



A Large-Scale Multimodal Dataset and Benchmarks for Human Activity Scene Understanding and Reasoning

Siyang Jiang¹, Mu Yuan¹, Xiang Ji¹, Bufang Yang¹
 Zeyu Liu², Lilin Xu³, Yang Li¹, Yuting He¹, Liran Dong¹, Wenrui Lu¹
 Zhenyu Yan¹, Xiaofan Jiang³, Wei Gao⁴, Hongkai Chen^{1,✉}, Guoliang Xing^{1,✉}
¹The Chinese University of Hong Kong, Hong Kong, ²University of Illinois Urbana-Champaign, United States,
³Columbia University, United States, ⁴University of Pittsburgh, United States.

Abstract

Multimodal human action recognition (HAR) utilizes complementary data for activity classification. Built on traditional HAR tasks, recent advances in Large Language Models (LLMs) enable detailed descriptions and causal reasoning of human actions, advancing new tasks of human action understanding (HAU) and human action reasoning (HARn). However, most LLMs, especially multimodal Large Vision-Language Models (LVLMs), struggle with modalities other than RGB images, like depth, IMU, or mmWave, due to a lack of large-scale `<data, caption>` datasets in these task domains. Existing HAR datasets provide only coarse-grained `<data, label>` annotations, insufficient for depicting the detailed action dynamics required in HAU and HARn tasks. Simply combining annotations and generating captions with LLMs often lacks necessary logical and spatiotemporal consistency. In this paper, we introduce CUHK-X, a large-scale multimodal dataset and benchmarks for HAR, HAU, and HARn. It includes 64,267 samples of 40 actions performed by 30 participants across two indoor environments, covering diverse daily scenarios. To address the challenge of spatiotemporal inconsistencies in captions, we propose a prompt-based scene creation method that leverages LLMs to generate logically connected activity sequences. CUHK-X also includes three benchmarks with six tasks to evaluate state-of-the-art models. Experimental results show average accuracies of 76.52% for HAR, 40.76% for HAU, and 70.25% for HARn. This large-scale multimodal dataset aims to empower the research community to apply, develop, and adapt data-intensive learning techniques for a wide range of human activity-related tasks.

CCS Concepts

• Human-centered computing → Ubiquitous and mobile computing systems and tools; • Computing methodologies → Activity recognition and understanding.

Keywords

Human action understanding, Large language models, Datasets

✉ denotes the corresponding author. {hkchen, glxing}@ie.cuhk.edu.hk.



This work is licensed under a Creative Commons Attribution 4.0 International License.
 MobiSys '26, Cambridge, United Kingdom
 © 2026 Copyright held by the owner/author(s).
 ACM ISBN 979-8-4007-2027-7/2026/06
<https://doi.org/10.1145/3745756.3809209>

ACM Reference Format:

Siyang Jiang¹, Mu Yuan¹, Xiang Ji¹, Bufang Yang¹, Zeyu Liu², Lilin Xu³, Yang Li¹, Yuting He¹, Liran Dong¹, Wenrui Lu¹, and Zhenyu Yan¹, Xiaofan Jiang³, Wei Gao⁴, Hongkai Chen^{1,✉}, Guoliang Xing^{1,✉}. 2026. A Large-Scale Multimodal Dataset and Benchmarks for Human Activity Scene Understanding and Reasoning. In *The 24th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '26)*, June 21–25, 2026, Cambridge, United Kingdom. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3745756.3809209>

1 Introduction

Human action recognition (HAR) tasks have advanced significantly, leveraging artificial intelligence to classify human activities from multimodal sensory data [17, 32, 37, 60]. Beyond classification, Human Action Understanding (HAU) and Human Action Reasoning (HARn) tasks provide richer and more detailed descriptions of human activities, enabling diverse applications across various domains such as healthcare, daily living assistance, and surveillance [16, 23, 58]. For example, in the management of Alzheimer’s disease, a coherent understanding of a patient’s longitudinal behaviors is crucial for monitoring daily routines, providing timely caregiver support, and preventing accidents [39]. As shown in Fig. 1, a traditional HAR task is limited to recognizing isolated human actions, such as “sleep” or “fall”, and lacks the ability to interpret a continuous sequence of actions. In contrast, the HAU task addresses this limitation by understanding and providing natural language descriptions of the sequence of actions. For example, describing a scene as “the subject is eating, holding a knife in his right hand and a fork in his left, repeatedly picking up and putting down utensils” provides valuable context for the early detection of cognitive decline. Moreover, the HARn task infers intentions from sequences of human actions and predicts future actions. A typical example is that if the system detects “a subject is walking toward the sofa”, it may predict the subsequent action as “attempting to watch television or sit down”, thereby triggering a preventative intervention.

In practice, HAU and intention prediction can rarely be done by a straightforward autoregressive process with conventional deep neural networks (DNNs). Instead, it requires capability for knowledge representation and logical reasoning that integrate environmental contexts and scene knowledge [59]. To obtain these capabilities, existing techniques usually fine-tune Large Language Models (LLMs) using high-quality datasets annotated with `<data, caption>` pairs. Furthermore, logical reasoning in LLMs can be explicitly elicited through methodologies such as Chain-of-Thought [34], Tree-of-Thought [62], or Graph-of-Thought [45].

However, most existing HAR datasets provide only coarse-grained `<data, label>` pairs for RGB images [54], thus are unsuitable for

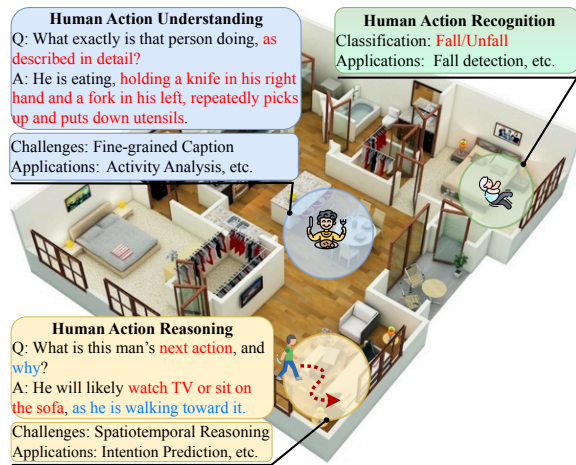


Figure 1: CUHK-X captures a multi-room home environment and supports three tasks: HAR (classification), HAU (captioning), and HARn (intention prediction). It integrates diverse modalities, including RGB, depth, thermal, infrared, IMU, skeleton, and mmWave.

fine-tuning LLMs for HAU and HARn tasks.¹ Although some recent datasets offer fine-grained $\langle \text{data}, \text{caption} \rangle$ pairs for RGB images [6, 19, 20], the fixed fields of view (FOVs) and limited mobility of RGB cameras hinder the timely capture of human behavior in many practical scenarios. Moreover, in privacy-sensitive scenarios such as daily home monitoring, RGB images pose risks by potentially containing sensitive personal data. Consequently, alternative sensor modalities, such as depth, thermal, IMU, and mmWave, are preferable.

Most current LLMs, particularly Large Vision Language Models (LVLMs), perform robustly with RGB and textual data, but they encounter significant difficulties when they are applied to other prevalent non-RGB modalities. The primary reason is the notable scarcity of large-scale datasets with $\langle \text{data}, \text{caption} \rangle$ pairs in non-RGB modalities [24], as most existing ones are confined to coarse-grained $\langle \text{data}, \text{label} \rangle$ annotations [31, 43, 65]. A naive approach to obtaining captioned datasets across multiple sensory modalities involves merging unimodal datasets, combining their coarse-grained labels, and using an LLM to generate captions. However, this method frequently yields captions that lack essential spatiotemporal consistency. For example, directly combining actions like brushing teeth and eating into a single scene is illogical, as these two actions typically occur independently in distinct contexts (e.g., a bathroom versus a dining room). Additionally, when generating captions, LLMs often fail to accurately infer human behavior contexts from the given coarse-grained labels, due to their limited representational capacity. This often results in incomplete, inaccurate, or even misleading captions (see §A.2 for details).

To address these gaps, we present CUHK-X, a large-scale multimodal dataset with seven synchronized modalities (RGB, depth, infrared (IR), thermal, skeleton, IMU, and mmWave) and three benchmarks for HAR, HAU, and HARn tasks. The CUHK-X dataset employs a ground-truth-first (GT-first) data collection scheme, where

¹We focus on two categories of $\langle \text{data}, \text{ground truth} \rangle$ data pairs: (1) $\langle \text{data}, \text{label} \rangle$, where the label is a discrete category, and (2) $\langle \text{data}, \text{caption} \rangle$, where the caption is a textual description.

target states are predefined before data recording, to ensure the acquisition of precisely aligned multimodal signals. To prevent spatiotemporal inconsistencies and guarantee accurate ground truth, the construction of CUHK-X begins with a Scene-based Caption Generation Framework. This framework categorizes human actions into seven thematic groups based on the American Time Use Survey (ATUS) [35, 63], from which 40 representative actions are carefully selected based on their frequency and relevance in prior benchmark datasets (e.g., HHAR [48], UCI [42], and Cosmo [37]). Subsequently, LLMs are used to logically connect these actions into semantically coherent captions that depict predefined scenes of daily living, such as living rooms, kitchens, bedrooms, and bathrooms. These captions incorporate varied contexts (i.e., performing actions in a relaxed, calm or hurried manner) to further enrich the narrative coherence. Lastly, we employ a human-verification stage to ensure that the generated captions are consistent with the ground truth by validating physical plausibility and temporal logic.

Using the generated captions as the ground truth, CUHK-X comprises over 64,267 daily activity samples from 30 participants, captured across two indoor settings using seven modalities, including RGB, depth, thermal, infrared, skeleton, IMU, and mmWave sensors. The sensor suite includes a Goermicro Vzense NYX 650 (RGB, depth, IR), Texas Instruments IWR6843ISK (mmWave radar), Hikvision TB4117 (thermal), and five WitMotion WT9011DCL-BT50 IMUs. Participants are instructed to understand and act out the generated captions, enabling the collection of high-quality, well-aligned data pairs.

To verify the dataset's utility, we provide benchmarks for six tasks spanning HAR, HAU, and HARn by evaluating state-of-the-art baselines of DNNs and LLMs for these benchmark tasks. These tasks include (1) HAR; (2-5) HAU tasks (caption comparison, context analysis, sequential action reordering, and action selection); and (6) HARn. The HAR benchmark validates the dataset's sufficient knowledge for recognition tasks. The HAU benchmarks assess caption comparison (against ground truth), context analysis (e.g., inferring speed of action), temporal ordering (for shuffled actions), and action identification (from a predefined set). The HARn benchmark evaluates an LLM's ability to infer intentions, causal relationships, and logical action progression. For HAR evaluation, we used state-of-the-art recognition models for each sensor modality, and analyze model performance under long-tail distributions and cross-subject situations. For HAU and HARn evaluation, we employed state-of-the-art baselines, including four captioning models (InternVL2.5-2B [11], InternVL2.5-8B [11], QwenVL2.5-3B [3], QwenVL2.5-7B [3]) and two reasoning models (VLLaVA-7B [29] and VR1Chat-7B [28]). The goal of these benchmarks is to explore the task performance and differentiability over different models and modalities (see §A.1 for overview illustration).

Experimental results demonstrate that fine-tuning models on CUHK-X significantly improves HAR accuracy compared to using pre-trained models alone, confirming that the dataset provides the necessary knowledge. Specifically, we achieve an average accuracy of 76.52% across seven modalities for HAR. Additionally, we achieve an average accuracy of 40.76% across all HAU tasks. Moreover, we achieve an average accuracy of 70.25% across three vision modalities for HARn. These results confirm that CUHK-X enables robust benchmarking and bridges the key gaps in existing datasets. The main contributions are summarized as follows:

- We introduce CUHK-X, a large-scale multimodal dataset comprising 64,267 samples collected from 30 participants using seven sensor modalities (RGB, depth, thermal, IR, IMU, mmWave, and skeleton) across two real-world environments. It provides diverse and realistic activity data and captions for advanced research in HAR, HAU, and HARn.
- To address challenges in logical consistency and spatiotemporal representation, we propose a prompt-based scene creation that leverages prompt-driven LLMs to generate logically connected actions in daily activity scenes with a human-verification stage.
- We establish three benchmarks containing six tasks to evaluate state-of-the-art baselines. Through rigorous analysis of task performance and differentiability over different models and modalities, we position CUHK-X as a cornerstone dataset for advancing research in HAR, HAU, and HARn. The project page and code are available at <https://openaiotlab.github.io/CUHK-X/> and <https://github.com/openaiotlab/CUHK-X>, respectively.

2 Motivation Study

This section outlines the tasks of HAU and HARn, followed by a discussion on the limitations of existing datasets in supporting these tasks.

