visibility, not from increased representational power. This helps control for model capacity when comparing interfaces; in Table 7, we further report a model-robustness study showing that the qualitative leakage ordering across G1–G3 remains stable across representative adversary classes.

5 Results and Analysis

We empirically trace how privacy leakage evolves as adversary *access* and *context* increase. We first separate *statistical capacity* from *algorithmic exploitation* by starting with randomized, context-free windows and then restoring temporal continuity and richer channels. The results follow the interface factors in Table 2: §5.1 studies granularity, temporal structure, and interface resolution on UTAH-STM-HAR; §5.2–§5.4 study transfer and phenomenon distinguishability; and §5.5 studies cross-modal alignment through G4. We summarize our research questions as follows. **RQ1** *Statistical vs. Algorithmic Leakage*: How much private information is carried by the observation channel itself (e.g., a binary inference) versus additional leakage unlocked when temporal continuity, multi-channel signals, or side context are restored? **RQ2** *Spatial Transfer*: When does leakage persist under sensor relocation (e.g., train on one body site, test on another), and when do richer channels (e.g., coupling the binary inference output with an accessible and/or compromised accelerometer) make leakage strongly placement-specific? **RQ3** *Cross-Dataset Transfer*: Do leakage patterns survive changes in device, population, and collection protocol (cross-dataset, zero-shot), and how do the privacy metrics behave relative to accuracy under such shifts? **RQ4** *Phenomenon Distinguishability*: When do binary encodings collapse distinct activities (e.g., walk vs. jog) and when does added coverage or increased sampling resolve them? and **RQ5** *Representation-Level Leakage*: Can cross-modal alignment (e.g., learned video–IMU embeddings) bypass sensing-layer privacy bounds, and what minimal supervision suffices to mount such attacks?

5.1 Statistical vs. Algorithmic Leakage (RQ1)

We separate the signal's *statistical capacity* from *algorithmic exploitation* by first stripping context and then progressively restoring it. Concretely, windowed segments $\mathcal{W}$ are drawn uniformly so that subject, placement, and anthropometrics carry no information, $P(\mathcal{W} \mid Sub, L, A) = P(\mathcal{W})$, and temporal dependence factorizes, $\prod_i P(\mathcal{W}_{t_i})$ ("context-free"). We then re-enable contiguous sessions (temporal coherence) and optionally add placement and anthropometrics ("context-aware"). We report Games 1–3 here because they operate at the sensing layer (binary inference → accelerometer/gyroscope → IMU); Game 4 is a representation-level result discussed in §5.5.

Capacity baseline (Game 1). For C=7 classes, perfect discrimination requires $\log_2 C = 2.81$ bits. Empirically, the mutual information between the binary inference stream and activity is ≈ 0.7 to 0.8 bits, implying a low context-free accuracy ceiling ($\sim$ 35% via Fano's inequality). This is consistent with the context-free Game 1 results, where Accuracy and NMI remain low under shuffled windows; varying window size has comparatively small effects relative to restoring temporal continuity (App. Table 12).

Context enables algorithmic leakage. Restoring temporal coherence yields marked gains across Games 1-3 (Table 5). In Game 1, Accuracy climbs from near-random to $\sim$ 56% with NMI $\sim$ 0.5; Games 2A/2B improve further; Game 3 (accel+gyro) rises again as cross-channel structure becomes available. Two ablations isolate what matters: (i) removing anthropometrics changes results negligibly; (ii) anonymizing placement has similarly small effect. The dominant driver is *temporal coherence*, which unlocks **algorithmic leakage**: learned models exploit residual correlations (run-lengths, edge rates, transition rhythms) that are invisible under randomized sampling but present once time is restored. Vulnerability complements these task metrics by measuring one-shot guessing confidence; it generally rises as the interface preserves more temporal or continuous structure, but may differ from Accuracy because it measures posterior concentration rather than average correctness.

RQ1 Takeaway. Binary outputs impose a *statistical* bound (capacity); restoring context (especially time) lets adversaries climb toward that bound, and adding continuous channels *expands* the exploitable structure. In short: *capacity sets the ceiling; context and channels determine proximity to it.*

Design takeaway. Together, these comparisons show that privacy is shaped by practical interface levers, including output coarsening, temporal continuity, sampling rate, window size, and channel breadth. Sampling rate produces visible changes in the main sensitivity analysis, while window-size effects in the context-free setting are smaller and reported in App. Table 12.

5.2 Cross-Placement Leakage (RQ2)

We next ask whether leakage persists when models are trained on one *body placement* and evaluated on another. We keep temporal coherence intact but constrain training to a single, anonymized placement; each trained model is then tested on all five placements.

Binary streams transfer strongly. In **Game 1** (binary inference only), cross-placement performance closely matches same-placement performance. Averaged over placements (Table 8), Accuracy is negligibly different with 67.6% (same) vs. 66.4% (cross), while NMI is virtually unchanged with 54.0% (same) vs. 55.0% (cross). The full train–test placement matrix is reported in App. Table 13. This indicates that the binary edge stream encodes placement-invariant motion rhythms (e.g., run lengths/edge rates) that generalize across body locations.

Continuous channels are placement-specific. As depicted in Table 8, adding raw channels makes leakage *stronger but less transferable*. **Game 2A** (Accel), Accuracy drops from 97.1% (same) to 76.0% (cross), **Game 2B** (Gyro) falls from 96.2% to 75.2%, and **Game 3** (Accel+Gyro) drops from 98.1% to 77.9%. Information-theoretic metrics mirror this trend: NMI shrinks by $\tilde{8}$–9 points (e.g., 68.6%→60.8% in Game 3), and NRKL collapses sharply (e.g., 82.4%→55.5% in Game 3). These gaps reflect *placement-specific* kinematics (e.g., ankle vs. pocket dynamics) that models exploit when available but fail to carry over spatially.

Takeaway for RQ2. Leakage *persists* under spatial transfer for binary outputs (placement-invariant rhythms), but becomes *location-bound* once richer channels are exposed: continuous signals boost intra-placement accuracy yet degrade cross-placement transferability by ≈20–21 points, tightening the privacy–utility trade-off.

