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

ABSTRACT

Multimodal human activity recognition (HAR) classifies activities by leveraging complementary modalities. While traditional methods focus on action recognition, advances in Large Language Models (LLMs) enable detailed descriptions and causal reasoning, advancing human action understanding (HAU) and reasoning (HARn). However, most LLMs, especially Large Vision-Language Models (LVLMs), excel at RGB data but struggle with other modalities like Depth, IMU, or mmWave due to limited large-scale "⟨data, caption⟩" datasets. Existing HAR datasets mostly offer coarse-grained "⟨data, label⟩" annotations, insufficient for capturing detailed action dynamics in HAU and HARn. A quickly fixing approach is combining coarse-grained annotations and generating captions via LLMs, but this often lacks the logical spatiotemporal consistencies needed for effective activity representation. In this paper, we introduce CUHK-X, a large-scale multimodal dataset and benchmarks for HAR, HAU, and HARn tasks. CUHK-X uses 40 actions performed by 30 participants in two indoor environments, providing 58445 samples. The actions cover diverse daily scenarios, ensuring practical relevance. To address spatiotemporal inconsistencies in captions, we propose a prompt-based scene creation method using LLMs to generate logically connected activity sequences. CUHK-X also provides three benchmarks with eight tasks to evaluate state-of-the-art baselines. Experimental results show CUHK-X achieves an average accuracy of 52.14%, 40.23% for HAU and 48.32% HARn for tasks. By introducing this large-scale multimodal dataset, we aim to empower the research community to apply, develop, and adapt data-intensive learning techniques for various human activity-related tasks.

1 INTRODUCTION

In the past years, Human Action Recognition (HAR) tasks have been well developed to leverage the power of AI to classify human activities from multimodal sensory data [37, 63]. Beyond classification HAR tasks, Human Action Understanding (HAU) and Reasoning (HARn) tasks further provide richer and detailed descriptions of human activities, facilitating diverse applications in different domains such as healthcare, daily living monitoring, and surveillance [23, 54]. For example, in the management of Alzheimer's disease (AD), a coherent understanding of the patient's longitudinal behaviors is crucial to monitoring daily routines, providing timely caregiver support, and preventing accidents [36, 56]. As shown in Fig. 1, a traditional HAR task, however, is limited to recognizing isolated human actions, like "*sleep*" or "*fall*", and lacks the ability to interpret a continuous sequence of actions as a whole. The HAU task, in contrast, addresses this limitation by understanding and providing natural language descriptions of the sequence of actions, such as "*the subject is eatting, holding a knife in his right hand and a fork in his left*", which provide valuable contexts for early detection of cognitive decline. HARn tasks, further, infer intentions from the sequence of human actions and predict future actions. A typical instance is that if a subject is

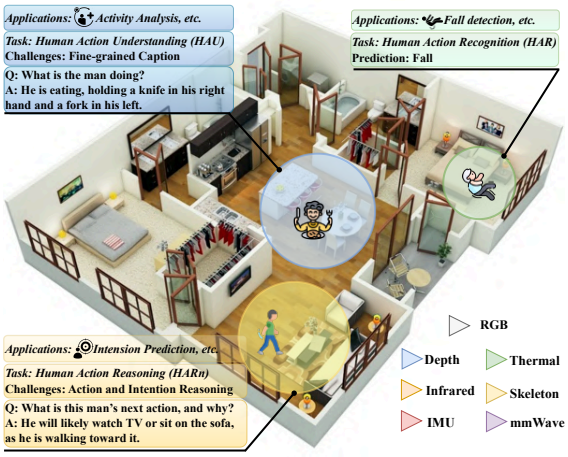


Figure 1: CUHK-X spans a multi-room home and supports three tasks: HAR, HAU (captioning task), and HARn (question answering task), integrating diverse modalities, including RGB, depth, thermal, infrared, IMU, skeleton, and mmWave, to enable robust perception and reasoning in complex indoor contexts.

observed as "*a subject is walking toward to sofa.*", the next action could be predicted as "*attempting to watch TV or sit on the sofa,*", which triggers a preventative intervention.