2.1 Applications of HAU and HARn

The CUHK-X dataset can be applied to HAU and HARn tasks across various domains, including smart health [59], smart home [9], and disease intervention [36]. It enables continuous, longitudinal monitoring and analysis of user behavior, such as in Alzheimer’s disease monitoring [10] and Parkinson’s disease management [49]. Beyond healthcare, CUHK-X can also support smart home systems by enhancing comfort and energy efficiency. For example, systems built on CUHK-X can use the predicted user actions to adjust room lighting or thermostat settings, thereby optimizing energy consumption while maintaining a comfortable environment. This enriched holistic understanding, facilitated by CUHK-X, is critical not only for improved caregiving but also for creating smarter and more efficient living spaces.

2.2 Limitations of Existing Datasets

Existing multimodal datasets often suffer from key limitations, such as small subject pools (e.g., USC [65], Shoaib [46], HHAR [48]) and a restricted range of activities (e.g., UTD [8], mRI [1], Thermal-IM [51]). While datasets like NTU-RGB+D [31] and Ego-Exo4D [20] are more extensive, they frequently lack modality diversity (e.g., most miss IR, thermal, or captions), as summarized in Table 2.

2.2.1 Limitations of existing coarse-grained HAR datasets. As shown in Table 2, many earlier datasets, such as USC [65], Shoaib [46], and HHAR [48], are constrained by small participant numbers (fewer than 15) and a narrow range of activities (e.g., only 6–12 actions). Similarly, datasets like Thermal-IM [51] and UTD [8] involve too few participants or a limited number of activities. Some recent datasets, such as NTU-60/120 [31, 43], and PKU-MMD [12], include larger participant cohorts (e.g., 66 people in PKU-MMD) and more activity classes (e.g., 60 actions in NTU-60). However, they primarily focus on RGB

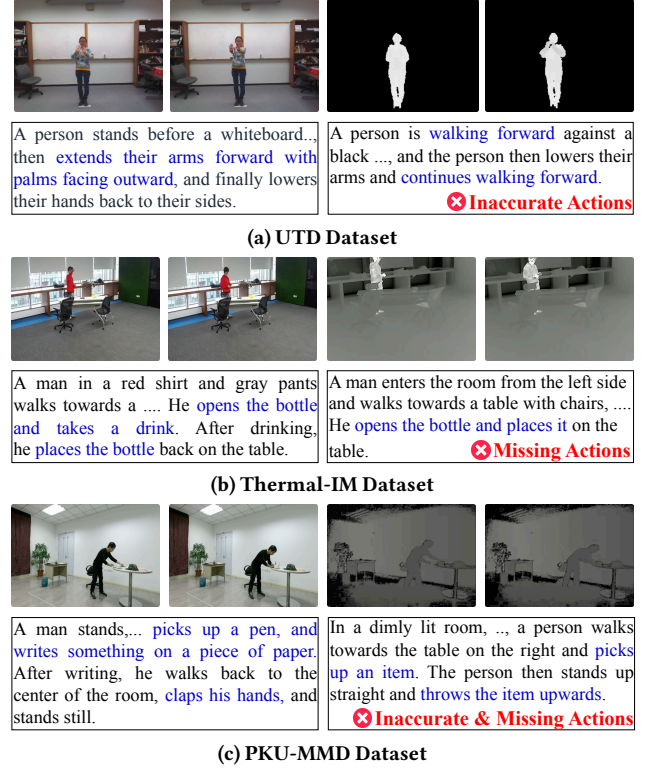


Figure 2: Limitations of SOTA LVLMs in HAU tasks.

and skeleton data, overlooking essential modalities like thermal, IR, and IMUs. For instance, action recognition relying solely on RGB data becomes challenging under occlusion or when a person faces away from the camera. Thus, a major limitation of coarse-grained datasets is their inability to provide sufficient detail for HAU or HARn tasks.

2.2.2 Limitations of existing fine-grained HAU datasets. Previous fine-grained datasets, such as Ego-4D [19] and Ego-Exo4D [20], enhance model capabilities for detailed human action understanding by providing rich data descriptions. Nonetheless, these datasets are limited in modality coverage. As shown in Table 2, Ego-4D and Ego-Exo4D include RGB data but lack other critical modalities, such as depth, thermal, infrared, and skeleton. Moreover, state-of-the-art captioning models, such as Tarsier [55] and Tarsier2 [64], are designed specifically for RGB data, leading to suboptimal performance in other modalities. To illustrate, as shown in Fig. 2, we conducted experiments on UTD [8], Thermal-IM [51], and PKU-MMD [12] datasets using Tarsier2 (see §A.2 for details).

2.3 Summary

In summary, existing coarse-grained datasets are unsuitable for HAU and HARn tasks due to insufficient descriptive detail, while existing fine-grained datasets fail to comprehensively cover multiple modalities. To address these gaps, CUHK-X provides a wider variety of modalities (e.g., RGB, depth, thermal, IR, skeleton, IMU, and mmWave), a more diverse range of activities, and a larger participant cohort. Following NTU-RGBD [31] and MM-Fi [61], CUHK-X treats skeleton as a modality, though the sequences are estimated from RGB and depth frames via MMPose rather than captured by

dedicated motion-capture hardware such as Vicon or OptiTrack. This comprehensive dataset not only supports HAR tasks but also enables fine-grained HAU tasks that demand a deeper contextual understanding, as well as reasoning-based tasks like HARn, which require understanding sequential actions and predicting future behaviors. Most importantly, it establishes a foundational resource for developing robust multimodal systems for sophisticated human behavior analysis in real-world scenarios.

3 Ground Truth-First Data Collection

This section describes our ground truth-first (GT-first) methodology for human action data collection, positioning it as a practically better alternative compared to conventional data-first approaches to human data collection.

3.1 Motivation for a GT-First Strategy

Data collection and annotation typically follow two paradigms: data-first [26] and ground-truth-first, or GT-first [1, 6, 31, 37, 61]. The *data-first* strategy, as the most intuitive method of data collection, first gathers large-scale real-world data from human subjects, and then applies annotations to the collected data afterwards. While conceptually reasonable, these strategies present significant limitations in practice. In particular, since data collection is conducted before annotations and hence lacks specific guidelines, the collected data requires intensive postprocessing and filtering to ensure efficient and proper annotation. These extra efforts make data collection highly ineffective and not scalable, especially when involving large groups of human subjects. In addition, since annotations have to be applied to collected data, it is expensive to fix any early mistakes in the human action set, raising extra privacy and consent risks when recording people and homes.

To avoid these limitations, most existing large-scale datasets, such as ImageNet [15] and NTU-RGBD [31], follow the GT-first strategy for data collection. Being different from the data-first strategy, the GT-first strategy defines the label space, annotation rules, and target scenes in advance, and then collects data based on those definitions. Prioritizing ground truth yields clear benefits: 1) *focus*, as only planned cases are recorded; 2) *efficiency*, as annotators verify checklists rather than create labels from scratch; 3) *scalability*, as the same collection scripts and annotation rules can be reused across sites.

However, the approach would introduce bias compared to the data-first approach. In the following, our objective is to analyze and mitigate three inherent biases of GT-first approaches in CUHK-X, i.e., *coverage*, *diversity*, and *discrepancy*, through several targeted strategies including: (1) refining class coverage within defined boundaries, (2) enhancing variation via intra- and inter-class combinations coupled with linguistic enrichment, and (3) incorporating human-in-the-loop verification to ensure physical and logical coherence. These methodological components are elaborated in §4.

3.2 Bias in GT-First Data Collection

GT-first approaches, i.e., those that derive supervision or text directly from predefined actions and associated templates, offer clear benefits in controllability and reproducibility to obtain the GT. We therefore adopt this strategy in CUHK-X. However, it also induces systematic artifacts that reflect the prior encoded by the ontology and the

Table 1: Evaluation on discrepancy.

BertScore-F1	Com-Cap.	Free-Act.	Instruct-Act.
Task-1	93.40%	95.50%	93.60%
Task-2	92.90%	94.10%	93.10%
Task-3	93.50%	94.20%	92.70%
Task-4	92.20%	93.80%	92.50%

templating process. In particular, they may bias models toward (i) *Coverage*, (ii) *Diversity*, and (iii) *Discrepancy*. In the following, we justify the biases in CUHK-X.

3.2.1 Coverage. We define coverage as the number of classes included in the dataset. It is hard for any dataset to include all human actions. In practice, some labels mean more than one thing, and they are not consistent across datasets. Consequently, in CUHK-X, we adopt a closed-world objective, i.e., we focus on a fine-grained subset of actions curated from prior research and informed by our experimental evidence. In particular, we coarse-grained our action selection using ATUS [35] and the action frequency across datasets (§4.1.1), and then fine-grained the selection of several significant actions based on prior studies (§4.1.2).

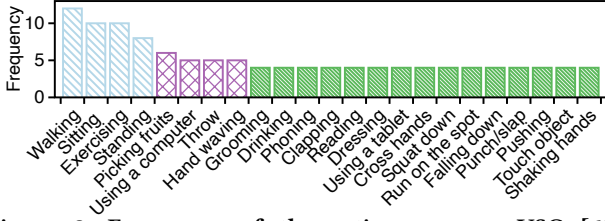
3.2.2 Diversity. Even when labels are well covered, their realizations can exhibit narrow lexicalization, limited contextual variety, and stereotyped co-occurrence patterns. We therefore define diversity, for each modality, as the extent to which samples within the same class differ from one another. For example, in each modality, instances of the class “drinking water” should vary in setting, poses, and previous or co-occurring actions. To mitigate this limitation, we first generate captions from diverse intra- and inter-class action combinations (§4.2.1) and then enrich those captions with linguistic variation to diversify lexical choice, syntax, and contextual framing (§4.2.2).

3.2.3 Discrepancy. In CUHK-X, discrepancy is denoted from two perspectives. Firstly, it is the mismatch between captions generated by an LLM and those authored by humans for the same meaning. To minimize this discrepancy, we incorporate a human-verification stage that enforces physical plausibility, logical coherence, and contextual appropriateness (§4.2.3), aiming to bring LLM captions close to human parity. Secondly, we consider whether the data reflect key actions performed naturally versus performed by following explicit instructions. Because CUHK-X focuses on daily activities, we regard any bias introduced by instruction-following as acceptable. In practice, we asked the volunteers to read the instructions first, then perform the action based on their understanding. We also conduct an experiment to assess this effect.

Specifically, we set a home office setting with four micro-tasks: drinking water (Task-1), listening to music with earphones (Task-2), writing notes (Task-3), and answering a phone call (Task-4). For each task, we recruited three volunteers (10% of the participants in CUHK-X) to act freely and annotate their recordings. Next, the same volunteers performed the actions following instructions derived from their free actions. We again collected three independent captions. To make the distinction concrete, we present a paired example from Task-1 (drinking water). In the **free-action** condition, the volunteer receives only the high-level prompt “Please drink some water.” In the **instructed** condition, the same volunteer receives a step-by-step prompt derived from their own free performance: “Walk to the desk. Use your right hand to pick up the white cup. Raise the cup to

Table 2: A summary of the related coarse-grained HAR and fine-grained HAU datasets (● indicates inclusion). DEP, THE, IR, SKE denotes depth, thermal, infrared, skeleton modalities, respectively.

Dataset	Years	# of Samples	# of Subjects	# of Activities	RGB	DEP	THE	IR	SKE	IMU	mmWave	Text Captions
USC [65]	2012	840	14	12	○	○	○	○	○	●	○	○
Shoaib [46]	2014	70	10	7	○	○	○	○	○	●	○	○
HHAR [48]	2015	3,240	9	6	○	○	○	○	○	●	○	○
UTD [8]	2015	3,444	8	27	●	●	○	○	●	●	○	○
UCI [42]	2016	180	30	6	○	○	○	○	○	●	○	○
NTU-60 [43]	2016	56,880	40	60	●	●	○	●	●	○	○	○
PKU-MMD [12]	2017	20,000	66	51	●	●	○	●	●	○	○	○
NTU-120 [31]	2019	114,480	106	120	●	●	○	●	●	○	○	○
Cosmo [37]	2022	3,434	30	14	○	●	○	○	○	●	●	○
mRI [1]	2022	300	20	12	●	●	○	○	○	●	●	○
Thermal-IM [52]	2023	783	2	24	●	○	●	○	○	○	○	○
MM-Fi [61]	2023	1080	40	27	●	●	○	○	●	○	●	○
XRF55 [54]	2024	42,900	39	55	●	●	○	●	○	○	●	○
ActivityNet [6]	2015	9,682	-	203	●	○	○	○	○	○	○	●
Ego-4D [19]	2022	5,831	923	146	●	○	○	○	○	●	○	●
Ego-Exo4D [20]	2024	5,035	740	689	●	○	○	○	○	●	○	●
CUHK-X	2025	64,267	30	40	●	●	●	●	●	●	●	●

**Figure 3: Frequency of the actions among USC [65], Shoaib [46], HHAR [48], UTD [8], ActivityNet [63], UCI [42], NTU [31, 43], PKU-MMD [12], Cosmo [37], mRI [1], Thermal-IM [51].**

your mouth and take a sip. Put the cup back on the desk.” Note that the instruction does not inject semantically alien content. We compared the semantic similarity of captions using closed-source LLMs, i.e., QwenVL3-235B, between the “free” and “instruct” conditions using BertScore. Across all four tasks, the two sets of captions are highly similar, indicating that both conditions effectively capture the underlying action information (Table 1, Com-Cap.). Furthermore, when compared the generated captions against human-annotated ground truth, the results are also accurate, as shown in Table 1 (Free-Act. and Instruct-Act.). These findings justify our focus on human actions, and the remaining discrepancies are acceptable within CUHK-X.