5.3 Cross-Dataset Transfer (RQ3)

We test whether the leakage patterns observed on UTAH-STM-HAR persist across populations and collection pipelines by training adversaries on public benchmarks (UCI-HAR at waist; MotionSense at front pocket) and evaluating zero-shot on all five STM placements, with no fine-tuning. We exclude non-overlapping classes (*lying* for

Metric: Accuracy (%)	Game 1	Game 2A (Accel)	Game 2B (Gyro)	Game 3
Placement + Temporal Coherence + Anthropometrics	56.1 ± 0.07	92.4 ± 0.04	82.0 ± 0.09	97.1± 0.04
Placement + Temporal Coherence	57.14 ± 0.06	97.14 ± 0.04	90.48 ± 0.07	98.7 ± 0.09
Temporal Coherence Only	67.2 ± 0.03	97.3 ± 0.08	87.6 ± 0.02	98.5 ± 0.07
Metric: F1-score (%)	**Game 1**	**Game 2A (Accel)**	**Game 2B (Gyro)**	**Game 3**
Placement + Temporal Coherence + Anthropometrics	52.9 ± 0.05	91.9 ± 0.04	81.5 ± 0.05	96.3 ± 0.09
Placement + Temporal Coherence	55.6 ± 0.06	97.1 ± 0.01	90.12 ± 0.04	93.08 ± 0.09
Temporal Coherence Only	64.03 ± 0.06	96.5 ± 0.03	87.5 ± 0.02	98.3 ± 0.03
Metric: NMI (%)	**Game 1**	**Game 2A (Accel)**	**Game 2B (Gyro)**	**Game 3**
Placement + Temporal Coherence + Anthropometrics	68.1 ± 3.8	92.53 ± 2.9	83.8 ± 6.5	99.01 ± 0.02
Placement + Temporal Coherence	68.2 ± 3.5	97.6 ± 2.7	90.4 ± 5.5	94.9 ± 5.4
Temporal Coherence Only	74.05 ± 5.29	97.9 ± 1.3	87.7 ± 2.2	98.9 ± 1.8
Metric: NRKL (%)	**Game 1**	**Game 2A (Accel)**	**Game 2B (Gyro)**	**Game 3**
Placement + Temporal Coherence + Anthropometrics	47.9 ± 4.4	78.7 ± 1.6	67.0 ± 1.1	81.8 ± 1.51
Placement + Temporal Coherence	51.4 ± 4.9	79.1 ± 1.9	72.1 ± 1.03	82.6 ± 3.0
Temporal Coherence Only	55.02 ± 8.1	82.5 ± 3.9	67.1 ± 1.8	82.2 ± 3.1
Metric: Vulnerability (%)	**Game 1**	**Game 2A (Accel)**	**Game 2B (Gyro)**	**Game 3**
Placement + Temporal Coherence + Anthropometrics	54.97 ± 0.01	66.20 ± 0.02	55.97 ± 0.02	68.23 ± 0.03
Placement + Temporal Coherence	62.28 ± 0.03	68.20 ± 0.02	59.76 ± 0.03	70.48 ± 0.04
Temporal Coherence Only	75.33 ± 0.03	73.46 ± 0.03	56.93 ± 0.01	70.69 ± 0.03

Table 5: UTAH-STM-HAR Dataset: Adversarial Performance across different context configurations for Games 1–3. Metrics are reported as mean ± std (%).

Freq	G1				G2A				G2B				G3			
	Acc	F1	NMI	Vul	Acc	F1	NMI	Vul	Acc	F1	NMI	Vul	Acc	F1	NMI	Vul
5 Hz	60.0	56.1	71.4	66.7	95.2	94.9	95.6	72.6	81.0	79.6	84.4	57.9	98.4	97.0	98.7	71.4
10 Hz	61.0	59.1	72.1	72.2	95.2	94.9	95.6	72.8	85.1	84.8	82.4	57.9	96.2	96.1	95.9	70.1
25 Hz	62.9	60.5	71.8	74.4	94.3	94.0	94.8	73.2	85.7	85.0	83.8	57.3	97.1	97.1	96.9	70.2
50 Hz	67.2	64.0	74.1	75.3	97.3	96.5	97.9	73.5	87.6	87.5	87.7	56.9	98.5	98.3	98.9	70.7

Table 6: Sampling-rate sensitivity. All metrics are reported in percentages.

Model	G1				G2A				G2B				G3			
	Acc	F1	NMI	Vul	Acc	F1	NMI	Vul	Acc	F1	NMI	Vul	Acc	F1	NMI	Vul
DT	58.1	56.1	71.6	80.0	81.9	80.7	84.7	84.0	76.2	74.9	82.3	67.5	88.6	88.5	88.0	81.2
RF	67.2	64.0	74.1	75.3	97.3	96.5	97.9	73.5	87.6	87.5	87.7	56.9	98.5	98.3	98.9	70.7
CNN	47.6	38.7	63.3	54.1	89.5	88.0	91.7	88.8	71.4	71.1	75.6	63.6	82.9	80.8	89.1	83.9
RNN	57.1	56.6	67.5	56.5	84.8	82.4	87.0	83.9	71.0	67.7	77.8	68.9	85.7	85.3	86.6	84.4

Table 7: Model robustness at 50 Hz across Games G1–G3. All metrics are reported in percentages.

Game (setting)	Acc	F1	NMI	NRKL	Vul
G1 (Same)	67.6	67.1	54.0	63.2	73.8
G1 (Cross)	66.4	65.3	55.0	58.8	73.4
G2A Accel (Same)	97.1	97.1	67.7	82.0	67.5
G2A Accel (Cross)	76.0	72.6	59.9	53.6	52.9
G2B Gyro (Same)	96.2	96.1	67.2	75.3	62.1
G2B Gyro (Cross)	75.2	72.6	59.4	57.6	52.1
G3 (Same)	98.1	98.1	68.6	82.4	66.8
G3 (Cross)	77.9	75.0	60.8	55.5	50.6

Table 8: UTAH-STM-HAR: Same-sensor vs. cross-sensor adversarial performance (percent, %).

MotionSense, *jogging* for UCI-HAR). Unless noted, the adversary is context-aware (contiguous sessions during training) and evaluated under *Game 1* to isolate leakage from the binary signal stream.

Table 9 shows that binary leakage transfers robustly despite device, placement, and population shifts. Average intra-dataset STM performance (same/cross placement averaged) is ≈66–67% Acc with NMI ≈54%. When transferring: (i) **UCI→STM**, accuracy is 48–53% across STM placements with NMI 59–63%; (ii) **MotionSense→STM**, accuracy is 59–63% with NMI 62–64%. While accuracy drops moderately (especially from waist-mounted UCI), NMI remains close to (and occasionally slightly higher than) STM intra-dataset NMI, reflecting distribution changes after class filtering. This indicates that the adversary preserves a comparable fraction of activity structure even across datasets.

The binary signal primarily captures coarse *motion rhythms* (edge rates, run-lengths, duty cycles) that are largely *placement-invariant* and population-agnostic. These statistics reflect physical regularities of human locomotion and posture rather than dataset idiosyncrasies, yielding stable NMI under domain shift.