In practice, such understanding and prediction of human actions are usually not a straightforward autoregressive process that can be enforced by conventional deep neural networks (DNNs). Instead, they require the capability of knowledge representation and logical reasoning that integrates environmental contexts and scene knowledge [24, 57]. To obtain such capabilities, existing techniques usually use high-quality datasets with annotations, in the form of "⟨data, caption⟩" pairs, to fine-tune Large Language Models (LLMs). Logical reasoning can further be triggered on LLMs via Chain-of-Thoughts [19, 34], Tree-of-Thoughts [53, 65], or Graph of-Thoughts [4, 43].

Most existing HAR datasets only provide coarse-grained '⟨data, label⟩' pairs in RGB images [6, 25], and cannot be used to fine-tune LLMs for the aforementioned HAU and HARn tasks. Some recent datasets provide fine-grained pairs of RGB images and the corresponding captions [7, 16, 17], but the fixed fields of views (FOVs) and limited mobility of RGB cameras hinder the timely capture of human behaviors in many practical scenarios. Furthermore, in privacy-sensitive scenarios such as daily home monitoring, RGB images may also contain sensitive personal data and hence raise privacy concerns. Instead, some alternative sensor modalities, such as Depth, Thermal, IMU, and mmWave, are better options. However, there is currently a significant lack of large-scale datasets with "⟨data, caption⟩" pairs in these non-RGB modalities, as most existing datasets in non-RGB modalities are still limited to coarse-grained "⟨data, label⟩" pairs [32, 42, 64]. While the LLMs, particularly Large

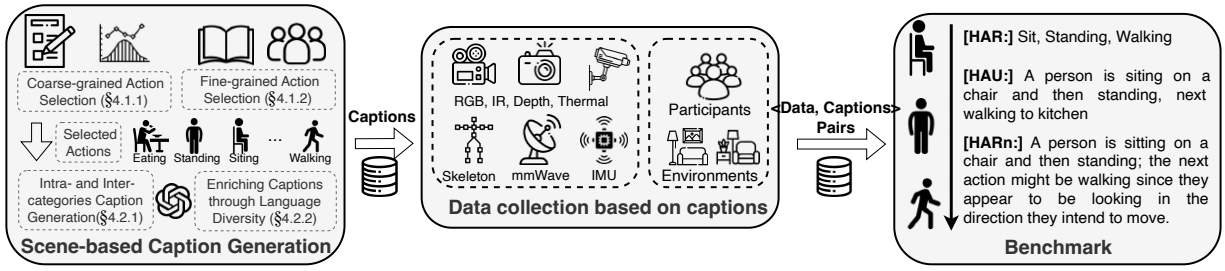


Figure 2: Overview of CUHK-X. We first obtain a set of action categories based on coarse & fine-grained action selection and then generate captions of these actions. We first create a scene based on selected actions and obtain the captions. Then, we collect data from 30 participants across seven modalities 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.

Vision Language Models (LVLMs), perform well in processing RGB and text data, they face significant challenges when applied to other common modalities, such as depth, thermal, and infrared (IR) data. The primary reason is the lack of large-scale training datasets with multimodal action data, i.e., "⟨data, caption⟩" pairs [24].[1]

To obtain captioned datasets across multiple sensory modalities, a naive approach is to merge unimodal datasets, combining coarse-grained labels and using an LLM to generate captions from them. However, the pairs produced by this combination often lack the necessary *spatiotemporal consistency*. For example, it would be inconsistent to directly combine actions like brushing teeth and eating into one scene, as these two actions typically occur independently in different contexts (i.e., bathroom and dining room). Besides, when generating captions, the LLMs often cannot precisely infer the actual human behavioral contexts from the given coarse-grained labels due to their limited representation power, hence resulting in incomplete, inaccurate, or even misleading captions (see §2.2.2 for details).