4 Scene-based Caption Generation

4.1 Prior knowledge-based Action Selection

CUHK-X is developed through a labor-intensive collection of multimodal data in real-world scenarios. In this section, we describe how CUHK-X incorporates typical daily actions with a two-stage selection process to select representative actions.

4.1.1 Coarse-grained Action Selections via Predefined Categories and Cross-dataset Frequency. Firstly, based on the ATUS [35] activity hierarchy and ActivityNet [6], we categorized the activity classes into seven top-level categories: Personal Care, Eating and

Drinking, Household, Caring and Helping, Working, Socializing and Leisure, and Sports and Exercise. Note that while the ATUS activity classification system was originally developed over a decade ago, it remains actively maintained with annual data releases through 2025, ensuring contemporary relevance for activity taxonomy [5]. In addition, ActivityNet is used as a data source to combine multiple data sources to address coverage gaps [66]. CUHK-X adopts a structured semantic framework that leverages hierarchical relationships between activities, ensuring the selection of typical and comprehensive real-world daily activities. Next, we analyzed the action frequency in 12 popular human action recognition datasets, such as NTU [31, 43], UTD [8], and UCI [42], which are primarily summarized in Table 2. Unlike prior datasets with label-level annotations, CUHK-X is the first to combine seven modalities (RGB, depth, thermal, infrared, skeleton, mmWave, and IMU) with rich caption-level textual descriptions, enabling both coarse-grained recognition and fine-grained semantic understanding of human activities. As illustrated in Fig. 3, we analyze the high-frequency occurrences within these datasets. Specifically, the total number of action classes is 349, which are consolidated into 127 classes by merging those with similar meanings. However, since most actions appear only once, we focus exclusively on high-frequency actions (#frequency > 4) for further analysis.

4.1.2 Fine-grained Action Selections via Prior Studies. Then, we carefully selected fine-grained, representative actions based on insights from previous research. Specifically, the *Personal Care* category (6 actions) was guided by the prior study [33]. The *Eating and Drinking* (6 actions) and *Household* (5 actions) categories were guided by findings from the previous work [41]. The *Working* category (6 actions) was inspired by [2], while the *Socializing and Leisure* category (5 actions) was shaped by [4]. Finally, the *Sports and Exercises* (9 actions) and *Caring and Helping* (3 actions) categories were supported by insights from Gerber et al. [18].

We select 40 actions divided into seven categories as follows: **(1) Personal Care**, which has 6 actions including Washing face, Brushing teeth, Combing hair, Undressing, Wiping hands, and Getting Dressed; **(2) Eating and Drinking**, which has 6 actions including Drinking, Eating, Grabbing utensils, Pouring, Stirring, and Peeling fruit; **(3) Household**, with 5 actions including Sweeping, Mopping, Washing dishes, Wiping surface, and Folding clothes; **(4) Working**, which includes 6 actions including Typing on a keyboard, Writing, Calling, Checking the time, Reading, and Turning a page; **(5) Socializing and Leisure**, with 5 actions including Taking a selfie, Playing board games, Watching TV, Using a phone, and Listening to music with headphones; **(6) Sports and Exercises**, which has 9 actions including Walking, Lunges, Sitting down, Lying down, Standing up, Stretching, Jumping jacks, Squats, and Running; and **(7) Caring and Helping**, with 3 actions including Taking medicine, Checking body temperature, and Massaging oneself.

4.2 Prompt-based Scene Creation with Human-driven Checking

In this subsection, we propose a prompt-based scene creation approach designed to logically connect the selected actions (§4.1) via constructing various daily living scenes.

4.2.1 Intra- and Inter-categories Caption Generation. Our goal is to connect as many selected actions as possible into a coherent and logical sentence that aligns with everyday life scenarios. To achieve this, we implement a two-stage prompt design primarily based on the selected actions, ensuring both diversity and relevance in the generated captions. Specifically, we design the prompt to encourage the LLM to combine multiple actions into a cohesive scene within each category. For instance, actions such as “Washing face,” “Brushing teeth,” “Combing hair,” “Dressing,” and “Wiping hands” can be combined to generate a detailed scene: *The user wakes up, opens the curtains, and stretches (Stretching). The user walks to the bathroom and washes their face (Washing face) with water or facial cleanser, then dries it with a towel. The user picks up a toothbrush, squeezes toothpaste, and begins brushing their teeth (Brushing teeth). After brushing, they rinse their mouth with water and clean the toothbrush. The user uses a comb to carefully brush their hair (Brushing hair), possibly tying it up or styling it. The user quickly wipes their hands (Wiping hands) with a towel or tissue. Finally, the user returns to the bedroom, selects clothes from the wardrobe, and completes the process of getting dressed (Dressing).* Similarly, actions from other categories are combined via LLMs to create contextually rich captions that reflect realistic and meaningful daily scenarios. This method ensures that the generated captions not only integrate multiple actions logically but also create a natural flow of events that mirrors real-life activity patterns.

4.2.2 Enriching Captions through Language Diversity. To further enhance the diversity of captions, we leverage LLMs to enrich them by expanding or substituting their sentence components. Specifically, a sentence is composed of several key elements, including the subject, predicate, object, attribute, adverbial, and complement. In the context of central actions in human action understanding, the subject, predicate, and object are typically predefined. Thus, we use LLMs such as GPT-4o [22] or DeepSeek [30], to add more attributes and adverbials. For instance, we can enrich the description of the

example in §4.2.1 by incorporating adverbs before the verb to provide additional context and nuance. Specifically, instead of a straightforward description, the user can *carefully squeeze a small amount of toothpaste onto the bristles and begin brushing their teeth (Brushing teeth)*, with added detail such as “quickly,” “smoothly,” or “slowly” to describe how the action is performed. By enriching the attributes and adverbials, the generated captions provide a more detailed and vivid depiction of actions, creating a natural flow between individual activities. This level of detail not only enhances the linguistic diversity of captions but also improves their utility in datasets for tasks such as human action understanding and multimodal learning.

4.2.3 Human Verification via Physical Knowledge and Logical Coherence. To reduce discrepancy and hallucination in LLM-generated captions, we implement a human-verification stage that enforces physical plausibility, scene logic, and dataset conventions before acceptance as ground truth. Four graduate-level raters conduct this review using a structured checklist and edit protocol.

We validate captions against the following criteria: (1) *Physical feasibility and kinematics.* Body-pose and object-state transitions must obey continuity. For example, “cup is empty” → “pouring into cup” → “cup becomes full,” not the reverse. (2) *Scene and environment consistency.* Actions and objects must be compatible with the room and floor plan (e.g., “brushing teeth” in a bathroom; “watching TV” requires a visible or plausibly placed TV). Captions must not assert observations outside a sensor’s FOV. (3) *Temporal and causal coherence.* Event order must be logically progressive (e.g., “grabbing utensil” precedes “eating”). (4) *Affordance and commonsense constraints.* Interactions must respect object affordances (e.g., “stirring with a fork/spoon,” not “stirring with a phone”). A caption may cover multiple actions, but each action span must be temporally localizable. Note that in CUHK-X, captions prioritize action/scene semantics over appearance details that are modality-incongruent. This human verification serves as a reliability gate, yielding captions that are (i) physically plausible within the recorded environments, (ii) temporally and causally coherent, and (iii) aligned with the action ontology.

5 Data Collection Setup

5.1 Hardware Setup

5.1.1 Ambient sensors setup. As shown in Fig. 5a, firstly, we use a Goermicro Vzense NYX 650 camera to capture RGB, depth, and IR data. Specifically, the Vzense NYX 650 cameras offer a 70° horizontal and 50° vertical field of view, operating at a frame rate of 10 frames per second. Leveraging 940nm infrared light, these cameras are well-suited for both indoor and outdoor environments, even under low-light or no-light conditions. Next, we use a Texas Instruments IWR6843ISK mmWave radar operating in the 60–64 GHz band. We configure it with a 20 fps frame rate, 0.044 m range resolution, a 5.03 m maximum unambiguous range, a 1.0 m/s maximum radial velocity, and a 0.13 m/s radial velocity resolution. This sensor excels in detecting objects, measuring distances, and tracking motion with high precision. In addition, we use a Hikvision TB4117 thermal imaging camera for precise temperature measurement. Featuring a 120×160-pixel resolution and compact 70×46×22.75 mm dimensions, this device measures temperatures from 30°C to 45°C with 25 fps, making it ideal for thermal monitoring. Lastly, we use a TSRV-Q9

AI Tracking Gimbal, a compact ($60 \times 70 \times 185$ mm) and lightweight (220 g) device designed for precise automatic tracking and stabilization. Powered by a 3.7V/1200mAh battery, it supports 3.5 hours of continuous tracking. Compatible with devices up to 12 mm thick, it features 360° horizontal rotation and 180° manual vertical adjustment, making it ideal for dynamic content creation. In practice, we fix the sensor's angle and position during data collection.

5.1.2 Wearable sensors setup. We use the Bluetooth 5.0-enabled WitMotion WT9011DCL-BT50 as our Inertial Measurement Unit (IMU) for precise tracking of acceleration ($\pm 16g$), angular velocity ($\pm 2000^\circ/s$), and magnetic field (± 2 Gauss). It supports output frequencies ranging from 0.2 Hz to 200 Hz and provides angular measurements of up to $\pm 180^\circ$ (X/Z) and $\pm 90^\circ$ (Y). Powered by a 130 mAh battery, it delivers up to 40 hours of continuous operation with a maximum transmission range of 50 m in open space with 10 samples per second. Measuring $32.5 \times 23.5 \times 11.6$ mm in size and weighing just 9 g, each participant was equipped with 5 of these devices, with sensors placed on the wrists, ankles, and waist using adjustable bands, shown in Fig. 5b.

5.2 Environmental Setup

We collected data from two indoor environments, with a focus on four common room settings: the living room, kitchen, bedroom, and bathroom. These rooms were selected to represent a diverse range of daily activities for studying and analyzing human behavior in realistic, everyday scenarios. To ensure systematic data collection, we documented all sensor locations in each room by marking their positions and taking photographic records. This meticulous annotation method provides thorough coverage of typical activities occurring within these spaces, facilitating robust data analysis. The floor plans, as depicted in Fig. 4, offer detailed spatial representations of the two indoor environments, with icons highlighting the exact locations of the sensors deployed for data collection. Specifically, ambient sensors were strategically positioned to optimize coverage and data reliability. Furthermore, each photo depicts a room and serves to visually contextualize the sensor data, providing a visual context to the collected data. Our environmental setup not only enables fine-grained monitoring of human activities but also supports the integration and analysis of data across multiple modalities.

5.3 Put All Things Together

5.3.1 Demographic Characteristics of Participants. We recruited 30 participants (40% male, 60% female) with an age range of 20–23 years. BMI ranged from 16.41 to 29.02, with a mean of 24.54. Additionally, we collected data on participants' activity habits, indicating an average session duration of approximately 22.67 minutes and an average exercise frequency of 1.7 times per week, where low, moderate, and high intensities are assigned scores of 1, 2, and 3, respectively. These metrics suggest a relatively balanced distribution of height and weight among participants and highlight their tendency toward low-frequency, short-duration exercise routines. This dataset provides a meaningful baseline for the development and evaluation of computational models or systems aimed at activity recognition and health monitoring, ensuring both generalizability and reliability in human-centric data.

5.3.2 Data Synchronization and Annotation. To ensure precise alignment across all modalities, we adopt the global time from the host computer as the reference for synchronization. We use a marker, i.e., a director's board, to define the start and end points of the alignment process, enabling consistent temporal boundaries for all recorded data. RGB data serves as the primary modality for synchronization due to its high temporal resolution and consistency. Radar and IMU data are recorded with timestamps rigorously aligned to the global time, ensuring that all data streams are temporally synchronized to a high degree of accuracy.