In contrast, when raw channels are exposed (Games 2A/2B/3), our within-STM cross-sensor results (Table 8) already show strong *placement specificity*. Same-placement accuracy is about 97–98%, whereas cross-placement accuracy is about 73–78%. Consistent with that trend, cross-dataset generalization for Games 2–3 is markedly

Experiment Setup	Accuracy (%)	F1 (%)	NMI (%)	NRKL (%)	Vulnerability (%)
UCI → UTAH-STM-HAR (Cross-Dataset)					
UCI → Left Ankle	50.4	35.6	60.1	67.2	71.6
UCI → Right Ankle	50.5	36.6	60.4	66.9	72.0
UCI → Left Wrist	51.5	36.5	62.6	64.3	70.0
UCI → Right Wrist	48.5	36.9	59.1	69.09	69.5
UCI → Right Pocket	52.9	40.0	60.5	66.5	72.0
MotionSense → UTAH-STM-HAR (Cross-Dataset)					
MotionSense → Left Ankle	59.3	52.1	62.3	57.4	81.7
MotionSense → Right Ankle	58.6	52.7	62.8	56.6	83.9
MotionSense → Left Wrist	62.6	56.7	62.9	56.5	86.6
MotionSense → Right Wrist	61.4	55.7	63.8	62.7	82.8
MotionSense → Right Pocket	60.7	54.1	63.9	56.4	83.4

Table 9: Adversarial performance (Game 1) under cross-sensor and cross-dataset transfer. For the UCI→UTAH-STM-HAR setting, the adversary is trained on UCI-HAR and evaluated separately on each UTAH-STM-HAR sensor placement; the MotionSense→UTAH-STM-HAR setting follows the same protocol.

Placement	Metric	G1	G2A	G2B	G3
Right Thigh (6 Activities)	Acc (%)	65.1	98.0	97.8	99.9
	F1 (%)	63.1	98.0	97.7	99.3
	NMI (%)	61.6	96.7	96.4	99.9
	NRKL (%)	51.8	98.3	98.0	99.9
	Vul (%)	67.1	94.6	95.1	96.3
Right Wrist (21 Activities)	Acc (%)	26.2	74.9	74.4	86.2
	F1 (%)	20.9	74.1	72.5	85.6
	NMI (%)	70.9	83.5	82.1	90.4
	NRKL (%)	19.6	99.2	99.2	99.2
	Vul (%)	39.5	40.6	37.1	40.1

Table 10: UTD-MHAD same-sensor adversarial performance (%) across Games G1–G3.

worse (not shown in Table 9). Distributional shifts in device dynamics and feature marginals (e.g., energy, skewness, and kurtosis) reduce transferability even when class sets overlap.

Takeaway (RQ3). Binary interfaces leak *universal* activity structure that survives dataset shifts, as reflected in high NMI and moderate accuracy. Exposing continuous channels increases *local*, placement- and device-specific structure. That boosts within-dataset accuracy but *hurts* transfer, so it broadens the adversary's attack surface mainly in-domain.

5.4 Phenomenon Distinguishability and Placement Sensitivity (RQ4)

We ask when activities remain distinguishable under a highly constrained *binary* on-device interface (Game 1), and when distinguishability requires richer sensing channels (Games 2–3). We use UTD-MHAD because it provides two contrasting placements: a *right-thigh* IMU capturing six lower-body activities and a *right-wrist* IMU capturing 21 gesture-like upper-body activities. We evaluate Games 1–3 per placement and report Accuracy, Macro-F1, NMI, and NRKL in Table 10.

Binary outputs preserve coarse structure for periodic thigh motion, but degrade on diverse wrist gestures. On the thigh (6 activities), Game 1 achieves 65.1% accuracy with 61.6% NMI, while Games 2A/2B (Accel/Gyro) rise to ∼98% and Game 3 (Accel+Gyro) to ≈99.9%. On the wrist (21 activities), Game 1 drops to 26.2% accuracy (vs. uniform random $1/21 \approx 4.8\%$), whereas Games 2A/2B recover to ∼75% and Game 3 to 86.2%. Thus, for both placements, exposing continuous channels substantially increases adversarial success, but the *binary-only* interface is far more limiting for the wrist setting.

Interpretation: phenomenon and placement shape what survives quantization. These results are consistent with a simple view: lower-body activities at the thigh often exhibit structured, repeatable dynamics (e.g., gait/posture transitions) that can still leave informative temporal patterns even after aggressive quantization, whereas the wrist setting comprises a broader and more heterogeneous set of gestures whose distinctions depend more on signal amplitude, axis structure, and inter-channel dynamics that a binary interface does not retain. Exposing accelerometer and gyroscope channels restores this richer structure, narrowing the gap between placements and substantially improving distinguishability.

5.5 Representation-Level Leakage via Multimodal Binding (RQ5)

G4 evaluates semantic inference through aligned representations rather than reconstruction of hidden sensor streams. Table 11 summarizes the IMU-only UTD-MHAD ablations.

From sensing to representation. Sections §5.1–§5.4 examined leakage at the *sensing layer* (binary → accel/gyro → IMU). We now ask whether representation learning can bypass those sensor-level bounds. We evaluate multimodal binding using MMBind [30] as an off-the-shelf mechanism for aligning heterogeneous modalities through a shared bridge. In our setting, IMU is the adversary-visible modality at inference time, while Skeleton/video-derived annotations provide high-fidelity semantic structure during auxiliary alignment.

Modality bridging (UTD-MHAD: IMU ⇒ Skeleton/RGB semantics). We instantiate a tgt-privacy game in which the adversary observes only IMU streams, or their embedding representation $f(\mathbf{x}_c)$, and predicts the activity semantics associated with high-fidelity Skeleton/RGB annotations. We set compromised modalities $\mathcal{S}_c = \{\text{Acc}, \text{Gyro}\}$ and private modalities $\mathcal{S}_p = \{\text{Skeleton}, \text{RGB}\}$. The target is the activity label induced by the full realization $\mathbf{x}(t) \in \mathcal{X}$, i.e., $\text{tgt}(\mathbf{x}(t)) \triangleq Y(t)$, where $Y(t)$ is defined by the Skeleton/RGB

Setting	Paired Data	Acc. / F1 (%)
Our IMU-only G4 ablations		
No-align IMU	0%	55.2 / 52.3
Aligned pairs	100%	67.9 / 64.9
Mismatched pairs	100%	65.0 / 62.1
Aligned pairs	50%	66.8 / 63.6
Aligned pairs	25%	57.7 / 55.0
Original MMBind reference		
MMBind Skeleton–IMU (orig. paper)	Reported	78.9 / 76.3

Table 11: G4 semantic leakage on UTD-MHAD. The adversary observes only IMU (Accel+Gyro) at inference time and predicts activity/gesture labels associated with high-fidelity motion annotations. *No-align* measures ordinary IMU-only leakage; *Aligned* uses correct cross-modal pairing; *Mismatched* preserves paired-data size but breaks correspondence; and reduced-pairing rows weaken the alignment bridge. The original MMBind row is included only for comparability and is not part of our IMU-only G4 ablation protocol.

