Breaking Sensor-Text Alignment: Cross-Modal Attack on Contrastive Multimodal Human Activity Recognition

Anonymous CVPR submission

Paper ID *****

Abstract

Embodied human activity recognition is a way to understand human behaviors through the connection of sensors and machines. In this setting, contrastive learning has become a key technique for aligning wearable sensor data with natural-language activity descriptions in human activity recognition (HAR). By learning a shared embedding space for sensor and text representations, these systems enable more interpretable and context-aware reasoning beyond simple classification. However, this alignment mechanism also creates a security risk: an attacker can manipulate the embedding space so that a sensor pattern is matched with an incorrect semantic description. In this paper, we present a cross-modal attack on sensor-text alignment in embodied HAR systems. Our method uses a search-based embedding poisoning attack at the text-embedding layer during inference and evaluates its impact on MotionSense, WISDM, HHAR, and WESAD. Across all datasets, the attack consistently reduces cosine similarity for correct sensor-text pairs, increases contrastive loss, and induces semantic drift in the shared representation space. Even very small perturbations, often below 5%, are sufficient to cause incorrect or inconsistent activity interpretations. These findings reveal vulnerability in embodied intelligence and highlight the need for more secure and robust alignment mechanisms in sensor-language systems.

1. Introduction

The integration of embedded intelligence, specifically sensor-based human activity recognition, with large language models, offers a new way to understand how machines understand human behavior. Traditionally, HAR systems rely heavily on machine learning models that use raw sensor signals from smartphones or wearables, such as accelerometers and gyroscopes, to identify daily activities like walking, sitting, or climbing stairs [10, 19, 38]. While these methods achieved high accuracy in controlled environments, they often struggled to capture the deeper context or meaning behind human actions. The large language models (LLMs) have changed this landscape by enabling connections between sensor data and semantic understanding [2, 15]. Through multimodal learning, models can now interpret not only **what** activity is occurring but also **why** and **how** it happens within a specific context (Fig. 1) [22, 40, 45].

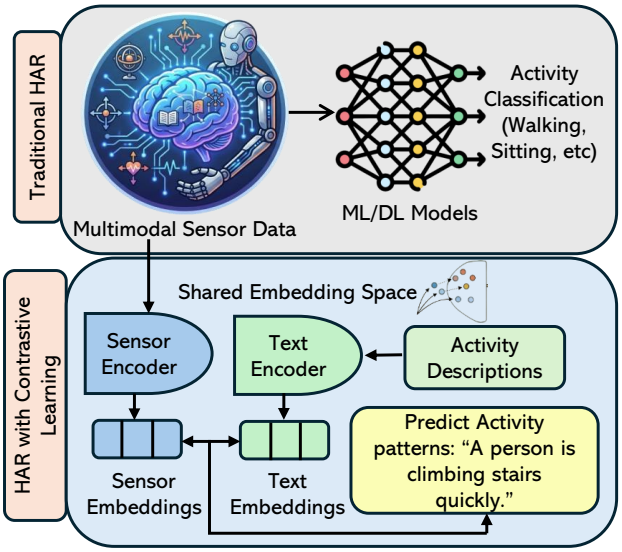


Figure 1. Traditional HAR directly classifies sensor data, whereas contrastive HAR aligns sensor and text embeddings in a shared space for activity recognition.

Another key method that enables this connection between sensor data and semantic interpretation is contrastive learning (CL), which aligns sensor signals with textual descriptions in a shared representation space. For example, data patterns from a smartwatch that record the acceleration and rotation of a person "walking upstairs" can be linked to the corresponding textual label or even a natural language sentence like "a person is climbing stairs quickly." Over time, the model learns to align semantically related

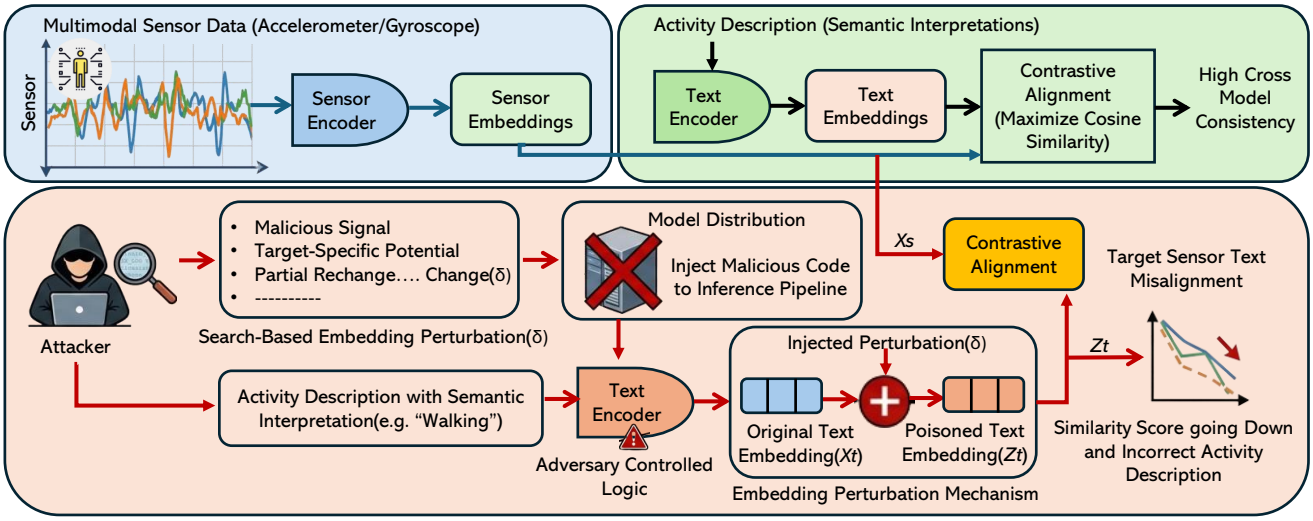


Figure 2. Overview of the proposed cross-modal embedding poisoning attack, where an adversary injects a target-specific perturbation into the text embedding during inference to disrupt sensor-text contrastive alignment and cause incorrect activity interpretation.

representations that describe the sensor patterns, such as those from the accelerometer and gyroscope, bringing them closer in the shared embedding space. This alignment enables the system to reason across modalities and generalize more effectively across users and environments, making it more adaptive, interpretable, and robust to unseen activities or noisy data [40].