For caption data annotation, captions are pre-generated (refer to §4 for more details), and subsequently used during data collection. This process ensures that the descriptions of each video segment are naturally aligned with the corresponding actions, minimizing annotation errors. In addition to caption-level alignment, individual action annotations are performed with meticulous care. Each video segment is manually labeled and segmented on a frame-by-frame basis to achieve the highest possible precision. Special attention is given to segment transitions and ambiguous actions to avoid misalignment or mislabeling, which can significantly impact downstream tasks. This manual process provides an accurate foundation for training and evaluating computational models. We provide more details of data statistics and data visualizations in §B.1 and §B.2, respectively.

6 Experimental Setup

Here, we describe our tasks, baseline, metrics, and implementation details. Note that in this paper, LLMs are used for generality without distinguishing between modalities.

6.1 Tasks Descriptions

6.1.1 HAR Task. HAR is a task focused on identifying and classifying human activities such as walking, running, sitting, and standing from sensor data. We define 40 classes across seven categories for recognition as our HAR tasks.

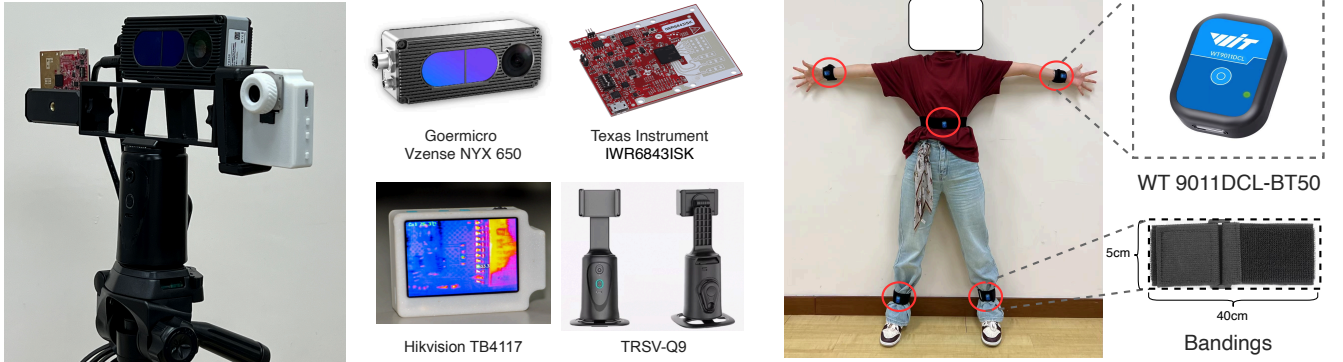
6.1.2 HAU Tasks. HAU goes beyond basic HAR by capturing richer semantic information. Unlike HAR, which focuses on predefined actions, HAU seeks to understand the context of action sequences, including spatiotemporal semantics, relationships between actions, their order, and interactions with objects or the environment. In particular, we define four sub-tasks in HAU as follows:

- **Caption Comparison:** This sub-task involves comparing the captions generated by the LLMs with the ground truth captions to evaluate the LLMs' capability for accurate description generation.
- **Context Analysis:** This task requires that the LLMs identify the correct context exhibited by the participants. In particular, we hope LLMs can recognize when actions are performed in a relaxed, calm, or hurried manner.
- **Sequential Action Reordering:** The model observes data containing actions in a shuffled order and accurately reorders them into the correct sequence.
- **Action Selection:** The LLMs observe the data to select the correct actions from a predefined pool of 40 actions.

6.1.3 HARn Task. HARn goes beyond HAU's semantic understanding by adding reasoning capabilities to infer intentions, causal relationships, and logical action sequences, which involves predicting



Figure 4: Environment Visualization (The left room is Room 1 and the right one is Room 2). Layout with room-wise visual annotations (Bedroom, Kitchen, Bathroom, and Living Room) showing corresponding example images and sensor placements. The icon indicates the location of the ambient sensor.



(a) Ambient sensors include the Vzense NYX 650 for image sensing, the Texas Instruments IWR6843ISK for radar sensing, and the Hikvision TB4117 for thermal imaging. TRSV-Q9 is an AI tracking gimbal.

(b) Wearable sensor includes 5 IMU sensor placements on the wrists, ankles, and waist using bandings, and the WitMotion WT9011DCL-BT50 IMU module.

Figure 5: Photos of our ambient and wearable sensor hardware.

outcomes. Specifically, the model must predict the next action from a provided list based on a series of preceding actions.

6.2 Baseline and Metrics

6.2.1 HAR Task. We use ResNet-50 [21] for its visual recognition effectiveness. Radar data is processed with PointNet [38], enhanced by feature engineering to capture spatial characteristics. Skeleton data employs MotionBERT [67], using a dual-stream transformer and a multilayer perceptron, with 17 3D joints extracted via Human3.6M-compliant pose estimation models [13]. IMU data is handled by a 1D-CNN [50] with three convolutional layers and transformer encoder layers [53] followed by a linear classification head. HAR task evaluation uses Accuracy, Precision, Recall, and F1-score.

6.2.2 HAU and HARn Tasks. We evaluate these two tasks via the latest video LLMs and visual reasoning LLMs, including InternVL2.5-2B

(InternVL-2B) [11], InternVL2.5-8B (InternVL-8B) [11], QwenVL2.5-3B/7B (QwenVL-3B/7B) [3], VideoLLaVA-7B (VLLaVA-7B) [29], and VR1Chat-7B [28]. For the caption comparison task in HAU, we use metrics including BERTScore (F1-Score), ROUGE-1, ROUGE-L, BLEU-1, and METEOR scores. Additionally, we use accuracy to evaluate the context analysis, sequential action reordering, action selection, and HARn task.

6.3 Implementation Details

For both training schemes, we set the learning rate to 0.001, use a batch size of 64, and update parameters with the Adam [25] optimizer. For HAR tasks, we randomly split 80% as the training set and 20% as the testing set in all modalities. For RGB, IR, thermal, and depth modalities, models are initialized with ResNet-50 pre-trained weights, while for IMU and radar modalities, baseline models are

initialized using Kaiming initialization [21] since no general pre-trained models are available for these modalities. Besides, for the skeleton modality, we initialize MotionBERT [67] as our backbone with weights pre-trained on the NTU RGB+D dataset [43].

We observed that directly using this raw data fails to effectively capture the spatial information of depth since the raw depth data is 16-bit. To address this, we process the raw depth data via the two types of house floor plans, shown in Figure 4. In particular, we filtered depth values outside a defined range, replacing them with zero to focus on relevant depth information for each specific environment. In Room 1 of Fig. 4, the depth ranges are set as follows: Living Room [500, 5000], Kitchen [500, 3300], Bedroom [500, 3200], and Bathroom [500, 2800]. In Room 2 of Fig. 4, the depth ranges are slightly different: Living Room [500, 4700], Kitchen [500, 3260], Bedroom [500, 3500], and Bathroom [500, 2000]. These ranges are tailored to accommodate the spatial characteristics of each scenario and room.

All video LLMs are evaluated under a zero-shot paradigm using their default configurations and task-specific system prompts. We directly use the whole video clip as our input. In HAU tasks, models receive different prompts: (1) captioning-“Describe what the person in the video is doing. You can briefly mention the background or setting, but focus mainly on understanding the person’s actions.”; (2) context analysis-“What emotion does the person experience while performing the activities?”; (3) sequential action reordering-“Please sort the following activity lists in chronological order based on the video content.”; (4) action selection-“What activity is the person performing in the video? You must choose only from the following activities: {Class Set}. You can choose multiple activities if necessary.”; (5) HARn-“What activity is the person likely to do next?”.

7 Benchmarks

We present three benchmarks in CUHK-X: HAR, HAU, and HARn. For HAU and HARn, we report results for vision only. The main reason is that strong and widely used LLMs exist for RGB, for example QwenVL, while comparable models for sensors such as IMU or mmWave are not yet available. Note that a direct comparison across these very different modalities would not be fair. Our results are meant to show that each modality contains useful task relevant information, not to rank the modalities or claim that one is the best.

7.1 Benchmark of HAR

To verify that the CUHK-X dataset contains sufficiently new knowledge for the HAR tasks in different data modalities, we provide a benchmark for the HAR task. The remaining content is structured around addressing the following two questions: (1) *Does the dataset contain valuable knowledge?* (2) *What are the challenges in this task?*

As shown in Table 3, performance varies by modality, with an overall accuracy of 76.52%. With standard supervised training, vision-based inputs work best. Thermal gives the top results, with accuracy 92.57% and F1 score 93.36%. Depth and RGB follow, with F1 scores of 90.93% and 91.28%. IR is lower than the other visual channels but still strong, with F1 90.46%. Skeleton features also perform well, with accuracy 79.08% and F1 84.17%. In contrast, other sensors are much lower: mmWave reaches 46.63% accuracy and 44.53% F1, and IMU reaches 45.52% accuracy and 38.32% F1. These gaps likely come from the lower spatial resolution and signal-to-noise ratio of mmWave,

Table 3: Overall Performance of HAR Task in CUHK-X.

Modality	Accuracy	Precision	Recall	F1-Score
RGB	90.89%	92.24%	91.02%	91.28%
Depth	90.46%	91.76%	90.75%	90.93%
IR	90.22%	91.53%	89.94%	90.46%
Thermal	92.57%	93.54%	93.50%	93.36%
mmWave	46.63%	48.29%	46.63%	44.53%
IMU	45.52%	40.84%	38.00%	38.32%
Skeleton	79.08%	91.46%	79.08%	84.17%

the sensitivity of IMU to placement and orientation, and larger shifts across trials for these sensors. Overall, the results show that the dataset provides useful information for HAR. We observe that certain modalities exhibit markedly different standalone accuracy: IMU and mmWave individually fall below 50%, whereas RGB achieves substantially higher performance. We attribute this disparity to modality-specific information ceilings, i.e., IMU and mmWave inherently encode less discriminative information for activity recognition than vision-based modalities such as RGB and Depth.

7.2 Benchmark of HAU

In HAU benchmark, our goal is to benchmark the task performance of different models and different modalities of the following four sub-tasks.

7.2.1 Results of Caption Comparison. As shown in Table 4, different models excel on different metrics and modalities. VLLaVA-7B achieves the best BERTScore F1 and the highest ROUGE-1/ROUGE-L on the depth and thermal modalities. QwenVL-3B attains the top ROUGE and BLEU-1 scores on RGB and IR and remains competitive elsewhere. VR1Chat-7B frequently yields the best BLEU-1 and METEOR (e.g., RGB and IR), indicating strong fluency. In contrast, InternVL-2B/8B obtain decent BERTScore F1 but lag behind on ROUGE and BLEU. Overall, higher-capacity models (e.g., 7B) tend to outperform the InternVL-2B/8B baselines, although the 3B Qwen model is a notable exception that matches or surpasses some 7B models on several metrics. We notice that BERTScore and n-gram metrics show a notable discrepancy since CUHK-X encourages diverse paraphrasing. Semantically equivalent captions often differ lexically, which n-gram metrics penalize. More abstractive models amplify this gap, scoring lower on ROUGE/BLEU despite preserving meaning.

7.2.2 Results of Context Analysis. As shown in Fig. 6a, the average accuracy is 50.52%. In particular, VLLaVA-7B delivers the best overall context accuracy across modalities, leading on RGB, IR, and Depth. QwenVL-3B follows closely and remains competitive on Thermal. InternVL-2B and InternVL-8B stay at the lower end, with accuracies roughly between 24 and 35 percent across modalities, suggesting that although these architectures work well for general captioning, they are weaker at modeling contextual cues. VR1Chat-7B achieves mid-range results, typically around 42 to 50 percent, with slightly better scores on RGB and Depth than on IR and Thermal. The spread between models is largest on IR and Depth, where the best systems approach 80 percent while the weakest remain below 30 percent, underscoring the challenge of interpreting non-RGB signals.

7.2.3 Results of Sequential Action Reordering. As shown in Fig. 6b, the average accuracy is 47.24% and no single model dominates across

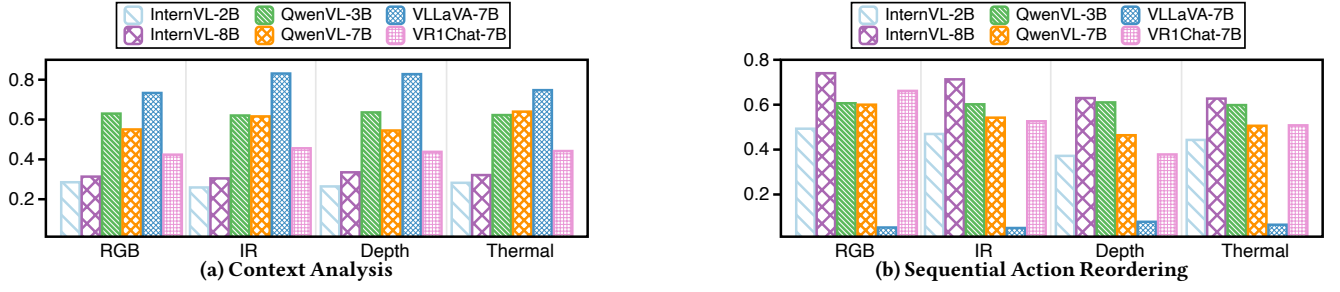


Figure 6: Accuracy of context analysis and sequential action reordering in CUHK-X.