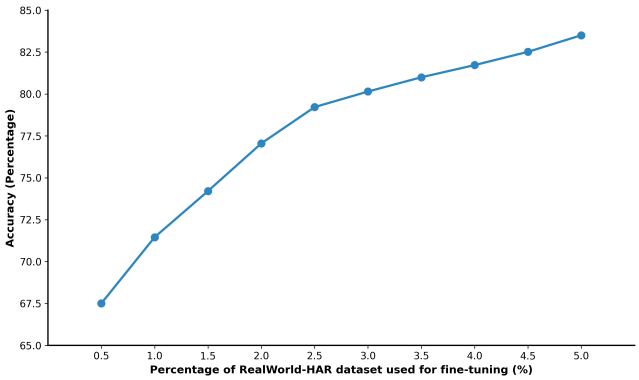

Figure 4: Representation-level leakage with MMBind: RealWorld accuracy vs. fraction of labeled data for fine-tuning. Severe leakage appears with $\leq 1\%$ labels.

annotation. In our notation, the observable interface is $Obs_{G4} = (f(\mathbf{x}_c(\cdot)), Aux)$, with optional access to $y(\cdot)$ when available. Here, f denotes the pretrained multimodal encoder used for binding.

Following MMBind [30], we use cross-node subject binding: IMU $\leftrightarrow$ Skeleton pairs are *not* directly seen in pretraining; alignment is induced indirectly through a shared channel. The aligned setting improves IMU-only activity inference over the no-alignment baseline, while mismatched and reduced-pairing settings weaken this effect. This indicates that cross-modal alignment contributes to semantic leakage beyond ordinary IMU-only inference, without implying reconstruction of hidden Skeleton/RGB streams.

Few-label thresholds for cross-domain alignment. We next probe how little labeled supervision suffices to unlock cross-domain leakage. We pretrain across siloed datasets using a shared modality/label space to create pseudo-pairs; examples include MotionSense (Acc+Gyro) and Shoaib (Acc+Mag). We then fine-tune on a tiny labeled slice of RealWorld (Acc+Gyro+Mag) and evaluate activity classification on the remaining RealWorld data. In this setting, auxiliary supervision aligns the embedding spaces, so even a small labeled slice enables a high-advantage probe despite differing raw sensing channels across datasets. Leakage emerges with *trivial* supervision. With only 0.5% labels, accuracy already exceeds 67%; at 1% it surpasses 71%; and by 2% it exceeds 77% (Fig. 4).[2]

Identity Linkage Under Domain Shift. We also investigated whether embeddings enable cross-dataset *user identity* linkage. We use RealWorld (Acc+Gyro+Mag) and MotionSense (Acc+Gyro), set $\mathrm{tgt}(\mathbf{x})$ to subject identity, and test whether embedding-space nearest neighbors transfer across datasets. After pretraining or fine-tuning with subject IDs retained, t-SNE clusters are dominated by *dataset* rather than user (App. Fig. 5). Cross-dataset nearest-neighbor matches are likewise sparse and weak (App. Fig. 6). Thus, under substantial hardware, placement, and protocol shift, identity linkage does *not* reliably transfer[3]. Matched devices and placements remain higher risk and are a priority for future tests.

Takeaway for RQ5. Multimodal binding enables *semantic modality bridging*: IMU-only access becomes more predictive of activity semantics associated with higher-fidelity modalities after cross-modal alignment. Leakage thresholds are also low ($\leq 1\%$ labeled data suffices to unlock strong cross-domain inference), while mismatched or reduced pairing weakens the representation-level bridge.

6 Discussion and Conclusion

On-Device Inference Does Not Imply Privacy. Our results show that restricting computation to the edge is insufficient to guarantee privacy: the dominant determinants are *what* the device exports and *what* temporal/contextual structure an adversary can observe. Binary outputs impose a statistical ceiling, but restoring temporal coherence and richer channels systematically enables algorithmic exploitation. This separation between *signal capacity* and *adversary context* helps explain why leakage persists even when raw streams never leave the device.

Scope and Extensibility of the Privacy Games. The four games stratify observable access regimes from sensing-layer interfaces (G1–G3) to representation-layer alignment (G4). This taxonomy organizes interface-conditioned leakage through quantization, temporal continuity, channel breadth, and cross-modal alignment. While not exhaustive, it is extensible: additional games can vary supervision, personalization, device heterogeneity, or auxiliary knowledge without changing the underlying advantage-based evaluation protocol.

Differential Privacy and Obfuscation. Differential privacy (DP) and output perturbation mitigate leakage by injecting randomness into training or release mechanisms. Our results address a complementary question: *given an exported interface*, how much target information remains inferable from the observable variables. DP can reduce overfitting-driven leakage and calibrate confidence, but it does not remove correlations induced by physics (e.g., periodic locomotion rhythms) unless utility degrades correspondingly. In this sense, DP/obfuscation interact with the same statistical capacity our

[2]Percentages reflect our fine-tuning sweep on RealWorld following the MMBind recipe; exact splits and seeds are provided in the appendix for reproducibility.

[3]See App. §C for methodology, diagnostics, and discussion of when identity linkage is more likely (e.g., matched devices/placements).

framework quantifies; the games remain applicable by measuring residual leakage under any perturbation budget. In our framework, release controls such as perturbation, aggregation, suppression, or rate limiting are treated as alternative interface designs; the same games can be rerun to measure residual leakage under the modified interface.

A Separability Proxy for Interface Risk. A recurring theme is that privacy improves when the observable channel is *less separable* with respect to the private target. This suggests a lightweight deployment-time proxy: estimate separability (e.g., via simple probes or information measures such as NMI) on the exported interface under representative temporal regimes. Interfaces whose outputs remain highly separable under context-aware observation should be treated as high-risk even if they are heavily quantized.