These embodied systems, although powerful, also introduce significant security risks and fairness concerns. For instance, wearable devices that rely on photoplethysmography (PPG) sensors can be sensitive to skin tone, leading to measurement disparities across users. Some studies report differences of less than 5 bpm across skin tones for darker-skinned users, and the error rate during vigorous exercise can exceed 20% [3, 27, 29]. Beyond fairness, embodied and multimodal systems may also be vulnerable to cross-modal attacks, similar to those studied in image-text settings [42]. In addition, real-world deployments in corporate environments have revealed system-level vulnerabilities that attackers can exploit to target their systems by injecting malicious code or poisoning the library [18]. Together, these issues show that embodied intelligence can fail in two complementary ways: it may behave inequitably across users and, at the same time, expose a broader attack surface through its multimodal interfaces and real-world deployment constraints.

Motivated by these issues and security risks in embodied intelligence, in our study, we introduce a related cross-modal attack that degrades the performance of HAR systems, highlighting the need for effective defense mechanisms. Specifically, we apply a search-based embedding poisoning attack [44] at the text-embedding layer during the model inference phase, by introducing carefully designed perturbations that bypass safe alignment mechanisms such as reinforcement learning in human feedback and supervised fine-tuning, rather than relying on traditional jailbreak attacks that manipulate the input query or poisoning attacks that modify the training data. This attack tampers with the model's execution flow by injecting malicious code that persists within the compromised model. To quantify the practical impact of this threat in realistic HAR settings, we evaluate the attack across widely used benchmark datasets and measure how small embedding-level changes propagate into cross-modal misalignment.

Through extensive experiments on benchmark HAR datasets such as MotionSense [24], WISDM [17], HHAR [4], and WESAD [30], we demonstrate that even minimal perturbations, less than 5% variation in text embedding, can substantially degrade the alignment accuracy and cause significant semantic drift between sensor and textual embeddings. This degradation ultimately leads to incorrect or inconsistent activity interpretations, revealing important vulnerabilities in current multimodal contrastive learning frameworks.

To summarize, our contributions are listed as follows.

- The paper introduces the cross-model attack specifically designed to exploit vulnerabilities in the contrastive loss function used for aligning multimodal sensor data and natural language text by using a search-based embedding poisoning attack. This addresses an important, previously unexplored security gap in sensor-text systems.
- The research provides a rigorous evaluation through multiple benchmarks to demonstrate attack generalizability.
- We demonstrate that cross-modal search-based embedding poisoning can degrade HAR performance, under-

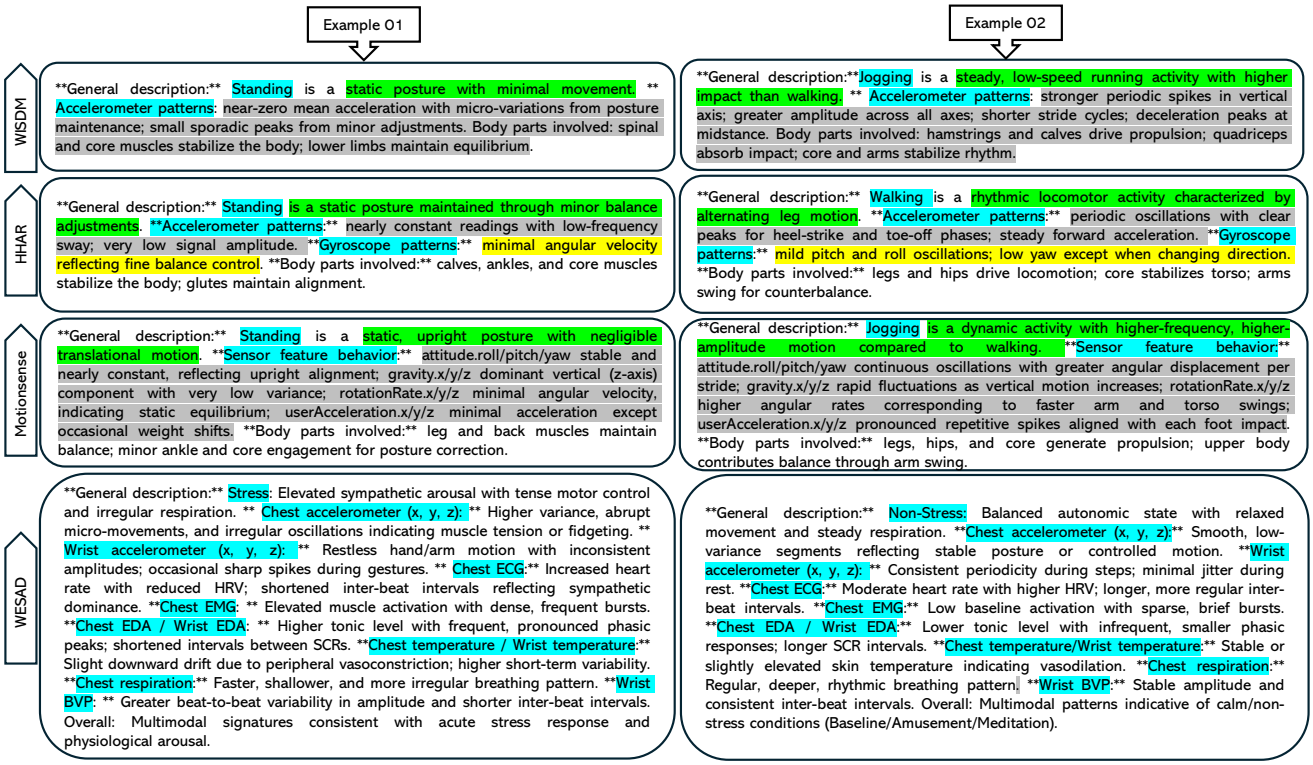


Figure 3. Semantic interpretations generated from ChatGPT of human activity and stress detection patterns across multiple sensor-based datasets: WISDM, MotionSense, HHAR, and WESAD.

scoring the need for robust defense mechanisms.