Therefore, we present CUHK-X, a large-scale multimodal dataset with seven modalities and three carefully designed benchmarks, to address the aforementioned limitations and advance research in human action understanding and reasoning, while also supports the conventional HAR task. To avoid spatiotemporal inconsistency and ensure accurate ground truth in the multimodal dataset, our construction of the CUHK-X dataset starts with a Scene-based Caption Generation Framework (Fig. 1 - Left). This framework categorizes human actions into seven thematic groups informed by the American Time Use Survey (ATUS) [35, 60], and deliberately selects 40 representative actions based on their frequency and relevance as observed in prior benchmark datasets such as HHAR [46],UCI [41], and Cosmo [37]. Afterwards, we use LLMs to logically connect the selected actions into semantically coherent captions that depict predefined scenes of practical daily living, such as the living room, kitchen, bedroom, and bathroom, and also incorporate varied emotional contexts (e.g., performing actions in a relaxed or hurried manner) to enrich the narrative coherence further.

By using the generated captions as the ground truth of action scenes, CUHK-X comprises data collected from 30 participants and includes more than 58,445 daily activity samples across seven distinct modalities, including RGB, depth, thermal, infrared, skeleton, IMU, and mmWave sensors, thereby offering a rich and diverse array of input signals in two different indoor settings. Specifically, we used the following sensors in data collection: the Goermicro Vzense NYX 650 for depth sensing, the Texas Instruments IWR6843ISK for RF radar sensing, the Hikvision TB4117 for thermal imaging, and 5 WitMotion WT 9011DCL-BT50 IMUs. Participants are instructed to understand and act out the generated captions, enabling the collection of well-designed pairs (Fig. 2 - Center).

To verify the practical usability of the CUHK-X dataset, we provide eight benchmark tasks of HAR, HAU, and HARn, by applying state-of-the-art baselines of DNNs and LLMs to these benchmark tasks (Fig. 2 - Right). These tasks include **(1)** HAR; HAU tasks such as **(2)** caption comparison, **(3)** context analysis, **(4)** sequential action prediction, and **(5)** action selection; and **(6)** HARn from the human action descriptions as the output of the HAU task. In HAR benchmarks, the cross-trial, cross-subject, and cross-domain HARs aim to validate that the ActScene dataset contains the necessary knowledge for the tasks. In HAU benchmarks, caption comparison evaluates the generated captions against the ground truth; emotion analysis infers emotional contexts from observed action behaviors; sequential action reordering tests understanding of temporal dependencies by reordering the shuffled actions, and action selection evaluates the identification and classification of actions from a predefined set. In HARn benchmark, we evaluate the LLM's ability to infer the underlying intentions, causal relationships, and the logical progression of human actions. To evaluate HAR task performance, we conduct both full-layer model training and last-layer fine-tuning across seven data modalities. For HAU and HARn tasks, we included six SOTA baselines, including four captioning models, namely InternVL2.5-2B [11], InternVL2.5-8B [11], QwenVL2.5-3B [3], QwenVL2.5-7B [3], and two reasoning models, namely VideoLLaVA-7B [29] and VideoChatR1-7B [28]. The goal of these benchmarks is to explore the tasks performance and differentiability over different models and modalities.

The HAR results show that fine-tuning the models with CUHK-X leads to significant accuracy improvements compared to the task performance achieved using pre-trained models alone, thereby verifying that the CUHK-X contains necessary knowledge in different data modalities to improve the HAR task. In particular, it achieves an average accuracy of 40.18% across seven modalities with full-layer model training, a relative improvement of 52.14% compared to last-layer fine-tuning. For HAU tasks, it achieves an average accuracy of 40.23% across four vision modalities, with the highest accuracy

[1]We consider datasets of pairs ⟨data,GT⟩, where GT denotes ground truth. We divide them into two categories: (i) ⟨data,caption⟩ datasets, where GT is a free-form text caption; and (ii) ⟨data,label⟩ datasets, where GT is a discrete label.