Beyond Discrete Outputs and Embedding Leakage. Although we instantiate with lightweight classifiers, the framework naturally extends to continuous and structured outputs produced by on-device foundation or generative models (embeddings, trajectories, token distributions). In such settings, success criteria can be phrased in divergence- or information-theoretic terms rather than accuracy alone. More importantly, embedding spaces can act as *semantic bridges*: cross-modal alignment can make low-fidelity sensor exports predictive of target-relevant structure associated with higher-fidelity modalities, even without reconstructing those modalities.

Design Guidance for Privacy-Preserving Sensing. The empirical gradients suggest three practical levers. First, temporal coherence is the dominant amplifier (RQ1): reducing long-range continuity (e.g., aggregation over short, jittered horizons) can reduce algorithmic leakage even when outputs remain locally useful. Second, channel breadth and placement matter (RQ2–RQ4): exposing multi-axis streams increases both within-placement leakage and placement-specific inference, tightening the privacy–utility tradeoff. Third, representation alignment can create a semantic bypass (RQ5): limiting unintended bridges, such as shared labels, ubiquitous pivot modalities, or paired-data exposure, can reduce cross-domain leakage without changing the on-device model.

Limitations and Failure Cases. Our experiments also identify conditions under which leakage weakens. Removing temporal continuity in the shuffled/context-free setting reduces algorithmic leakage (§5.1); coarsening the observable interface to G1 limits the available structure compared with G2/G3 (§5.1); and breaking or weakening cross-modal alignment through mismatched or reduced paired data reduces G4 semantic inference (§5.5). These cases show that leakage is conditioned on the observable interface and available context, rather than being an unavoidable property of edge inference. Our empirical claims remain grounded in motion-sensing datasets, the evaluated adversary classes, and the considered contexts; other modalities, richer auxiliary information, or stronger adversaries may change the measured leakage.

Validity, Transfer, and Measured Conservatism. We evaluate across multiple datasets, placements, and transfer settings to separate collection artifacts from physical regularities. Where results downtrend (e.g., cross-placement for rich IMU channels), the framework exposes *why* placement-specific dynamics rather than merely reporting lower accuracy. These choices make the measurements conservative for settings where stronger temporal continuity, tighter device homogeneity, or richer supervision may increase leakage under the same interface-conditioned evaluation.

Future Directions. Two directions follow naturally: (i) defense co-design that explicitly budgets statistical capacity against task utility under temporal constraints, and (ii) representation hygiene that detects and limits unintended cross-modal bridges during continual on-sensor learning. By grounding evaluation in target-specific privacy games, systems can iterate toward interfaces whose leakage is quantified and commensurate with their utility.

Conclusion. Edge inference changes *where* data are processed, but not necessarily *what* is leaked. Across sensing-layer games, we find that quantization limits capacity, yet temporal continuity and channel breadth unlock algorithmic leakage that remains measurable under transfer. At the representation layer, multimodal binding can make low-fidelity or seemingly public signals more predictive of private semantic attributes with minimal supervision, weakening privacy-by-locality assumptions. Together, these results argue for privacy specifications and evaluations that are *interface-conditioned*, reporting what an adversary can infer given the exported channel and context, not merely whether raw data stay local. These findings should be read as interface-conditioned measurements for motion sensing, not as a universal claim about all edge systems.

References

[1] Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. 2001. *Estimation with applications to tracking and navigation: theory algorithms and software.* John Wiley & Sons.

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

[3] Flavio P Calmon, Ali Makhdoumi, and Muriel Médard. 2015. Fundamental limits of perfect privacy. In *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 1796–1800.

[4] Thomas Carr and Depeng Xu. 2024. User Privacy in Skeleton-based Motion Data. In *2024 IEEE International Conference on Big Data (BigData)*. 8219–8221. https://doi.org/10.1109/BigData62323.2024.10825650

[5] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*. 168–172. https://doi.org/10.1109/ICIP.2015.7350781

[6] Hyunsung Cho, Akhil Mathur, and Fahim Kawsar. 2022. FLAME: Federated Learning across Multi-device Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (Sept. 2022), 1–29. https://doi.org/10.1145/3550289

[7] Thomas M Cover. 1999. *Elements of information theory.* John Wiley & Sons.

[8] Antreas Dionysiou and Elias Athanasopoulos. 2023. SoK: Membership Inference is Harder Than Previously Thought. *Proceedings on Privacy Enhancing Technologies* 2023, 3 (July 2023), 286–306. https://doi.org/10.56553/popets-2023-0082

[9] Antreas Dionysiou, Vassilis Vassiliades, and Elias Athanasopoulos. 2023. Exploring Model Inversion Attacks in the Black-box Setting. *Proceedings on Privacy Enhancing Technologies* 2023, 1 (Jan. 2023), 190–206. https://doi.org/10.56553/popets-2023-0012

[10] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. 2020. Ask the experts: What should be on an IoT privacy and security label?. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 447–464.

[11] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. 2020. Ask the Experts: What Should Be on an IoT Privacy and Security Label?. In *2020 IEEE Symposium on Security and Privacy (SP)*. 447–464. https://doi.org/10.1109/SP40000.2020.00043 ISSN: 2375-1207.

[12] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2021. An informative security and privacy "nutrition" label for internet of things devices. *IEEE Security & Privacy* 20, 2 (2021), 31–39.

[13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind One Embedding Space to Bind Them All. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457

[14] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. MOMENT: a family of open time-series foundation models.

In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 642, 38 pages.

[15] Bin Hu, Yan Wang, Jerry Cheng, Tianming Zhao, Yucheng Xie, Xiaonan Guo, and Yingying Chen. 2023. Secure and Efficient Mobile DNN Using Trusted Execution Environments. In *Proceedings of the ACM Asia Conference on Computer and Communications Security*. ACM, Melbourne VIC Australia, 274–285. https://doi.org/10.1145/3579856.3582820

[16] Ibrahim Issa, Aaron B Wagner, and Sudeep Kamath. 2019. An operational approach to information leakage. *IEEE Transactions on Information Theory* 66, 3 (2019), 1625–1657.

[17] Matthew Jagielski, Stanley Wu, Alina Oprea, Jonathan Ullman, and Roxana Geambasu. 2022. How to Combine Membership-Inference Attacks on Multiple Updated Models. https://doi.org/10.48550/arXiv.2205.06369 arXiv:2205.06369 [cs].

[18] Jacob Leon Kröger, Philip Raschke, and Towhidur Rahman Bhuiyan. 2019. Privacy implications of accelerometer data: a review of possible inferences. In *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy* (Kuala Lumpur, Malaysia) *(ICCSP '19)*. Association for Computing Machinery, New York, NY, USA, 81–87. https://doi.org/10.1145/3309074.3309076

[19] Ripan Kumar Kundu and Khaza Anuarul Hoque. 2025. SecretVR: Differential Privacy Defense Against Membership Inference Privacy Attacks in Virtual Reality. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Saint Malo, France, 1266–1267. https://doi.org/10.1109/VRW66409.2025.00279

[20] Oscar D Lara and Miguel A Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials* 15, 3 (2012), 1192–1209.