2. Related Works

Contrastive Learning for Sensor-based HAR. Contrastive learning is a self-supervised machine learning technique where the model learns by comparing pairs of data samples to understand which ones are similar and which are different. And CL addresses the limitation of the availability of labeled data for training; it teaches the model to learn meaningful representations by comparing similar (positive) and dissimilar (negative) data pairs. Tang et al [35]. investigate the use of contrastive learning, particularly the SimCLR framework, to improve HAR using wearable sensor data. Liu et al [23] introduce semi-supervised contrastive pre-training for training an encoder using both real and self-labeled data. Here, they efficiently integrate labeled and unlabeled data in a unified contrastive framework, achieving superior performance and label efficiency for HAR tasks. Chen et al [6]. propose the self-supervised framework that learns meaningful sensor representations from unlabeled time-series data, reducing dependence on costly manual labeling. Here, they construct positive pairs from temporally adjacent windows and negative pairs from distant ones and use InfoNCE loss to maximize similarity between positive pairs and minimize it for negatives.

Sensor-Text Alignment for HAR. Sensor-text alignment maps wearable sensor signals and natural language descriptions into a shared embedding space, bringing matched sensor-text pairs closer while pushing unrelated pairs apart [22]. It typically uses a sensor encoder to embed time-series data and a text encoder to embed descriptions. Recent foundation models, such as SensorLM [45], address the difficulty of applying standard LLMs to high-dimensional sensor streams by generating multi-level captions (statistical, structural, and semantic) and pretraining with contrastive and generative objectives (CLIP-style). Related work also uses LLMs to produce semantic interpretations of sensor readings and activity labels, aligning them in a language-based semantic space to improve recognition and generalization to new datasets and activities [40].

Attack on Encoders. Text-encoder attacks compromise the module that maps text to embeddings, so attacker-crafted inputs yield attacker-chosen representations that mislead downstream models. Zhao et al. [46] propose Buster, which embeds a backdoor in the text encoder of text-to-image diffusion models to redirect unsafe prompts to benign concepts instead of filtering outputs. Hu et al. [12] attack contrastive-learning encoders (SimCLR, MoCo-v2, Sim-

Siam) by adding imperceptible perturbations during pre-training and optimizing them with PGD/SGD. Yang et al. [41] study black-box poisoning by injecting a small number of samples for single-target image, single-target label, and multi-target label attacks. Wu et al. [39] show that text-embedding-based recommenders (BERT/RoBERTa) can leak more privacy than traditional ID-based systems. Lu et al. [21] introduce a training-free, data-free Trojan that implants hidden malicious behavior into a diffusion model's text encoder bottleneck MLP, triggered by specific words or symbols. Rocamora et al. [1] highlight that CLIP-style models have comparatively fragile text encoders under small perturbations (typos/character swaps) and propose Levenshtein-efficient adversarial fine-tuning to improve robustness.

2.1. Attacks on Sensor-Based HAR

Sensor systems are easy targets because they constantly sense the physical world, send data (often wirelessly), and operate with limited power and computing resources, sometimes in open, uncontrolled settings. That combination makes them vulnerable to attacks such as data tampering, spoofing, signal injection, and eavesdropping, which can compromise the integrity, confidentiality, and availability of the collected data [33, 34]. This risk is especially serious for sensor-based HAR. Attackers can manipulate sensor readings to trick ML models into predicting the wrong activity [16, 25]. Some attacks happen at inference time (evasion), where tiny perturbations are added to inputs [43]. Others happen during training (poisoning), where corrupted or mislabeled samples are injected to reduce accuracy or enable targeted misbehavior [26, 32]. Because HAR is widely used in healthcare, fitness, and wearable applications, these attacks can cause real harm, misleading decisions, false alerts, or unauthorized manipulation. This makes robust defenses and secure learning frameworks critical for trustworthy sensor-based AI [31].

Summary. From the literature review, we conclude that although several studies have investigated attacks on text encoders, they have not focused on objectives targeting sensor-text alignment through search-based embedding poisoning.

3. Proposed Method

In this section, we provide a detailed discussion of the search-based embedding poisoning attack on the cross-modal architecture as in Fig. 2

3.1. Threat Model

Target System. The victim system targeted by this attack is a deployed cross-modal HAR architecture built on a CL framework that uses sensor modalities [22]. The system is designed to align two distinct data types: continuous time-series sensor data, processed by the **sensor encoder**, and discrete semantic text descriptions of activities, processed by the **text encoder**. The model employs a standard contrastive loss to map both modalities into a shared embedding space, maximizing the similarity between correctly matched sensor-text pairs. For the attack, the system is assumed to be fully trained and operating solely in inference mode, meaning the model's parameters remain fixed throughout the attack.

Attacker Goal. The attackers' primary objective is to disrupt the similarity alignment between sensor and text pairs in the CL environment by introducing carefully crafted perturbations into the embeddings through search-based embedding poisoning. This attack is similar to a model poisoning attack in inference time, which modifies the models' parameters or structure.

Attacker Capabilities. We assume that the attacker can inject malicious code into the target model's execution flow to introduce carefully crafted perturbations that induce malicious behavior. This assumption is consistent with prior research on model poisoning attacks, in which adversaries manipulate training data or fine-tuning procedures to compromise the integrity of machine learning models.