Table 4: Results of caption comparison. We report BERTScore F1 (B.-F1), ROUGE-1 (R.-1), ROUGE-L (R.-L), BLEU-1 (B.-1), and METEOR (MET.).

Model	B.-F1	R.-1	R.-L	B.-1	MET.
RGB					
InternVL-2B	84.39%	4.33%	3.64%	0.61%	3.97%
InternVL-8B	84.07%	3.04%	2.53%	0.72%	3.63%
QwenVL-3B	86.22%	18.40%	13.80%	21.46%	19.89%
QwenVL-7B	85.47%	14.79%	12.05%	18.04%	22.21%
VLLaVA-7B	86.40%	16.12%	12.77%	12.86%	11.58%
VR1Chat-7B	86.24%	17.42%	13.66%	21.62%	23.18%
Depth					
InternVL-2B	84.09%	4.63%	3.95%	0.53%	3.67%
InternVL-8B	83.95%	2.89%	2.35%	0.73%	3.52%
QwenVL-3B	85.03%	15.00%	10.76%	18.58%	16.89%
QwenVL-7B	84.55%	12.70%	10.65%	16.64%	18.98%
VLLaVA-7B	85.94%	16.31%	13.53%	12.07%	10.37%
VR1Chat-7B	84.69%	14.17%	11.49%	17.73%	19.39%
IR					
InternVL-2B	84.24%	4.38%	3.68%	0.58%	4.09%
InternVL-8B	84.22%	3.05%	2.54%	0.64%	3.68%
QwenVL-3B	86.49%	18.56%	13.99%	22.17%	18.85%
QwenVL-7B	85.38%	14.71%	11.93%	18.25%	21.50%
VLLaVA-7B	86.25%	15.64%	12.81%	12.25%	10.78%
VR1Chat-7B	86.03%	16.51%	12.96%	21.23%	22.05%
Thermal					
InternVL-2B	84.30%	5.05%	4.28%	0.38%	3.93%
InternVL-8B	83.85%	2.74%	2.24%	0.73%	3.37%
QwenVL-3B	85.04%	14.78%	11.40%	18.13%	16.95%
QwenVL-7B	84.48%	12.44%	10.31%	15.41%	20.24%
VLLaVA-7B	85.85%	16.24%	13.32%	11.31%	10.16%
VR1Chat-7B	84.94%	14.23%	11.51%	17.90%	19.86%

all modalities. We notice that InternVL-8B achieves the best accuracy on RGB and IR, reaching roughly three quarters in both cases. QwenVL-3B edges out the others on Depth, with scores in the mid-sixties. VR1Chat-7B is strongest on Thermal, around sixty percent. QwenVL-7B is steady across modalities but does not lead any of them. VLLaVA-7B performs poorly, near chance on every modality, while InternVL-2B also sits toward the lower end. Across modalities, IR and RGB are generally the most informative for this task, Depth is in the middle, and Thermal is the most challenging. These results suggest that architecture–modality compatibility matters: capacity helps in places, but the top results come from different models on different sensors.

7.2.4 Results of Action Selection. As shown in Fig. 7, the average accuracy is 24.54%. We notice that QwenVL-7B generally attains the best or near-best scores across all three metrics and modalities, indicating strong action-selection capability, VLLaVA-7B is consistently competitive, with notable strength on infrared and thermal inputs. InternVL-8B delivers mid-range results, with its strongest performance on RGB, while InternVL-2B trails the other models, reflecting the limitations of a smaller-capacity model. Across modalities, RGB and infrared typically outperform depth and thermal, suggesting that these signals provide more task-relevant information. These findings underscore the joint importance of model scale and sensing modality for robust action selection.

7.2.5 Impact of Model Scaling. In addition, we extended our evaluation with a subset of CUHK-X, comprising one-tenth of the full dataset, to include models of varying capability, namely Qwen-35B, QwenVL3-235B, and Doubao-seed-2.0. We observe that scaling model size does not uniformly improve HAU performance. For example, within the Qwen family, the 235B variant yields mixed results relative to the 35B model. It improves on tasks such as Thermal Context Analysis and IR Task Caption, yet degrades notably on Depth Context Analysis, dropping from 0.422 to 0.286, and even slightly declines on RGB Action Selection. Meanwhile, the strongest model, Doubao-seed-2.0, achieves the best overall results but still leaves substantial room for improvement, e.g., reaching only 0.552 on Depth Context Analysis. These findings confirm that the challenges posed by CUHK-X, particularly for non-standard modalities and complex temporal reasoning, persist even with significantly stronger VLMs, and that naively increasing model capacity does not resolve modality-specific difficulties inherent in HAU.

7.3 Benchmark of HARN

We evaluate the performance of LVLMs in inferring intentions and causal relationships in human action sequences.

7.3.1 Results of HARN. As shown in Fig. 8, the average accuracy is 70.25%. We notice that VLLaVA-7B leads on RGB and infrared and remains competitive on depth. QwenVL-7B performs consistently well across all modalities and is close to the top overall. QwenVL-3B stands out on depth, surpassing several larger models. We found that InternVL-8B and VR1Chat-7B deliver mid-range results, while InternVL-2B trails the others. On average, depth and infrared yield higher accuracy than RGB, indicating that these signals carry more information for HARN. These findings highlight the combined impact of model capacity and sensing modality on robust human-activity reasoning.

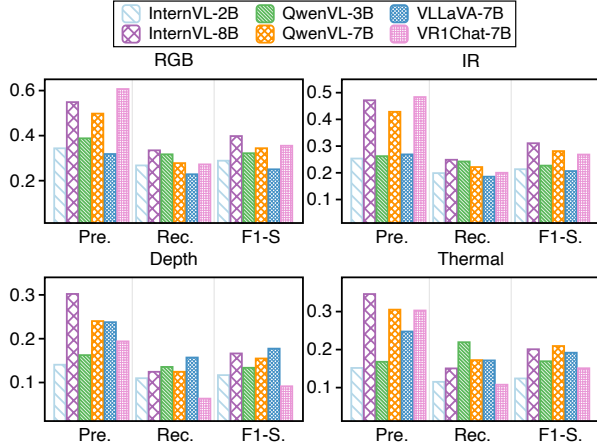


Figure 7: Accuracy of action selection tasks in HAU.

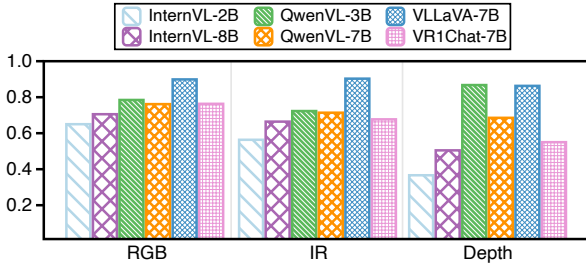


Figure 8: Results of HARn.

7.3.2 Why reasoning models work in HARn? Fig. 9 illustrates why reasoning models outperform captioning models in HARn. Captioning models such as Qwen-7B and InternVL-8B tend to misinterpret superficial cues, whereas the reasoning model VR1Chat leverages contextual understanding and logical inference to associate observed actions (e.g., interacting with items on a table) with the most probable next action (e.g., “Getting Dressed”). It also handles ambiguity better and provides transparent explanations, improving interpretability.

A related question is whether the gains stem from the structured captions or the sensor data. Our captions encode rich context (body posture, object interaction, temporal progression) far beyond categorical labels in existing datasets, and are intentionally designed to ground LLM/LVLM reasoning in fine-grained descriptions aligned with sensor evidence. Replacing them with simple action labels (e.g., “drinking”) would reduce annotation granularity to that of conventional datasets, likely degrading performance; feeding captions alone without sensor input would remove grounding signals needed for ambiguous activities. The consistently strong results when both components are present suggest that structured captions and sensor data are complementary rather than redundant.

8 Discussion

Due to the page limitation, we have provided more discussion of CUHK-X in our Appendix §C.

Limitations. Though developed over a decade ago, the ATUS activity classification system remains actively maintained. We combine ATUS with ActivityNet to leverage complementary strengths: ATUS provides granular, population-grounded activity descriptions

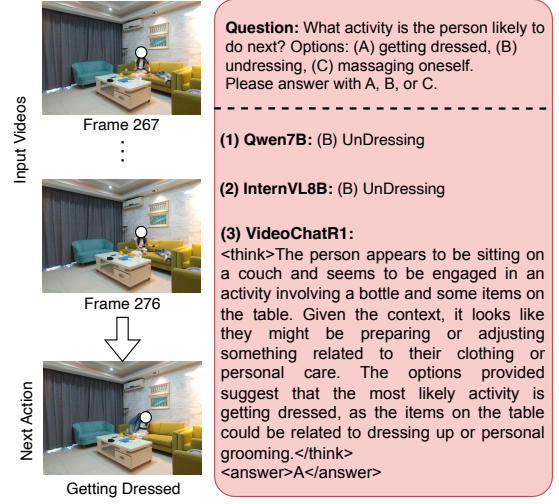


Figure 9: An illustration analyzing why the reasoning model performs better than the captioning model.

across 400+ coded categories [47], while ActivityNet contributes 203 visually-distinct activity classes optimized for video recognition [44]. This hybrid approach aligns with recent multimodal dataset practices that combine multiple sources to address coverage gaps [66]. CUHK-X is a controlled dataset collected in two indoor environments with 30 participants aged 20 to 23. As such, it lacks population-level and ecological diversity, and generalization to other settings, long-horizon activities, or populations with different motor patterns, such as children, older adults, individuals with mobility impairments, is uncertain. Because actions are elicited from scene-level captions, instruction-following may shift motion distributions relative to naturalistic behavior despite preliminary, small-scale evidence of caption similarity. In future releases, we may refine our adjudication protocol, such as reporting inter-rater reliability for the initial independent annotations, which can provide a more comprehensive evaluation.

Modality-Specific Benefits and Trade-offs. Among the modalities in CUHK-X, vision-based ones (RGB and Depth) individually carry the richest discriminative information for activity recognition, as they capture fine-grained spatial and temporal appearance cues. However, they are susceptible to occlusion, lighting variation, and privacy concerns. In contrast, IMU and mmWave individually yield lower standalone accuracy, yet they provide complementary cues that are robust to visual degradation. Specifically, IMU captures body-worn kinematic signals that are invariant to environmental conditions, while mmWave offers contactless motion sensing that operates reliably in low-light or occluded scenarios. When fused with vision modalities, these non-visual sensors consistently improve robustness under challenging conditions and enhance cross-subject generalization, as also observed in prior multimodal fusion studies [37, 57]. Therefore, the inclusion of these modalities is motivated not by their standalone discriminability, but by their complementary value in practical deployments where no single modality is universally reliable.

Instruction-following bias and cross-dataset comparison. Volunteers in CUHK-X follow written instructions rather than act freely, potentially introducing systematic bias in recorded motions and captions. A controlled micro-experiment (§3.2.3) reveals high semantic similarity between free-action and instructed captions (Table 1), indicating the bias is small for the daily activities in CUHK-X. Two limitations remain: both conditions share the same cohort, and only four tasks in one environment are tested. We consider this acceptable because common indoor routines rely on well-practised motor habits that leave little room for instructions to alter action structure, and natural cross-volunteer variation already matches the free-vs-instructed difference. Broader validation across more activities and a cross-cohort design is left to future work. We also note that HAR, HAU, and HARN are independent tasks whose performances do not affect one another. Directly comparing CUHK-X with existing datasets is difficult because those datasets provide only categorical action labels, supporting a fundamentally different task formulation. A more controlled alternative is to vary annotation granularity within CUHK-X while holding sensor data fixed, isolating the contribution of caption structure without cross-dataset confounds.