[21] Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Plötz. 2024. 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 (Aug. 2024), 1–32. https://doi.org/10.1145/3678545

[22] Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. 2014. From the information bottleneck to the privacy funnel. In *2014 IEEE Information Theory Workshop (ITW 2014)*. IEEE, 501–505.

[23] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. 2018. Protecting Sensory Data Against Sensitive Inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems* (Porto, Portugal) *(W-P2DS'18)*. ACM, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/3195258.3195260

[24] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile Sensor Data Anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation* (Montreal, Quebec, Canada) *(IoTDI '19)*. ACM, New York, NY, USA, 49–58. https://doi.org/10.1145/3302505.3310068

[25] Shentong Mo, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. 2024. IoT-LM: Large Multisensory Language Models for the Internet of Things. https://doi.org/10.48550/arXiv.2407.09801 arXiv:2407.09801 [cs].

[26] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2CLIP: Language-grounded Motion Sensor Translation with Multimodal Contrastive Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 13246–13253. https://doi.org/10.18653/v1/2023.findings-emnlp.883

[27] Andreas Mueller. 2019. *Modern robotics: Mechanics, planning, and control [bookshelf]*. Vol. 39. IEEE. 100–102 pages.

[28] Sayedeh Leila Noorbakhsh, Binghui Zhang, Yuan Hong, and Binghui Wang. 2024. Inf2Guard: an information-theoretic framework for learning privacy-preserving representations against inference attacks. In *Proceedings of the 33rd USENIX Conference on Security Symposium* (Philadelphia, PA, USA) *(SEC '24)*. USENIX Association, USA, Article 135, 18 pages.

[29] Alan V Oppenheim. 1999. *Discrete-Time Signal Processing*. Pearson Education India.

[30] Xiaomin Ouyang, Jason Wu, Tomoyoshi Kimura, Yihan Lin, Gunjan Verma, Tarek Abdelzaher, and Mani Srivastava. 2025. Mmbind: Unleashing the potential of distributed and heterogeneous data for multimodal learning in iot. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. 491–503.

[31] Eunji Park, Duri Lee, Yunjo Han, James Diefendorff, and Uichin Lee. 2024. Hide-and-seek: Detecting Workers' Emotional Workload in Emotional Labor Contexts Using Multimodal Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (Aug. 2024), 1–28. https://doi.org/10.1145/3678593

[32] Jorge Reyes-Ortiz, Alessandro Anguita, Davide adn Ghio, Luca Oneto, and Xavier Parra. 2013. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C54S4K.

[33] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. 2023. SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning. https://doi.org/10.48550/arXiv.2212.10986 arXiv:2212.10986 [cs].

[34] Mahadev Satyanarayanan. 2017. The emergence of edge computing. *Computer* 50, 1 (2017), 30–39.

[35] Sinem Sav, Abdulrahman Diaa, Apostolos Pyrgelis, Jean-Philippe Bossuat, and Jean-Pierre Hubaux. 2023. Privacy-Preserving Federated Recurrent Neural Networks. https://doi.org/10.48550/arXiv.2207.13947 arXiv:2207.13947 [cs].

[36] Reza Shokri, George Theodorakopoulos, and Carmela Troncoso. 2016. Privacy Games Along Location Traces: A Game-Theoretic Framework for Optimizing Location Privacy. *ACM Trans. Priv. Secur.* 19, 4, Article 11 (Dec. 2016), 31 pages. https://doi.org/10.1145/3009908

[37] Akash Deep Singh, Brian Wang, Luis Garcia, Xiang Chen, and Mani Srivastava. 2024. Understanding factors behind IoT privacy–A user's perspective on RF sensors. *arXiv preprint arXiv:2401.08037* (2024).

[38] Geoffrey Smith. 2009. On the Foundations of Quantitative Information Flow. In *Proceedings of the 12th International Conference on Foundations of Software Science and Computational Structures - Volume 5504*. Springer-Verlag, Berlin, Heidelberg, 288–302.

[39] STMicroelectronics. 2022. LSM6DSOX: machine learning core. https://www.st.com/resource/en/application_note/an5259-lsm6dsox-machine-learning-core-stmicroelectronics.pdf. Application note AN5259, Rev. 7. Accessed: 2026-02-28.

[40] STMicroelectronics. 2026. SensorTile.box PRO. https://www.st.com/en/evaluation-tools/steval-mkboxpro.html. Accessed: 2026-02-28.

[41] STMicroelectronics. 2026. STMems Finite State Machine Shake Detection Example. https://github.com/STMicroelectronics/STMems_Finite_State_Machine/tree/master/application_examples/lsm6dsv32x/Shake%20detection. Accessed: 2026-02-28.

[42] Petre Stoica, Randolph L Moses, et al. 2005. *Spectral analysis of signals*. Vol. 452. Pearson Prentice Hall Upper Saddle River, NJ.

[43] Xinyu Tang, Milad Nasr, Saeed Mahloujifar, Virat Shejwalkar, Liwei Song, Amir Houmansadr, and Prateek Mittal. 2022. Machine Learning with Differentially Private Labels: Mechanisms and Frameworks. *Proceedings on Privacy Enhancing Technologies* 2022, 4 (Oct. 2022), 332–350. https://doi.org/10.56553/popets-2022-0112

[44] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2021. Demystifying Membership Inference Attacks in Machine Learning as a Service. *IEEE Transactions on Services Computing* 14, 6 (Nov. 2021), 2073–2089. https://doi.org/10.1109/TSC.2019.2897554

[45] Xingchen Wang, Haoyu Wang, Feijie Wu, Tianci Liu, Qiming Cao, and Lu Su. 2024. Towards Efficient Heterogeneous Multi-Modal Federated Learning with Hierarchical Knowledge Disentanglement. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. ACM, Hangzhou China, 592–605. https://doi.org/10.1145/3666025.3699360

[46] Xi Wu, Matthew Fredrikson, Somesh Jha, and Jeffrey F. Naughton. 2016. A Methodology for Formalizing Model-Inversion Attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*. IEEE, Lisbon, 355–370. https://doi.org/10.1109/CSF.2016.32

[47] Hanshen Xiao and Srinivas Devadas. 2025. PAC Privacy and Black-Box Privatization. *IEEE Security & Privacy* 23, 4 (2025), 2–7. https://doi.org/10.1109/MSEC.2025.3563116