3.2. Motivation Examples

A realistic attacker can exploit the trust in model distribution to inject malicious code to induce target sense-text misalignment.

Scenario. Consider a mobile health application running on Android or iOS and paired with a smartwatch. The system continuously collects inertial sensor data, including accelerometer and gyroscope readings, and feeds each sensor window into a cross-modal contrastive HAR model. The model aligns the sensor representation with a set of natural-language activity prototypes, such as walking, sitting, or climbing stairs, to infer the user's current activity.

Threat Example. In this setting, an adversary, "Eve," compromises the deployed HAR model and embeds malicious logic into its inference pipeline. Eve locally searches for a carefully crafted perturbation associated with selected target tokens and embedding dimensions. When the mobile health application processes a sensor window, the malicious code is triggered and applies a perturbation to the activity prototype's text embedding. Consequently, the model aligns the incoming sensor signal with an incorrect activity description, causing targeted sensor-text misalignment and degrading cross-modal consistency between the sensor and text encoders.

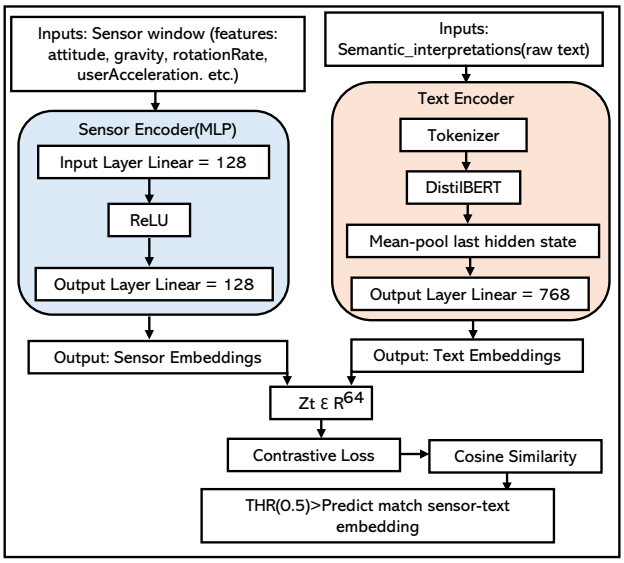


Figure 4. Dual-encoder architecture for training sensor–text pairs using contrastive learning.

3.3. Target Embedding Perturbation

The attack targets the *text-side representation* used for alignment. Specifically, the adversary perturbs the text embedding produced by the text encoder (or its input-embedding layer) to manipulate the resulting representation used in the contrastive matching stage. In our implementation, each activity class is associated with a natural-language *activity description* that serves as a text prototype for alignment (Fig. 3). These descriptions are generated using ChatGPT and then encoded by $\mathrm{Encoder}_t$ to obtain z_t.

For example, for the "Stairsdown" activity in the HHAR dataset, the prototype description is:

"General description: Stairsdown involves controlled descent with eccentric muscle actions to absorb impact and regulate speed. Accelerometer patterns: periodic impacts with sharper deceleration peaks at foot placement; vertical oscillations with slightly higher impact signatures than ascent; AP axis exhibits braking components. Gyroscope patterns: pitch decreases as the body lowers each step, there is a subtle roll from lateral stabilization, and there are occasional yaw adjustments at landings. Body parts involved: quadriceps and calves eccentrically control descent; hip stabilizers and core maintain balance; ankle strategy mitigates impact."

The attacker aims to induce a targeted mismatch by pushing the representation of this description toward that of a chosen target class "Walking" A representative target prototype is:

"General description: Walking is a rhythmic locomotion involving alternating leg movements. Accelerometer patterns: periodic oscillations with moderate amplitude; consistent peaks in vertical and AP axes representing heel strike and toe-off; low lateral variations indicate a steady gait. Body parts involved: lower limbs generate propulsion; hip flexors and extensors alternate; the core stabilizes torso motion."

Formally, given the benign embedding z_t for "Stairsdown" and the target embedding z_t^{target} for "Walking," the attacker computes a small perturbation δ_t such that $z_t^{\text{adv}} = z_t + \delta_t$ moves closer to z_t^{target} in the embedding space (and/or causes the sensor embedding z_s to rank the target description above the true description). This attack is therefore best interpreted as a *targeted alignment/retrieval attack* on the sensor-text matching mechanism, rather than a direct attack on a standalone HAR classifier.

Algorithm 1 attacks the text branch of a contrastive sensor-text model while keeping the sensor input fixed: it computes $z_s = \mathrm{Encoder}_s(x_s)$ and iteratively adds a bounded perturbation δ_t to the text embedding $z_t = \mathrm{Encoder}_t(x_t)$ to obtain $z_t^{\text{adv}} = z_t + \delta_t$. At each iteration, it maximizes the contrastive loss using projected gradient ascent, $\delta_t \leftarrow \mathrm{clip}\big(\delta_t + \alpha\,\mathrm{sign}(\nabla L_{\text{contrastive}}), -\epsilon, +\epsilon\big)$, and returns δ_t^* after T steps. In practice, updates are applied in the continuous token-embedding space (optionally followed by a decoding/projection step), and attack strength is governed by α, ϵ, T, and the negative sampling used in $L_{\text{contrastive}}$.