Ecological validity and script usage. CUHK-X prioritizes precise multimodal synchronization, controlled recording conditions, and reliable ground-truth annotations, which inevitably reduces behavioral naturalness compared with large-scale in-the-wild corpora. We view the two paradigms as complementary: controlled benchmarks enable fine-grained analysis under known conditions, while naturalistic datasets capture broader real-world variability. Future releases will incorporate longer, less scripted interaction routines, wider participant demographics, and more diverse environments. Regarding the use of LLM-drafted scripts, we stress that they served solely as pre-production guides for scenario coverage and were human-reviewed for physical and temporal plausibility before filming. The evaluated VLMs receive only recorded video frames and never access any script text. Ground-truth next-action labels were annotated from the captured footage by human annotators, and natural actor deviations from the scripted plan were preserved. HARN therefore evaluates a model’s ability to anticipate human actions from visual evidence, not to reproduce LLM-generated text.

Future Directions. CUHK-X can scale along two axes: (1) incorporating inter-participant interactions and longer, multi-step routines; and (2) augmenting the current seven modalities with complementary signals such as audio, tactile/contact sensing, and lightweight physiological measures (e.g., heart rate, EEG), as well as expanding to a wider variety of indoor environments. HAU and HARN already cover vision modalities such as Depth, IR, and Thermal. However, there are not yet stable, privacy-friendly LLMs for evaluating 1D/point-cloud sensor data (e.g., IMU, mmWave). Therefore, evaluating such privacy-friendly modalities remains a promising direction for future work. We also plan to broaden the participant pool and environments to strengthen generalizability. In parallel, CUHK-X serves as a standard benchmark for HAR, and as a testbed for LLM-based action understanding and reasoning. Its tightly synchronized multimodal streams make it a practical educational resource for teaching sensor fusion, and multimodal reasoning.

9 Related Works

Human Action Recognition Datasets. HAR analyzes and classifies human actions using various sensors. Vision-based datasets mainly utilize RGB and RGB-D data to capture activities. For example, NTU-60 [43] provides 56,880 videos of daily and health-related actions, while UTD [8] records 27 activities from 8 subjects for classification. Similarly, PKU-MMD [12] and NTU-120 [31] leverage RGB, depth, and skeleton data, supporting 51 and 120 actions, respectively. Sensor-based datasets use wearable or environmental sensors like IMUs, gyroscopes, or radar. UMAFall [7] employs IMU sensors on the chest, waist, wrist, and ankle for fall detection, while Epic-Kitchen [14] combines IMU, RGB, and optical flow to analyze over 90,000 action segments in kitchen environments. Smaller datasets, such as USC [65], Shoaib [46], HHAR [48], and UCI [42], focus on common activities like walking using IMU data. Radar-based datasets, such as HuPR [27], integrate radar and RGB for privacy-preserving action recognition. The emerging dataset Thermal-IM [51] employs thermal imaging and multimodal data to address challenges such as lighting variations and occlusion, enabling effective long-term tracking. MM-Fi [61] integrates RGB, depth, and radar, offering over 320,000 samples for 27 activities conducted by 40 subjects. While limited in scale, mRI [1] also combines IMU, covering 12 activities performed by 20 subjects. However, existing datasets lack comprehensive HAR data from diverse IoT devices.

Human Action Understanding Datasets. HAU involves comprehending actions through perceptual, contextual, and experiential integration, covering recognition, intention, and narrative understanding. Multimodal datasets such as PKU-MMD [12] and Ego-Exo4D [20] support the evaluation of algorithms for understanding complex activities. Captioning-based datasets such as Ego-4D [19] and Ego-Exo4D include 3,670 hours of videos with narrations to enrich activity understanding, while Tarsier2 [64] uses large language models for detailed descriptions. Reasoning-based datasets like ActivityNet-QA [63] and Next-QA [56] focus on spatial, temporal and causal reasoning, with annotated question-answer pairs to enhance deeper video content understanding. DailySTR [40] further leverages the VirtualHome-AIST simulator to create a video-based dataset comprising a total of 80,573 question-answer (QA) pairs. However, these datasets often focus on single modalities or involve high annotation costs. CUHK-X addresses these gaps as the first multimodal dataset for HAU, integrating understanding and reasoning across modalities to advance human action comprehension. In addition, some aforementioned large-scale, naturalistic datasets have demonstrated clear advantages in behavioral diversity, ecological validity, and generalization to unconstrained settings. CUHK-X is designed to complement, rather than replace, such resources by providing tightly synchronized multi-modal streams with controlled ground truth.

10 Conclusion

In this paper, we present CUHK-X, a large-scale multimodal dataset and benchmark in HAR, HAU, and HARN, offering 64,267 samples across seven modalities and two environments. With three carefully designed benchmarks encompassing six tasks, our results demonstrate the robustness of CUHK-X in validating state-of-the-art models across three tasks.

Acknowledgment

We gratefully acknowledge Dr. Jamie Du, Zhijiang Chen, Runju Fan, Yajing Feng, Peipei Li, Yutang Wei, Jiamin Wu, Yixin Xu, and Danni Yuanyong from Huizhou University, as well as Dr. Yunqi Guo from The Chinese University of Hong Kong, for their assistance in collecting the dataset. We also thank Ruijun Xia, Bohan Liu, and Guangyu Chen from The Chinese University of Hong Kong for their help with the paper artifacts. This work is supported by the Research Grants Council (RGC) of Hong Kong, China, under the grants Hong Kong RGC Strategic Topics Grant (STG) STG1/E-403/24-N, SRF5 2526-4S02, GRF 14201425, and CRF C4032-25E.

References

- [1] Sizhe An, Yin Li, and Umit Ogras. 2022. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in neural information processing systems* 35 (2022), 27414–27426.
- [2] Godfred Anakpo, Zanele Ngwayibana, and Syden Mishu. 2023. The impact of work-from-home on employee performance and productivity: a systematic review. *Sustainability* 15, 5 (2023), 4529.
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaozhai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).
- [4] Daumantas Bočkus, Timo Tammi, Elli Vento, and Raija Komppula. 2023. Wellness tourism service preferences and their linkages to motivational factors: a multiple case study. *International Journal of Spa and Wellness* 6, 1 (2023), 78–108.
- [5] Bureau of Labor Statistics. [n.d.]. *American Time Use Survey*. U.S. Department of Labor. https://fraser.stlouisfed.org/files/docs/releases/atus/atus_20250626.pdf
- [6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nibbles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–970.
- [7] Eduard Casilari, Jose A Santoyo-Ramón, and Jose M Cano-García. 2017. Umafall: A multisensor dataset for the research on automatic fall detection. *Procedia Computer Science* 110 (2017), 32–39.
- [8] 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)*. IEEE, 168–172.
- [9] Hongkai Chen, Sirajum Munir, and Shan Lin. 2022. RFCam: Uncertainty-aware fusion of camera and wi-fi for real-time human identification with mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–29.
- [10] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. 2019. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2145–2155.
- [11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 24185–24198.
- [12] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. 2017. PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding. *arXiv preprint arXiv:1703.07475* (2017).
- [13] MMPose Contributors. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. <https://github.com/open-mmlab/mmpose>.
- [14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision* (2022), 1–23.
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [16] Di Duan, Shengzhe Lyu, Mu Yuan, Hongfei Xue, Tianxing Li, Weitao Xu, Kaishun Wu, and Guoliang Xing. 2025. Argus: Multi-view egocentric human mesh reconstruction based on stripped-down wearable mmwave add-on. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. 1–14.
- [17] Di Duan, Huanqi Yang, Guohao Lan, Tianxing Li, Xiaohua Jia, and Weitao Xu. 2023. Emgsense: A low-effort self-supervised domain adaptation framework for emg sensing. In *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 160–170.
- [18] Leah R Gerber, Zachary Reeves-Blurton, Nika Gucci, Gwennlian D Iacona, JA Beaudette, and Teri Pipe. 2023. Practicing mindfulness in addressing the biodiversity crisis. *Conservation Science and Practice* 5, 7 (2023), e12945.
- [19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.
- [20] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19383–19400.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.
- [22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [23] Siyang Jiang, Xian Shuai, and Guoliang Xing. 2024. ArtFL: Exploiting data resolution in federated learning for dynamic runtime inference via multi-scale training. In *2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 27–38.
- [24] Siyang Jiang, Bufang Yang, Lilin Xu, Mu Yuan, Yeerzhati Abudunuer, Kaiwei Liu, Liekang Zeng, Hongkai Chen, Zhenyu Yan, Xiaofan Jiang, et al. 2025. An LLM-Empowered Low-Resolution Vision System for On-Device Human Behavior Understanding. *arXiv preprint arXiv:2505.01743* (2025).
- [25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision* 128, 7 (2020), 1956–1981.
- [27] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. 2023. HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 5715–5724.
- [28] KunChang Li, Yanan He, Yi Wang, Yizhou Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).
- [29] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning unified visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).
- [30] Aixun Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
- [31] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 10 (2019), 2684–2701.
- [32] Shengzhe Lyu, Yongliang Chen, Di Duan, Renqi Jia, and Weitao Xu. 2024. Earda: Towards accurate and data-efficient earable activity sensing. In *2024 IEEE Coupling of Sensing & Computing in AIoT Systems (CSCAIoT)*. IEEE, 1–7.
- [33] Michelle E Mlinac and Michelle C Feng. 2016. Assessment of activities of daily living, self-care, and independence. *Archives of Clinical Neuropsychology* 31, 6 (2016), 506–516.
- [34] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18798–18806.
- [35] U.S. Department of Labor. 2013. American Time Use Survey. <http://www.bls.gov/tus/>. Accessed: 2025-04-23.
- [36] Xiaomin Ouyang, Xian Shuai, Yang Li, Li Pan, Xifan Zhang, Heming Fu, Sitong Cheng, Xinyan Wang, Shihua Cao, Jiang Xin, et al. 2024. ADMarker: A Multi-Modal Federated Learning System for Monitoring Digital Biomarkers of Alzheimer's Disease. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 404–419.
- [37] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 324–337.
- [38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.

- [39] Wenhao Qi, Xiaohong Zhu, Bin Wang, Yankai Shi, Chaoqun Dong, Shiying Shen, Jiaqi Li, Kun Zhang, Yunfan He, Mengjiao Zhao, et al. 2025. Alzheimer’s disease digital biomarkers multidimensional landscape and AI model scoping review. *npj Digital Medicine* 8, 1 (2025), 366.
- [40] Yue Qiu, Shusaku Egami, Ken Fukuda, Natsuki Miyata, Takuma Yagi, Kensho Hara, Kenji Iwata, and Ryusuke Sagawa. 2024. DailySTR: A Daily Human Activity Pattern Recognition Dataset for Spatio-temporal Reasoning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 357–363.
- [41] TJ Quinn, K McArthur, G Ellis, and DJ Stott. 2011. Functional assessment in older people. *Bmj* 343 (2011).
- [42] Jorge-L. Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-aware human activity recognition using smartphones. *Neurocomputing* 171 (2016), 754–767.
- [43] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1010–1019.
- [44] Jungpil Shin, Najmul Hassan, Abu Saleh Musa Miah, and Satoshi Nishimura. 2025. A comprehensive methodological survey of human activity recognition across diverse data modalities. *Sensors* 25, 13 (2025), 4028.
- [45] SungUk Shin and Youngjoon Kim. 2025. Enhancing Graph Of Thought: Enhancing Prompts with LLM Rationales and Dynamic Temperature Control. In *The Thirteenth International Conference on Learning Representations*.
- [46] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors* 14, 6 (2014), 10146–10176.
- [47] Force Statistics, Shelley Kristina, and Kristina J Shelley. 2005. Developing the American time use survey activity classification system. *Monthly Labor Review* (2005), 3.
- [48] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.
- [49] Yue-meng Sun, Zhi-yun Wang, Yuan-yuan Liang, Chen-wei Hao, and Chang-he Shi. 2024. Digital biomarkers for precision diagnosis and monitoring in Parkinson’s disease. *NPJ digital medicine* 7, 1 (2024), 218.
- [50] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Michael Blumenstein, and Jing Jiang. 2020. Omni-scale cnns: a simple and effective kernel size configuration for time series classification. *arXiv preprint arXiv:2002.10061* (2020).
- [51] Zitian Tang, Wenjie Ye, Wei-Chiu Ma, and Hang Zhao. 2023. What Happened 3 Seconds Ago? Inferring the Past with Thermal Imaging. In *CVPR*.
- [52] Zitian Tang, Wenjie Ye, Wei-Chiu Ma, and Hang Zhao. 2023. What happened 3 seconds ago? inferring the past with thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17111–17120.
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [54] Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. 2024. Xrf55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–34.
- [55] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. 2024. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634* (2024).
- [56] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9777–9786.
- [57] 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*. 1–14.
- [58] Bufang Yang, Yunqi Guo, Lilin Xu, Zhenyu Yan, Hongkai Chen, Guoliang Xing, and Xiaofan Jiang. 2025. Socialmind: Llm-based proactive ar social assistive system with human-like perception for in-situ live interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 1 (2025), 1–30.
- [59] Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–29.
- [60] Huanqi Yang, Mingda Han, Xinyue Li, Di Duan, Tianxing Li, and Weitao Xu. 2024. iRadar: Synthesizing Millimeter-Waves from Wearable Inertial Inputs for Human Gesture Sensing. *arXiv preprint arXiv:2412.15980* (2024).
- [61] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2023. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems* 36 (2023), 18756–18768.
- [62] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.
- [63] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, Vol. 33. 9127–9134.
- [64] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. 2025. Tarsier2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding. *arXiv preprint arXiv:2501.07888* (2025).
- [65] Mi Zhang and Alexander A Sawchuk. 2012. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 1036–1043.
- [66] Yumin Zhang and Yanyong Wang. 2025. A comprehensive survey on RGB-D-based human action recognition: algorithms, datasets, and popular applications. *EURASIP Journal on Image and Video Processing* 2025, 1 (2025), 15.
- [67] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. Motionbert: A unified perspective on learning human motion representations. In *CVPR*. 15085–15099.