[48] Depeng Xu, Weichao Wang, and Aidong Lu. 2024. A Review of Motion Data Privacy in Virtual Reality. In *2024 International Conference on Meta Computing (ICMC)*. 186–194. https://doi.org/10.1109/ICMC60390.2024.00027

[49] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. ACM, Coimbra Portugal, 220–233. https://doi.org/10.1145/3485730.3485937

[50] Lilin Xu, Chaojie Gu, Rui Tan, Shibo He, and Jiming Chen. 2023. MESEN: Exploit Multimodal Data to Design Unimodal Human Activity Recognition with Few Labels. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. ACM, Istanbul Turkiye, 1–14. https://doi.org/10.1145/3625687.3625782

[51] Xin Yang and Omid Ardakanian. 2025. PrivDiffuser: Privacy-Guided Diffusion Model for Data Obfuscation in Sensor Networks. *Proceedings on Privacy Enhancing Technologies* 2025, 4 (2025), 40–55. https://doi.org/10.56553/popets-2025-0118

[52] Qingsong Yao, Yuming Liu, Xiongjia Sun, Xuewen Dong, Xiaoyu Ji, and Jianfeng Ma. 2024. Watch the Rhythm: Breaking Privacy with Accelerometer at the Extremely-Low Sampling Rate of 5Hz. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA, 1776–1790. https://doi.org/10.1145/3658644.3690370

[53] Tianya Zhao, Ningning Wang, and Xuyu Wang. 2025. Membership Inference Against Self-supervised IMU Sensing Applications. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. ACM, UC Irvine Student Center. Irvine CA USA, 268–281. https://doi.org/10.1145/3715014.3722060

Acknowledgments

This work was supported in part by the National Institutes of Health under Award Number R61MH135109 and by the National Science Foundation under Award No. 2425711. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NSF, the NIH, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereafter.

In addition to this support, we gratefully acknowledge our STMicroelectronics co-authors for their valuable technical sponsorship and advisory role in this work, particularly in the design and implementation of the embedded sensing platform, the configuration of the on-device machine-learning pipeline, and the experimental validation. Their expertise and close collaboration were essential in generating the UTAH-STM-HAR dataset and in grounding our privacy analysis in realistic edge-device settings. We also sincerely thank the broader STMicroelectronics team for their constructive feedback, technical expertise, and continued partnership throughout the project.

A Ethical Principles

Human Subjects and Consent. UTAH-STM-HAR was collected under Institutional Review Board (IRB) oversight at the authors' institution. All participants provided informed consent prior to data collection. The protocol (inertial sensing during everyday activities and collection of coarse, non-identifying anthropometric attributes) was approved by the IRB. We did not collect direct identifiers (e.g., names, contact information), and all records were stored with randomized subject codes. The additional datasets used in this paper are UCI-HAR, MotionSense, and UTD-MHAD. Each is publicly released in de-identified form by its respective provider.

Risk Minimization and Data Handling. IMU traces can encode sensitive behavioral information; accordingly, our study is designed to minimize risk. Our analyses do not attempt to re-identify individuals or infer protected attributes. All experiments are performed on de-identified data stored on secured institutional infrastructure with access restricted to the research team. Our evaluation focuses on *population-level* leakage characteristics of observable interfaces and access regimes, rather than targeted attacks on specific individuals or real-world deployments.

Responsible Communication of Methodology. This work characterizes interface-induced leakage properties of sensing pipelines and does not disclose previously unknown vulnerabilities in specific deployed products, vendors, or APIs. We describe methods at the level needed for scientific scrutiny and reproducibility, but avoid releasing materials that would directly enable harm (e.g., turnkey exploitation instructions for particular devices). Any future release of UTAH-STM-HAR-derived artifacts (data or trained models) will follow participant consent and an additional anonymization/review step.

B On-Sensor Machine Learning Pipeline and Configuration

We collected all motion data using the SensorTile.box PRO [40], which incorporates the LSM6DSOX IMU. The accelerometer (±4 g) and gyroscope (±500 dps) were recorded at 50 Hz. From these data, a representative subset was sampled to serve as the labeled examples for training the on-sensor shake/no-shake classifier. This training subset was then processed through the LSM6DSOX Machine Learning Core (MLC) pipeline [39, 41].

After the raw data were loaded into the LSM6DSOX, the feature-extraction stage was configured to compute windowed statistics over the accelerometer magnitude and axes using a 1.0–1.2 s window (32–40 samples at 50 Hz). Features such as variance, peak-to-peak amplitude, zero-crossings, and signal energy were selected because they reliably separate oscillatory shake patterns from idle or slow movement, enabling the pipeline to treat shake and no-shake segments in the logs as two distinct motion categories. These windowed features were then exported as an ARFF dataset and used to train a compact binary decision tree.

The resulting tree—typically only a few levels deep—captures simple threshold-based distinctions (e.g., high energy and frequent sign changes indicate shake, while low-amplitude low-variance segments correspond to no-shake). Within the MLC, this produces an algorithm that first evaluates whether the peak-to-peak magnitude or variance exceeds a small threshold and, if so, checks whether the zero-crossing count reflects an oscillatory motion pattern; windows that do not satisfy these conditions are assigned to the no-shake class. Once finalized, the decision tree and feature configuration were compiled into an ST MLC configuration file (.ucf), which encodes all required register settings, including sensor routing, filters, feature generators, and tree nodes. The resulting decision tree was then used to generate binaries for all the data in our datasets. A similar procedure was applied to the other public datasets, using the same feature-extraction and MLC training pipeline to ensure consistency across all evaluations.

C Identity Linkage Under Cross-Dataset Domain Shift

We examine whether MMBind's data-binding process unintentionally enables *identity linkage* across heterogeneous datasets. Concretely, we use *RealWorld* (Acc+Gyro+Mag) and *MotionSense* (Acc+Gyro), retaining subject identifiers throughout pre-training and fine-tuning (as in §5.5). After training the joint embedding model, we evaluate whether an adversary could match a RealWorld user's embedding to MotionSense subjects via cross-dataset nearest-neighbor search.

Evaluation.

(1) **Embedding geometry.** We compute t-SNE projections of the learned embeddings and color points by dataset and user ID. This reveals whether user-level structure persists across domains or whether embeddings separate primarily by dataset.

(2) **Cross-dataset retrieval.** For each RealWorld embedding, we retrieve the most similar MotionSense embeddings using cosine similarity. We report qualitative observations on nearest-neighbor matches and similarity distributions.

(3) **Within-dataset calibration.** We repeat the retrieval procedure within each dataset as a reference point, since inertial signals are known to contain biometric signatures (e.g., gait, motion style). This comparison helps quantify how much identifiability—if any—transfers across datasets.