Algorithm 1 Cross-Modal Attack on Contrastive Loss

Require: x_s (sensor input), x_t (text input), $\mathrm{Encoder}_s$, $\mathrm{Encoder}_t$, step size α, perturbation bound ϵ, number of steps T

Ensure: Adversarial text perturbation δ_t^*

Initialize perturbation $\delta_t \leftarrow 0$
Compute sensor embedding $z_s \leftarrow \mathrm{Encoder}_s(x_s)$
for $i = 1$ to T **do**
  Compute perturbed text embedding $z_t^{adv} \leftarrow \mathrm{Encoder}_t(x_t) + \delta_t$
  Compute contrastive loss
  $L_{\text{contrastive}} \leftarrow -\log \frac{\exp(\mathrm{Sim}(z_s, z_t^{adv})/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{Sim}(z_s, z_j)/\tau)}$
  Compute gradient with respect to text embedding
  $g_t \leftarrow \nabla_{x_t} L_{\text{contrastive}}(z_s, z_t^{adv})$
  Update perturbation using gradient ascent
  $\delta_t \leftarrow \delta_t + \alpha \cdot \mathrm{sign}(g_t)$
  Project perturbation to satisfy the bound
  $\delta_t \leftarrow \mathrm{clip}(\delta_t, -\epsilon, \epsilon)$
end for
return $\delta_t^* \leftarrow \delta_t$

4. Experimental Setup

Datasets. We evaluate our attack on four benchmarks: three motion-based HAR datasets (WISDM [46], Motion-

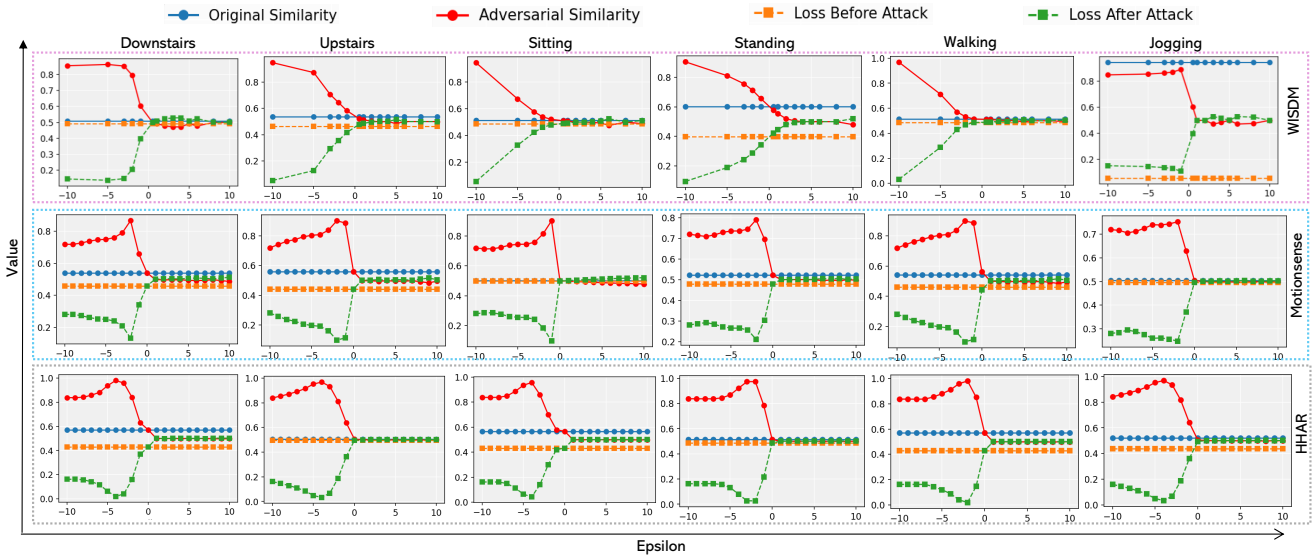
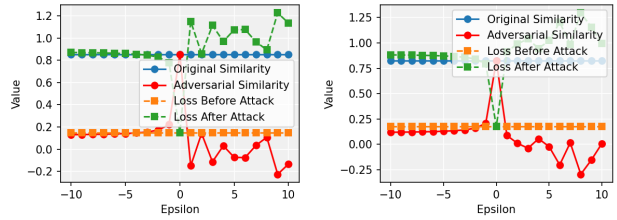


Figure 5. Attack behavior across datasets and activities: original vs. adversarial sensor-text cosine similarity and contrastive loss (before/after attack) as the perturbation budget ϵ increases. Each row corresponds to a dataset (WISDM, MotionSense, and HHAR), and each column to an activity; successful attacks reduce similarity and increase loss.



Stress Non-stress

Figure 6. Sensor–text encoder adversarial attack results on the WESAD dataset, showing original vs. adversarial similarities and corresponding contrastive loss values before and after perturbation.

Sense [24], HHAR [4]) and one physiological dataset (WESAD [30]). WISDM contains smartphone accelerometer data for common activities, including walking, jogging, stairs, sitting, and standing. MotionSense provides smartphone accelerometer and gyroscope signals for similar activities, and HHAR includes smartphone/smartwatch data to test device heterogeneity. WESAD captures multimodal biosignals such as ECG, EDA, respiration, temperature, and BVP under stress/non-stress conditions, enabling evaluation under different sensor modalities and semantics.

Rationale and Assumptions. This combination allows us to assess whether the proposed attack can (i) transfer across datasets that share similar motion patterns like Motionsense, HHAR, and WISDM, and (ii) remain effective on a dataset with distinct sensor modalities like WESAD and activity semantics. We assume small perturbations in text embeddings can shift the contrastive similarity boundary across modalities [47], and that consistent degradation across motion datasets indicates a mechanism-level vulnerability rather than dataset-specific bias [11, 28]. For sensor-text training, we generate short semantic descriptions for each label using ChatGPT-5 (Fig. 3).

Metrics. We measure sensor-text similarity using cosine similarity for vectors, $\mathbf{a}$ and $\mathbf{b}$, which is defined as $\frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|}$, and when both are unit-normalized, it reduces to their dot product. This metric is well-suited for contrastive learning because it captures angular (semantic) closeness in the shared embedding space [9]. In our experiments, we use a similarity threshold of 0.5: values above 0.5 indicate the pair is considered similar [7, 48].