Appendix

A Overview and Details

A.1 Overview of CUHK-X

We introduce CUHK-X, a large-scale multimodal dataset designed to advance human action recognition (HAR), understanding (HAU), and next-action reasoning (HARn). As shown on the left of the figure, we first build a Scene-based Caption Generation framework to avoid spatiotemporal inconsistency and provide precise ground truth. Guided by ATUS, we define seven action themes and select 40 representative actions with reference to HHAR, UCI, and Cosmo. Large language models then compose coherent captions that stitch these actions into everyday scenes—living room, kitchen, bedroom, and bathroom—augmented with emotional styles such as relaxed or hurried. Using these captions as ground truth (center), 30 participants enact the scenes in two indoor environments, yielding over 64,267 samples across seven synchronized modalities: RGB, depth, thermal, infrared, skeleton, IMU, and mmWave. The setup includes a Vzense NYX 650 (depth), TI IWR6843ISK (mmWave), a Hikvision TB4117 (thermal), and five WitMotion WT9011DCL-BT50 IMUs. Each recording is paired with its caption, forming rich data–caption pairs. On the right, CUHK-X supports three benchmarks implemented as six tasks. HAR performs single-label classification. HAU assesses caption comparison with ground truth, context analysis (e.g., quickly, smoothly, calmly), sequential action reordering to test temporal reasoning, and action selection from a predefined set. HARn predicts the next action from textual descriptions. We evaluate across all modalities using SOTA models, InternVL2.5-2B/8B, QwenVL2.5-3B/7B (captioning) and VLLaVA-7B, VR1Chat-7B (reasoning), and analyze long-tail and cross-subject effects.

A.2 Motivation Details

As shown in Fig. 2, we provide an illustration of the limitations in current LLMs, such as Tarsier [55] and Tarsier2 [64], which face challenges in achieving accurate HAU with depth and thermal modalities, while RGB performs accurate results. We observed that the SOTA captioning models often make the following mistakes: providing inaccurate action descriptions, missing actions, or sometimes both, as shown in Fig. 2. The main reason is that these models are not trained on this modality or designed for the HAU task. Thus, it is necessary to provide datasets with multimodal synchronized “(data, caption)” pairs to enable models to understand such information effectively.

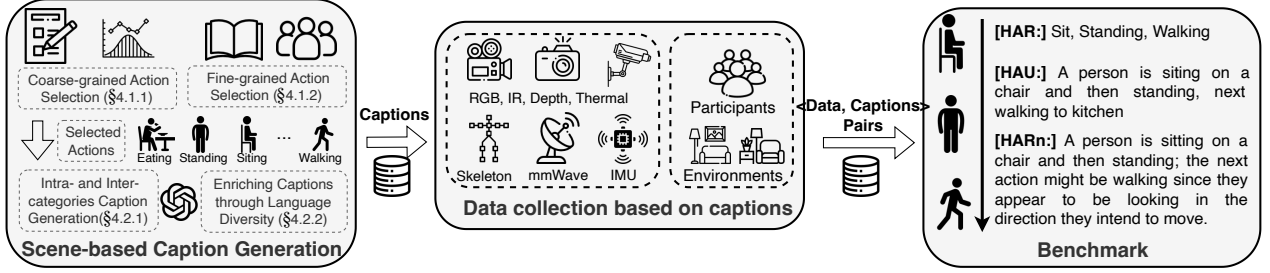


Figure 10: Overview of CUHK-X. We first obtain a set of action categories based on coarse & fine-grained action selection and then generate captions for these actions. We first create scenes based on selected actions and obtain captions. Then, we collect data from 30 participants across seven modalities and obtain the “(data, caption)” pairs. Lastly, these data can support HAR, HAU, and HARn tasks. The distinction between these tasks lies in their objectives: HAR focuses on single-label classification, HAU generates detailed captions, and HARn predicts the next action within a spatiotemporal context.

B Data Description

B.1 Data Statistics

Here, we provide a statistical description of our dataset. As shown in Fig. 11, we show a clear frequency imbalance across human actions. High-frequency actions such as walking, eating, sitting down, and drinking water dominate the dataset, with occurrences exceeding 200, reflecting their ubiquity in daily life. Moderately frequent actions, including pouring a drink, stirring utensils, checking time, and standing up, appear between 50 and 150 times, indicating their importance in routine behaviors while being less universal. In contrast, actions such as folding clothes, watching TV, and playing a game are sparsely represented, with fewer than 20 occurrences, likely due to their specific or context-dependent nature. The dataset follows a long-tail distribution, where a small number of actions account for a large proportion of occurrences, while the majority are infrequent. This imbalance is a common characteristic of real-world datasets, which naturally prioritize capturing frequent, everyday behaviors. Despite this, the dataset spans a diverse range of categories, including basic daily activities, work-related tasks, household chores, and physical exercises, providing a rich foundation for human activity recognition. In particular, in CUHK-X, each participant contributes over 30 minutes of footage with more than 100 samples. For example, vision modalities include 4,029 clips, with a total duration of 19 hours and 29 minutes.

B.2 Data Visualizations

We also provide a visualization of our multimodal data. Fig. 12 illustrates the multimodal data from both ambient and wearable sensors. We present the visualization of common activities, including sitting, walking, eating, drinking water, and pouring a drink. The data includes RGB, Depth, Thermal, IR, Radar, 3D Skeletons, and IMU signals, representing a comprehensive set of modalities for activity recognition. In particular, RGB serves as the primary visual modality, capturing color and spatial context, which is essential for understanding the environment and participants’ actions. Depth data provides spatial structure, highlighting the distance and geometry of objects and participants, while Thermal data visualizes temperature distributions, offering insights into heat signatures that may not be visible in other modalities. Infrared (IR) enhances visibility in low-light or dark environments, complementing RGB and Thermal data. Radar

visualizes motion and spatial dynamics, making it particularly useful for detecting movement patterns. In addition, the 3D Skeletons, extracted using mmPose [13], provide key body joint positions and orientations, enabling precise pose estimation and body movement tracking. These skeletons are overlaid on RGB images for better interpretability of the captured actions. IMU data, collected from five body locations (right arm, left arm, waist, right leg, left leg), includes acceleration, angular velocity, and angles across the X, Y, and Z axes, offering fine-grained motion analysis. In CUHK-X, each modality not only complements the others but is also capable of functioning independently.

C Discussion Details

C.1 Bias Checks of CUHK-X

Because actions in CUHK-X are collected from scene-level captions, instruction-following may shift motion distributions relative to naturalistic behavior. As preliminary evidence, we previously reported (§3.2) a small four-task study indicating high caption similarity. However, caution is warranted when extrapolating to unconstrained, long-horizon activities or to populations with different motor patterns (e.g., children, older adults, or individuals with mobility impairments).

C.2 Scope Limitations

CUHK-X is a controlled dataset and benchmarks for multimodal perception, language grounding, and action reasoning in everyday indoor scenes. The current release covers 30 participants aged 20–23 and two indoor environments. Therefore, we do not claim population-level or broad ecological diversity in CUHK-X. Our main goal is to provide synchronized modalities (RGB/depth/IR/thermal/skeleton/IMU/mmWave) with scene-level ground truth that stress-test cross-modal grounding and temporal reasoning. Expanding demographics and environments is planned future work.

C.3 Verification Procedures

In this release, we do not report inter-rater reliability (IRR), e.g., Cohen’s κ for binary checklist items or Krippendorff’s α for ordinal coherence ratings, because the final annotations are consensus labels produced after adjudication, making post-hoc IRR on the gold set uninformative. Although the intended tasks (HAR, HAU, HARn) primarily target scene semantics and temporal dependencies, we

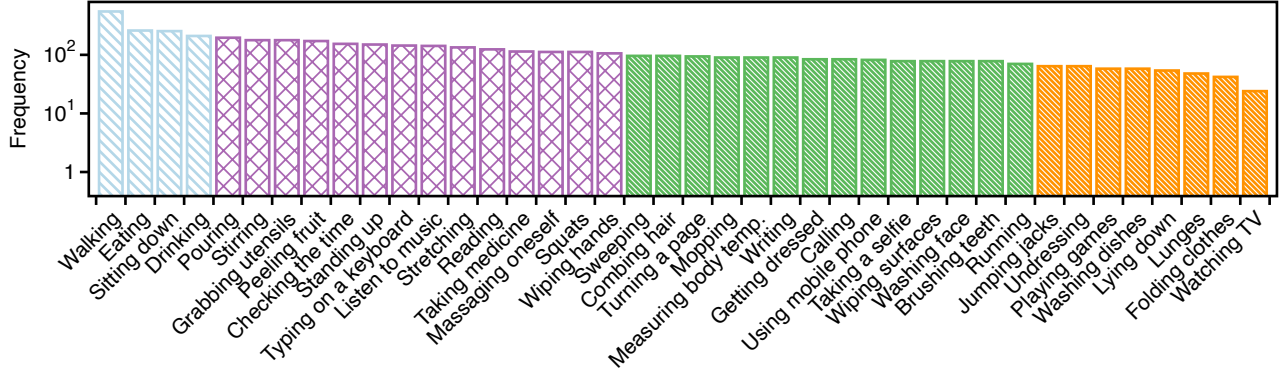


Figure 11: Data Statistics of CUHK-X.

view formal IRR as valuable for auditability; in future releases, we plan to report IRR on a double-coded holdout and to release the rubric and adjudication protocol to facilitate external review.

C.4 More Actions and Modalities.

A key future direction for CUHK-X is addressing its scalability across actions and modalities. Firstly, while CUHK-X is comprehensive, it could be expanded to include more actions that involve interactions between multiple participants. In addition, while the current version includes data from seven modalities, future expansions could incorporate other modalities, such as audio, tactile sensors, heart rate, or EEG. These additional modalities would provide deeper insights into human actions by capturing complementary information, such as emotional states, physiological responses, or fine-grained tactile interactions, enriching the dataset’s multimodal nature. Moreover, while CUHK-X currently includes data from 30 participants, expanding to a larger, more diverse pool of individuals is critical for improving generalizability.

C.5 More Discussions

C.5.1 Cross-subject and cross-domain HAR. Real deployments inevitably face domain shifts that go beyond subject identity. In CUHK-X, shifts arise from (i) environment (different apartments and room layouts), (ii) sensor placements and view angles, (iii) lighting and thermal conditions, (iv) background clutter and occlusions, and (v) subject attire and execution styles. Nevertheless, we believe it is feasible to improve model robustness for both cross-subject and cross-domain HAR.

C.5.2 Discussion of zero-shot and fine-tuning. Our HAU and HARn evaluations intentionally use zero-shot LVLs with task-specific prompts to expose modality gaps and avoid confounds from small-scale fine-tuning. This design choice surfaces two realities: (i) current LVLs are primarily optimized for RGB, and (ii) reasoning over non-RGB modalities remains challenging without targeted adaptation. At the same time, our HAR results show that task-specific fine-tuning on CUHK-X materially improves recognition compared to off-the-shelf backbones (average 76.52% across modalities), motivating a nuanced view of zero-shot vs fine-tuning. In this version, our goal is that CUHK-X can be a benchmark tool to evaluate the performance of different LVLs.

C.6 Broader Impact of CUHK-X

We hope that CUHK-X can make a meaningful impact across several fields. First, CUHK-X can serve as a benchmark to support conventional HAR algorithms, including, but not limited to, evaluating multimodal algorithms, cross-subject approaches, and cross-domain methods. Additionally, it provides a benchmark for assessing the capabilities of current LLMs in action understanding and reasoning. Second, CUHK-X offers synchronous multimodal sensor data, making it easier for researchers and practitioners to explore and work with various sensors and tools. This feature makes it a valuable educational resource, serving as a standard dataset for teaching essential topics such as sensor fusion, data annotation, and multimodal reasoning.