Results. Figure 5 shows that RealWorld and MotionSense embeddings form clearly separated clusters, driven by differences in hardware, sensor placement, and data-collection protocol. User IDs do not form coherent cross-dataset groupings. Cosine similarity retrieval (Figure 6) identifies only a few potential "matches," and these exhibit weak similarity scores with no consistent same-user alignment.

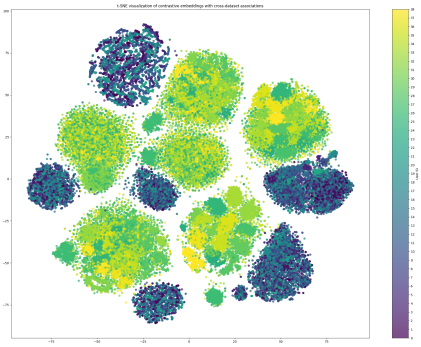

Figure 5: t-SNE of MotionSense and RealWorld embeddings (colored by user ID). Embeddings partition primarily by dataset, not identity, under cross-dataset shift.

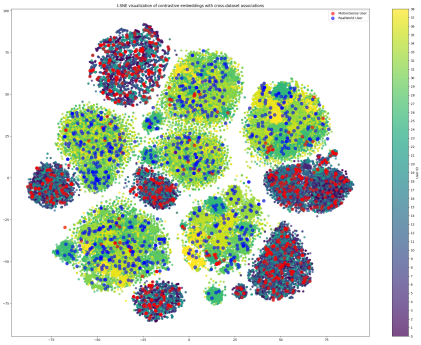

Figure 6: Cross-dataset nearest neighbors. Occasional elevated similarities appear but are inconsistent/weak, indicating limited adversarial leverage for identity linkage across datasets.

Implication. These findings suggest that under significant domain shift, MMBind's embeddings are unlikely to expose user identity across datasets. The embedding space is shaped far more by dataset-level factors than by individual movement signatures, making consistent cross-dataset biometric matching difficult for an adversary.

That said, this robustness is *not* guaranteed in general. Identity leakage may become more feasible when datasets are better aligned—for example, when they use similar devices, placements, sampling rates, or collection protocols. Such matched conditions represent higher-risk scenarios and should be a priority for future privacy evaluations.

D Window-size Sensitivity

Table 12 reports the non-contextual window-size sweep used to check whether the context-free Game 1 baseline is sensitive to segmentation length. Across the tested window sizes, Accuracy and NMI remain comparatively stable relative to the much larger gains observed when temporal continuity is restored in the main RQ1 analysis. We therefore use this table as a supporting sensitivity check rather than a primary main-body result.

	Accuracy				F1-score				NMI				NRKL				Vulnerability			
Window	G1	G2-A	G2-B	G3	G1	G2-A	G2-B	G3	G1	G2-A	G2-B	G3	G1	G2-A	G2-B	G3	G1	G2-A	G2-B	G3
0.5 s	34.8	59.3	50.3	65.2	22.1	58.4	48.5	64.3	26.8	35.1	29.7	38.1	38.3	99.1	98.9	99.0	30.1	68.9	58.1	69.7
1.0 s	35.3	61.9	51.6	65.3	23.3	61.2	49.8	64.4	28.7	36.9	30.6	38.8	37.6	99.0	98.8	99.1	30.2	69.9	60.6	69.6
1.5 s	34.5	63.2	52.1	67.8	23.1	62.6	50.4	67.0	26.9	37.5	31.0	40.1	38.9	99.1	98.8	99.3	30.3	70.5	61.4	69.3
2.0 s	30.4	66.0	53.6	69.5	24.3	65.5	51.9	68.7	28.8	39.1	31.9	41.2	37.0	99.5	98.6	99.8	30.4	70.8	62.4	69.5

Table 12: Non-contextual 50 Hz comparison across temporal window sizes (in seconds) for Games G1–G3. All metrics are reported as percentages.

E Full Game 1 Cross-Placement Matrix

Table 13 provides the full train–test placement matrix for the Game 1 binary-output adversary. The main text reports the same-vs.-cross placement averages; this appendix table shows that the same qualitative pattern holds across individual placement pairs: binary output streams transfer strongly across body locations, unlike richer continuous interfaces that become more placement-specific.

Train → Test Placement	Accuracy (%)	F1 (%)	NMI (%)
Left Ankle → Left Ankle	**61.9±7.4**	**61.4±8.1**	**55.7±2.9**
Left Ankle → Cross-Placement Avg.	69.0±7.6	67.0±8.6	56.6±3.3
Left Wrist → Left Wrist	**66.7±7.1**	**66.8±8.0**	**54.8±3.2**
Left Wrist → Cross-Placement Avg.	65.7±7.2	63.4±8.1	55.5±3.1
Right Ankle → Right Ankle	**71.4±6.9**	**71.1±8.3**	**53.6±2.5**
Right Ankle → Cross-Placement Avg.	64.1±7.6	63.0±8.4	54.8±2.9
Right Pocket → Right Pocket	**71.4±8.6**	**71.2±10.0**	**52.4±2.5**
Right Pocket → Cross-Placement Avg.	69.6±8.2	67.5±9.0	55.1±3.0
Right Wrist → Right Wrist	**66.7±7.7**	**65.1±8.2**	**53.7±2.7**
Right Wrist → Cross-Placement Avg.	67.2±7.9	65.6±8.5	54.7±2.9

Train → Test Placement	NRKL (%)	Vulnerability (%)
Left Ankle → Left Ankle	**57.2±21.9**	**72.3±5.5**
Left Ankle → Cross-Placement Avg.	62.0±17.7	71.9±1.2
Left Wrist → Left Wrist	**61.0±22.3**	**74.3±4.6**
Left Wrist → Cross-Placement Avg.	59.0±18.1	74.0±1.1
Right Ankle → Right Ankle	**70.1±23.8**	**75.1±5.3**
Right Ankle → Cross-Placement Avg.	58.8±18.5	72.0±2.6
Right Pocket → Right Pocket	**63.2±23.3**	**76.8±3.8**
Right Pocket → Cross-Placement Avg.	58.9±18.9	79.5±0.7
Right Wrist → Right Wrist	**64.3±18.8**	**70.4±3.9**
Right Wrist → Cross-Placement Avg.	60.4±16.1	69.8±0.7

Table 13: UTAH-STM-HAR Cross-Sensor Adversarial Performance for Game 1 (mean ± std). "Cross-Placement Avg." denotes the mean across all other test placements. Metrics are reported as mean ± std (%).