4.1. Initial Setup

We learn a shared sensor-text embedding space on WISDM, MotionSense, and HHAR using contrastive learning so that matched sensor-text pairs are close and mismatched pairs are separated. We train on $N = 60{,}000$ pairs $\{(s_i, t_i, y_i)\}_{i=1}^{N}$, where s_i is a sensor segment, t_i is its ChatGPT-generated description and $y_i \in \{0, 1\}$ indicates aligned vs. non-aligned pairs. We use a sensor MLP encoder, f_θ, and a DistilBERT text encoder, g_ϕ, to map the inputs into a shared 64-dimensional space, as shown in Fig. 4. Specifically, $\mathbf{z}_s = f_\theta(s_i) \in \mathbb{R}^{64}$ and $\mathbf{z}_t = g_\phi(t_i) \in \mathbb{R}^{64}$. Alignment is measured by cosine similarity $\mathrm{sim}(\mathbf{z}_s, \mathbf{z}_t)$ and optimized with a margin-based contrastive loss ($m = 0.5$). We train for 30 epochs with batch size 4 and learning rate

Table 1. Mean Cosine Similarity for Positive and Negative Pairs Across Datasets During Training (Before Attack)

Pairs	WISDM	MotionSense	HHAR	WESAD
Positive Pairs	0.5715	0.6009	0.5769	0.7956
Negative Pairs	0.5805	0.6315	0.5896	0.8099

10^{-4}, training f_θ from scratch and fine-tuning g_ϕ.

4.2. Attack Setup

After training, we generate adversarial text embeddings using projected gradient ascent (PGD) by adding a perturbation to the original text embedding, yielding $\mathbf{z}_t^{\text{adv}} = \mathbf{z}_t + \boldsymbol{\delta}$. The perturbation is constrained by an L_∞ budget, $\|\boldsymbol{\delta}\|_\infty \leq |\epsilon|$, and is iteratively updated with step size α for K steps, with clipping after each update to ensure that it remains within the L_∞ ball. To evaluate attack strength, we sweep ϵ over the range $[-10, 10]$ while using the representative PGD setting $(\alpha, K) = (0.005, 1000)$ for MotionSense, HHAR, WISDM, and WESAD. For a 64-dimensional embedding space, the corresponding L_2 bound is $\|\boldsymbol{\delta}\|_2 \leq \sqrt{64}\,|\epsilon| = 8|\epsilon|$.

5. Results and Evaluation

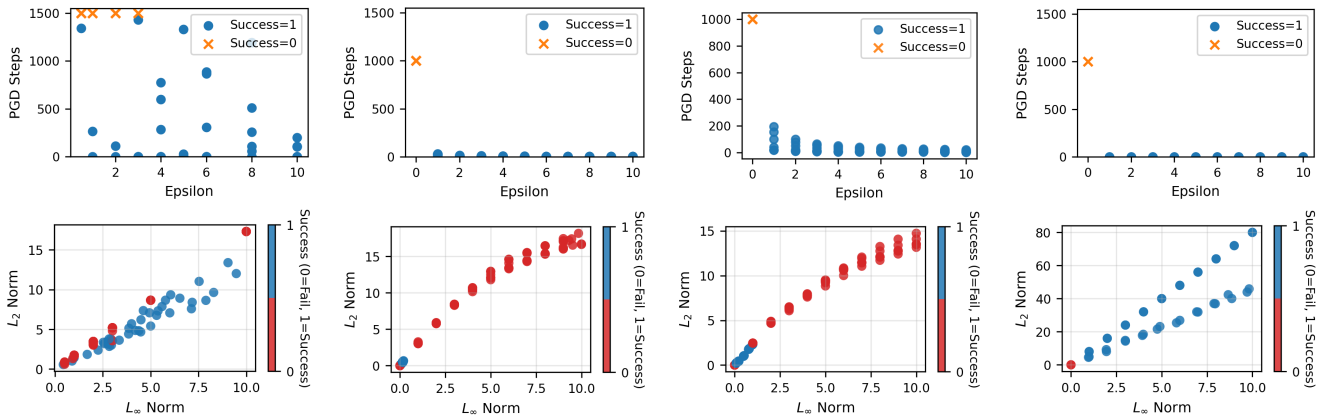
Results on the WISDM Dataset. Table 1 shows that both positive and negative cosine similarities are > 0.5 before attack (our detection threshold). In Fig. 5, the attack begins to succeed at different perturbation levels across activities. Specifically, the first successful attack occurs at $\epsilon = 0.5$ for *Downstairs*, $\epsilon = 1$ for *Jogging*, $\epsilon = 3$ for both *Sitting* and *Upstairs*, and $\epsilon = 4$ for *Standing* and *Walking*. This indicates that *Downstairs* is the most vulnerable activity, whereas *Standing* and *Walking* are comparatively more robust and require larger perturbations for successful misalignment. In Fig. 7 failed attacks are mostly concentrated at the maximum PGD budget of 1500 steps, while successful attacks require fewer steps as ϵ increases. The L_∞ and L_2 norms also grow approximately linearly, indicating that stronger perturbations lead to more reliable attack success in WISDM.

Results on the Motionsense Dataset. As in Table 1 before the attack, the mean cosine similarity of positive pairs is 0.6009, while the similarity of negative pairs is 0.6315. And after the attack, we can see in Fig. 5 the earliest successful attack occurs at $\epsilon = 1$ for every activity, and all change from failure to success at this point. After that, the attack remains successful for larger ϵ values. In fig. 7 failed attacks remain at the maximum PGD budget of 1000 steps, while successful attacks require only a few steps once it ϵ becomes positive. The L_∞ and L_2 norms also increase steadily, showing that MotionSense is highly sensitive to small effective perturbations.