D Artifact Appendix

This appendix provides instructions for the Artifact Evaluation Package of the paper “A Large-Scale Multimodal Dataset and Benchmarks for Human Activity Scene Understanding and Reasoning”. It aims to facilitate the independent evaluation of the reproducibility of the paper. [Please refer to this Google Docs \[2\] for step-by-step instructions.](#)

D.1 What are included

This artifact supports the reproduction of the key experimental results presented in the CUHK-X paper, including benchmarks for Human Action Recognition (HAR), Human Action Understanding (HAU), and Human Action Reasoning (HARn). The artifact package includes the CUHK-X dataset (64,267 samples from 30 participants across seven synchronized modalities: RGB, depth, thermal, infrared, skeleton, IMU, and mmWave), pre-trained model checkpoints, evaluation scripts, and a scene-based caption generation pipeline. Using these resources, evaluators can reproduce the three core benchmark results reported in the paper (Table 3–4, Fig. 6–9). Furthermore, to facilitate reproducibility, we provide remote access to a dedicated server (2×48 GB vGPU) along with detailed documentation. Reviewers are welcome to use this environment to reproduce all experimental results. We have prepared the working environment, data, models, and instructions. Please refer to this [Google Doc](#) for detailed instructions on how to reproduce the results.

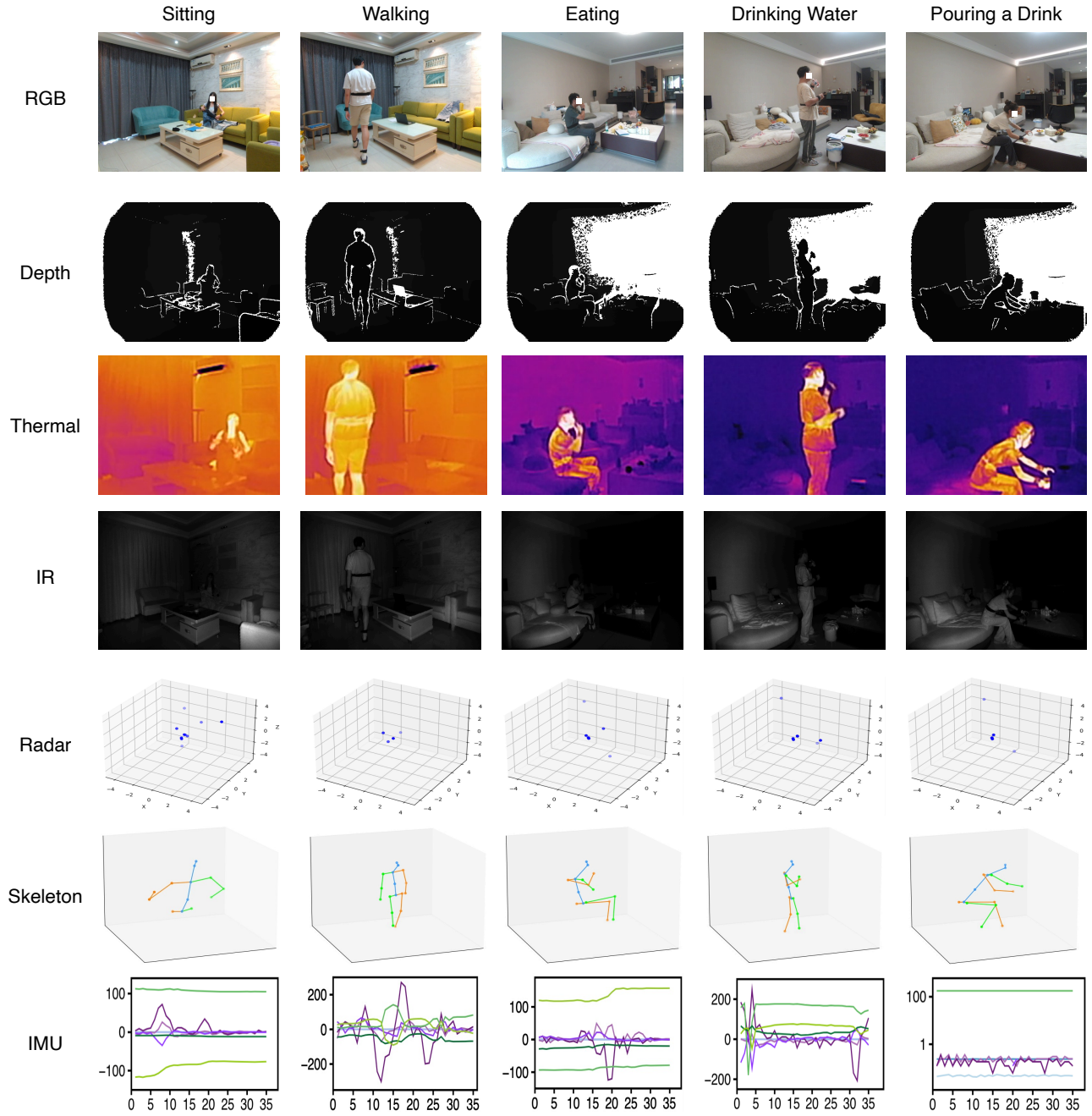


Figure 12: Visualization on ambient and wearable sensor data.

D.2 Artifact check-list (meta-information)

- **Algorithm:** Supervised HAR training with per-modality backbones; zero-shot LLM evaluation for HAU and HARn.
- **Program:** Python 3.9+, PyTorch 2.0+, Hugging Face.
- **Compilation:** NA.
- **Transformations:** Raw 16-bit depth data is range-filtered per room but it is processed well in the released dataset.

- **Binary:** For HAR tasks, all models must be trained from scratch using the provided scripts. For LLM models, we have downloaded the checkpoints in remote access.
- **Dataset:** 64,267 samples, 30 participants, 40 actions in 7 categories, 7 modalities. 18 non-RGB data is available at <https://openaiotlab.github.io/CUHK-X/>. All data is available via remote access. Reviewers can login using `ssh -p 13803 reviewer1@connect.westd.seetacloud.com`. The password can be found in this [Google Doc](#).
- **Run-time environment:** Ubuntu 20.04 LTS; CUDA 11.8+; Python 3.9+ with conda environment.

- **Hardware:** A GPU server with at least one NVIDIA GPU with 24 GB VRAM (e.g., RTX 4090) is recommended for HAR training. For HAU and HARn tasks, two NVIDIA GPUs with 96 GB VRAM are recommended (e.g., two A6000).
- **Run-time state:** Once launched, all experiments run fully automatically without any runtime intervention.
- **Execution:** Shell scripts and Python entry points are provided for each benchmark task (HAR training, HAU zero-shot evaluation, HARn zero-shot evaluation).
- **Metrics:** HAR: Accuracy, Precision, Recall, F1-Score. HAU caption comparison: BERTScore-F1, ROUGE-1, ROUGE-L, BLEU-1, METEOR. HAU context analysis / action reordering / action selection: Accuracy. HARn: Accuracy.
- **Output:** Per-modality classification reports; caption comparison score tables; accuracy scores for HAU sub-tasks and HARn; experiment logs are saved to the `results/` directory.
- **Experiments:** (1) HAR benchmark across 7 modalities with standard 80/20 split, long-tail resampling ablation, and cross-subject evaluation; (2) HAU benchmark: four sub-tasks (caption comparison, context analysis, sequential action reordering, action selection) across vision modalities; (3) HARn benchmark across RGB, IR, and depth.
- **How much disk space required (approximately)?** At least 500 GB to store the full CUHK-X dataset across all seven modalities. An additional ~100 GB is required for model checkpoints and evaluation outputs.
- **How much time is needed to prepare workflow (approximately)?** We have already prepared the environment, data, and LLM checkpoints using remote access for this artifact. For workflow preparation from scratch, 1 hour for checkpoint downloading and 1 hour for environment setup.
- **How much time is needed to complete experiments (approximately)?** Full HAR benchmark takes at most 12 hours per modality on a single GPU (A6000 or equivalent/higher-end device). Full HAU+HARn benchmark: around 20 days depending on GPU availability (2× A6000 or equivalent/higher-end device).
- **Publicly available?** Yes. However, due to an ongoing Kaggle competition, we provide reviewers with remote access to our code and data. For the public dataset, only data from 18 participants is released; the remaining data is reserved for the Kaggle CUHK-X competition evaluation and will be made fully available after the competition concludes.
- **Code licenses:** MIT License.
- **Data licenses:** Data Use Agreement (DUA) – Non-Commercial Research Only.
- **Workflow automation framework used?** Shell scripts. A dedicated `scripts/` folder contains one `.sh` file per experiment (e.g., HAR per modality, HAU sub-tasks, HARn), allowing each benchmark to be launched independently with a single command.
- **Archived (provide DOI)?** Yes. 10.5281/zenodo.19144427.

D.3 Description

D.3.1 How to access. *Obligatory:* Please use `ssh` to access our remote server. Please see the detailed documentation in [Google Docs](#).

D.3.2 Hardware dependencies.

- **HAR:** A single GPU with at least 12 GB VRAM is sufficient for training and evaluating all per-modality HAR models (ResNet-50, MotionBERT, 1D-CNN, PointNet).
- **HAU:** At least two NVIDIA GPUs with 48 GB VRAM each are required for running the captioning models (InternVL2.5-2B/8B, QwenVL2.5-3B/7B) across all four modalities.

- **HARn:** At least two NVIDIA GPUs with 48 GB VRAM each are required for VLLaVA-7B and VR1Chat-7B inference across RGB, IR, and depth modalities.

A CPU with at least 32 GB RAM is recommended for data preprocessing. The original data collection hardware (Goermicro Vzense NYX 650, Texas Instruments IWR6843ISK, Hikvision TB4117, and five WitMotion WT9011DCL-BT50 IMUs) is only required for new data collection, not for reproducing the benchmark results.

D.3.3 Software dependencies.

- Python 3.9+, PyTorch 2.0+, CUDA 11.8+.
- Standard scientific Python stack: NumPy, scikit-learn, pandas, matplotlib.
- Full dependency list provided in `requirements.txt` in the repository.

D.3.4 Datasets. CUHK-X contains 64,267 samples collected from 30 participants performing 40 daily actions across 7 synchronized modalities (RGB, depth, thermal, infrared, skeleton, IMU, mmWave) in two indoor environments. Each sample is paired with a scene-level caption as ground truth.

D.3.5 Models.

- **HAR:** ResNet-50 for RGB/depth/IR/thermal; PointNet for mmWave; MotionBERT for skeleton; 1D-CNN + Transformer encoder for IMU.
- **HAU and HARn:** InternVL2.5-2B, InternVL2.5-8B, QwenVL2.5-3B, QwenVL2.5-7B, VLLaVA-7B, VR1Chat-7B. The checkpoint of InternVL2.5-2B can be found at https://huggingface.co/OpenGVLab/InternVL2_5-2B; other checkpoints can also be found on Hugging Face at <https://huggingface.co/>.

D.4 Installation

Not required for artifacts. For your convenience, we have set up an online server with a pre-configured environment, so no additional installation is necessary. We kindly ask reviewers to use the provided platform for evaluation. Please refer to this [Google Doc](#) for detailed access instructions. After the full dataset is released, users will be able to access it through our GitHub repository. Below, we outline the key steps for setting up the environment.

- (1) Clone the repository: `git clone https://github.com/openaiotlab/`
`CUHK-X.git`
- (2) Create and activate a conda environment: `conda create -n`
`cuhkx python=3.9 && conda activate cuhkx`
- (3) Install dependencies: `pip install -r requirements.txt`
- (4) For HAU/HARn tasks, please download the six modality checkpoints from Hugging Face.
- (5) Directly testing.

D.5 Experiment workflow

For the artifact evaluation, we have provided comprehensive documentation. Please refer to this [Google Doc](#) for detailed setup and reproduction instructions.

D.6 Evaluation and expected results

We have provided three benchmarks for evaluation.

- **HAR:** the average accuracy across all seven modalities.

- **HAU**: the average accuracy across the four sub-tasks and four vision modalities.
- **HARn**: the average accuracy across RGB, IR, and depth.

D.7 Notes

During the reproduction process, we observed that the performance of certain models (e.g., InternVL-8B) deviates slightly from the reported results. Upon investigation, we hypothesize that this discrepancy may stem from differences in GPU architectures and numerical precision across hardware environments [1]. Due to time

constraints, a thorough root-cause analysis was not feasible. Nevertheless, the vast majority of results are successfully reproduced. To ensure transparency, we provide the original results from our paper for comparison. Please kindly find our [Google Docs](#)[2].

[1] Changing the GPU is changing the behaviour of your LLM.

[2] Google Docs Link: https://docs.google.com/document/d/1tF6Q3d0fL_tPl1Zk0QJzL9c8HGOCsy6feq6nEYLvkyU/edit?usp=sharing