Results on the HHAR Dataset. Before the attack, HHAR shows a cosine similarity of 0.5769 for positive pairs and 0.5896 for negative pairs (Table 1). Figure 5 shows that the attack consistently weakens sensor-text alignment, as the similarity decreases and the contrastive loss increases. Negative perturbations always fail and consume the full 1000 PGD steps, whereas positive perturbations succeed for all activities starting at $\epsilon = 1$. After the attack succeeds, the loss rises above 0.5, and the required PGD steps decrease, indicating that even small positive perturbations can reliably disrupt sensor-text alignment. Finally, Fig. 7 shows that failed attacks remain concentrated at the maximum PGD budget, while successful attacks require fewer steps as ϵ increases; meanwhile, the L_∞ and L_2 norms grow approximately linearly, suggesting that stronger perturbations lead to more reliable attack success in HHAR.

Results on the WESAD Dataset. Before the attack, both stress and non-stress pairs show high cosine similarity, with original similarities of 0.8516 and 0.8251, respectively, as shown in Table 1. Figure 6 shows that the attack consistently reduces sensor-text similarity and increases the contrastive loss for both classes. For the stress class, the adversarial similarity drops to -0.2260 at $\epsilon = 9$, while the loss increases from 0.1484 to 1.2260. For the non-stress class, the adversarial similarity reaches -0.2964 at $\epsilon = 8$, and the loss rises from 0.1749 to 1.2964. These results show that the attack is highly effective for both physiological states, although the strongest degradation occurs at different perturbation levels. Figure 7 further shows that nearly all nonzero perturbations succeed in a single PGD step, while only $\epsilon = 0$ a few fail. This indicates that the WESAD embedding space is highly sensitive to even small perturbations. In addition, the perturbation norms follow an approximately linear relationship, with $L_2 \approx 8L_\infty$. Since the attack succeeds almost uniformly whenever $L_\infty > 0$, the results suggest that WESAD is highly vulnerable to embedding-level adversarial perturbations.

Summary of Results. Across all four datasets, the attack consistently disrupted sensor-text alignment by reducing cosine similarity and increasing contrastive loss, demonstrating the vulnerability of contrastive multimodal HAR systems to embedding-level perturbations. Among the datasets, *MotionSense*, *HHAR*, and *WESAD* were particularly sensitive, as successful attacks emerged with very small perturbations, often at $\epsilon = 1$. In contrast, *WISDM* showed more variation across activities, where *Downstairs* was the most vulnerable and *Standing* and *Walking* required comparatively larger perturbations for successful misalignment. Overall, failed attacks were concentrated at the maximum PGD budget, while successful attacks required fewer optimization steps as ϵ increased, and both L_∞ and L_2 norms grew approximately linearly, indicating that stronger



(a) WISDM (b) MotionSense (c) HHAR (d) WESAD

Figure 7. Comparison of adversarial attack behavior across datasets. The top row shows the effect of perturbation strength (ϵ) on the number of PGD steps required for successful attacks. The bottom row shows the relationship between L_∞ and L_2 norms across datasets, illustrating attack success characteristics.

perturbations lead to more reliable attack success.

6. Conclusion and Future Directions

This work introduces a cross-modal adversarial attack that targets the contrastive objective used to align sensor and text embeddings in LLM-based HAR systems. By applying small perturbations in the text embedding space, the attack disrupts sensor-text alignment and induces incorrect activity predictions. Experiments on WISDM, MotionSense, and HHAR show that even minimal perturbations can reliably mislead the model, revealing a key vulnerability in multimodal contrastive learning and motivating defenses for safety-critical deployments (e.g., healthcare and daily-life monitoring). Future work will develop robust countermeasures that detect and mitigate embedding-level attacks without degrading alignment quality, including adversarial-aware training, more robust contrastive objectives, and detection/sanitization of corrupted text embeddings [13, 37]. We will also explore stabilization techniques such as Lipschitz-constrained encoders, randomized smoothing, and curvature-controlled embedding manifolds [8, 14, 20], as well as embedding purification methods (e.g., denoising autoencoders, diffusion-based purification, and vector-space projection defenses) to remove perturbations before alignment [5, 36].

References

[1] Elias Abad Rocamora, Christian Schlarmann, Naman Deep Singh, Yongtao Wu, Matthias Hein, and Volkan Cevher. Robustness in both domains: Clip needs a robust text encoder. *arXiv e-prints*, pages arXiv–2506, 2025. 4

[2] Hymalai Bello. Unimodal and multimodal sensor fusion for wearable activity recognition. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 364–365. IEEE, 2024. 1

[3] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, 3(1):18, 2020. 2

[4] Bhattacharya Blunck, Henrik, Sourav, Prentow, Thor, Kjrgaard, Mikkel, Dey, and Anind. Heterogeneity Activity Recognition. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5689X. 2, 6

[5] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4562–4572, 2023. 8

[6] Xiaobing Chen, Xiangwei Zhou, Mingxuan Sun, and Hao Wang. Temporal contrastive learning for sensor-based human activity recognition: A self-supervised approach. *IEEE Sensors Journal*, 2024. 3

[7] Inseop Chung, Daesik Kim, and Nojun Kwak. Maximizing cosine similarity between spatial features for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1351–1360, 2022. 6

[8] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019. 8

[9] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 6

[10] Mohammad Mehedi Hassan, Shamsul Huda, Md Zia Uddin,

Ahmad Almogren, and Majed Alrubaian. Human activity recognition from body sensor data using deep learning. *Journal of medical systems*, 42(6):99, 2018. 1

[11] Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. Crosshar: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–26, 2024. 6

[12] Aodi Hu, Zhiyong Zhang, Gaoyuan Quan, and Xinxin Yue. Idpa: Indiscriminate data poisoning attacks targeting pretrained encoder based on contrastive learning. *ACM Trans. Priv. Secur.*, 28(4), 2025. 3

[13] Jiacheng Huang and Long Chen. Defense against adversarial attacks via textual embeddings based on semantic associative field. *Neural Computing and Applications*, 36(1):289–301, 2024. 8

[14] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 8

[15] Haru Kaneko and Sozo Inoue. Toward pioneering sensors and features using large language models in human activity recognition. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pages 475–479, 2023. 1

[16] Malithi Wanniarachchi Kankanamge, Abdur R. Shahid, and Ning Yang. Poster: Leveraging large language models for zero-shot detection and mitigation of data poisoning in wearable AI systems. In *GenAI for Health: Potential, Trust and Policy Compliance*, 2024. 4

[17] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011. 2

[18] Piergiorgio Ladisa, Merve Sahin, Serena Elisa Ponta, Marco Rosa, Matias Martinez, and Olivier Barais. The hitchhiker's guide to malicious third-party dependencies. In *Proceedings of the 2023 workshop on software supply chain offensive research and ecosystem defenses*, pages 65–74, 2023. 2

[19] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3):1192–1209, 2012. 1

[20] Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B Grosse, and Jörn-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. *Advances in neural information processing systems*, 32, 2019. 8

[21] Xindi Li, Zhe Liu, Tong Zhang, Jiahao Chen, Qingming Li, Jinbao Li, and Shouling Ji. TWIST: Text-encoder weight-editing for inserting secret trojans in text-to-image models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11025–11041, Vienna, Austria, 2025. Association for Computational Linguistics. 4

[22] Z Li, S Deldari, L Chen, H Xue, and FD Salim. Sensor-llm: Aligning large language models with motion sensors for human activity recognition. arxiv 2024. *arXiv preprint arXiv:2410.10624*. 1, 3, 4

[23] Dongxin Liu and Tarek Abdelzaher. Semi-supervised contrastive learning for human activity recognition. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 45–53. IEEE, 2021. 3

[24] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, pages 49–58, New York, NY, USA, 2019. ACM. 2, 6

[25] WKM Mithsara, Abdur R Shahid, and Ning Yang. Zero-shot detection and sanitization of data poisoning attacks in wearable ai using large language models. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 1510–1515. IEEE, 2024. 4

[26] WK Mithsara, Ning Yang, Ahmed Imteaj, Hussein Zangoti, and Abdur R Shahid. Adaptive and robust data poisoning detection and sanitization in wearable iot systems using large language models. *arXiv preprint arXiv:2511.02894*, 2025. 4

[27] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 284–285, 2020. 2

[28] Xin Qin, Yiqiang Chen, Jindong Wang, and Chaohui Yu. Cross-dataset activity recognition via adaptive spatial-temporal transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3 (4):1–25, 2019. 6

[29] Borja Sañudo, Moisés De Hoyo, Alejandro Muñoz-López, John Perry, and Grant Abt. Pilot study assessing the influence of skin type on the heart rate measurements obtained by photoplethysmography with the apple watch. *Journal of Medical Systems*, 43(7):195, 2019. 2

[30] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, page 400–408, New York, NY, USA, 2018. Association for Computing Machinery. 2, 6

[31] Abdur R. Shahid, Ahmed Imteaj, Peter Y. Wu, Diane A. Igoche, and Tauhidul Alam. Label flipping data poisoning attack against wearable human activity recognition system. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 908–914, 2022. 4

[32] Abdur R. Shahid, Ahmed Imteaj, Shahriar Badsha, and Md Zarif Hossain. Assessing wearable human activity recognition systems against data poisoning attacks in differentially-private federated learning. In *2023 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 355–360, 2023. 4

[33] Amit Kumar Sikder, Hidayet Aksu, and A Selcuk Uluagac. {6thSense}: A context-aware sensor-based attack detector for smart devices. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 397–414, 2017. 4

[34] Amit Kumar Sikder, Giuseppe Petracca, Hidayet Aksu, Trent Jaeger, and A Selcuk Uluagac. A survey on sensor-based threats and attacks to smart devices and applications. *IEEE Communications Surveys & Tutorials*, 23(2):1125–1159, 2021. 4

[35] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020. 3

[36] Kaibo Wang, Xiaowen Fu, Yuxuan Han, and Yang Xiang. Diffhammer: Rethinking the robustness of diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 37:89535–89562, 2024. 8

[37] Yibin Wang, Yichen Yang, Di He, and Kun He. Robustness-aware word embedding improves certified robustness to adversarial word substitutions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 673–687, 2023. 8

[38] Qingxin Wei, Jiaming Huang, Yi Gao, and Wei Dong. One model to fit them all: Universal imu-based human activity recognition with llm-assisted cross-dataset representation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9 (3), 2025. 1

[39] Zhiyong Wu and Yin Chen. Tram-s: A switch transformer-based textual recommendation attack model. In *2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 1866–1869. IEEE, 2025. 4

[40] Hua Yan, Heng Tan, Yi Ding, Pengfei Zhou, Vinod Namboodiri, and Yu Yang. Language-centered human activity recognition. *arXiv preprint arXiv:2410.00003*, 2024. 1, 2, 3

[41] Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 4

[42] Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning*, pages 39299–39313. PMLR, 2023. 2

[43] Guolin Yin, Junqing Zhang, Xinping Yi, and Xuyu Wang. Evasion attacks and countermeasures in deep learning-based wi-fi gesture recognition. *IEEE Transactions on Mobile Computing*, 2025. 4

[44] Shuai Yuan, Zhibo Zhang, Yuxi Li, Guangdong Bai, and Wang Kailong. Embedding poisoning: Bypassing safety alignment via embedding semantic shift. *arXiv preprint arXiv:2509.06338*, 2025. 2

[45] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed Metwally, Jinhua Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 3

[46] Xin Zhao, Xiaojun Chen, Yuexin Xuan, and Zhendong Zhao. Buster: Incorporating backdoor attacks into text encoder to mitigate nsfw content generation. *arXiv e-prints*, pages arXiv–2412, 2024. 3, 5

[47] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6311–6320, 2023. 6

[48] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10306–10315, 2021. 6