of 68.27% (QwenVL-7B in IR) and the lowest of 18.13% (InternVL-2B in RGB). In HARn task, it achieves an average accuracy of 48.23% across four vision modalities, with the highest accuracy of 78.78% (VideoR1Chat-7B in depth) and the lowest of 18.13% (InternVL-2B in depth). These results also demonstrate that CUHK-X allows robust benchmarking, and is able to bridge the key gaps in existing datasets and benchmarks. To the best of our knowledge, CUHK-X is the first large-scale dataset and benchmarks that well support HAU and HARn tasks. The main contributions are summarized as follows:

- We introduce a large-scale dataset, namely, CUHK-X, with 58.445 samples from 30 participants across seven modalities in two environments, including RGB, Depth, Thermal, IR, IMU, mmWave, and Skeleton, offering diverse and realistic activity data.
- To solve the challenges of logical consistency and spatiotemporal representation, we propose a prompt-based scene creation that leverages prompt-driven LLMs to generate daily activity scenes composed of a series of logically connected actions.
- We provide three benchmarks with eight tasks to evaluate state-of-the-art baselines, establishing CUHK-X as a robust foundation for advancing HAR, HAU, and HARn research.

2 MOTIVATION STUDY

2.1 Applications of HAU and HARn

CUHK-X can be applied to HAU and HARn tasks in various domains, including smart health [57], smart homes [23, 45], and disease intervention [36], enabling continuous, longitudinal monitoring and analysis of user behavior. For example, in Alzheimer's Disease (AD) monitoring [10], long-term reasoning and action understanding are essential to track subtle behavioral changes that indicate cognitive decline. CUHK-X can identify patterns such as a patient frequently forgetting to complete daily routines, wandering, or repeating tasks, which may suggest disease progression. Similarly, for Parkinson's Disease (PD) [47], it can detect nuanced motor impairments, such as a patient attempting to stand, struggling to maintain balance, and then sitting down again, signaling potential worsening of motor symptoms and providing insights for early intervention. Beyond healthcare, CUHK-X can also be applied in smart home systems to enhance comfort and energy efficiency. For instance, it can predict user actions such as adjusting room lighting or thermostat settings [24], optimizing energy consumption while maintaining a comfortable environment. This enriched understanding, enabled by CUHK-X, is critical not only for improved caregiving but also for creating smarter, more efficient living spaces.

2.2 Limitation of the Existing Datasets

The existing dataset generally has limitations in existing multimodal datasets, such as small subject pools (e.g., USC, Shoaib, HHAR) and limited activities (e.g., UTD, mRI, Thermal-IM). While datasets like NTU-RGBD and Ego-Exo4D are extensive, they often lack of modality diversity (e.g., most datasets lack IR, thermal, or captions) as shown in Table 1, HAU-X is designed to provide a comprehensive, multimodal dataset with diverse subjects, rich activities, and additional modalities (e.g., captions) to support advanced research in human activity understanding.

2.2.1 Limitation of the existing coarse-grained HAR datasets. As shown in Table 1, many earlier datasets, such as USC [64], Shoaib [44], and HHAR [46], are limited by small participant numbers (fewer than 15) and a narrow range of activities (e.g., only 6–12 actions). Similarly, datasets like Thermal-IM [49] and UTD [9] involve too few participants or a limited number of activities. Some recent datasets, such as NTU-60/120 [32, 42], and PKU-MMD [12], include much more participants (e.g., 66 people in PKU-MMD) and activity classes (e.g., 60 actions in NTU-60), but they focus mainly on RGB and skeleton data, overlooking key modalities like thermal, IR, and IMUs. For instance, recognizing actions using only RGB data is challenging in cases of occlusion or when a person faces away from the camera. Thus, a key limitation of coarse-grained datasets is their inability to provide reasonable and detailed information for HAU or HARn.

2.2.2 Limitation of the existing fine-grained HAU datasets. Previous fine-grained datasets such as Ego-4D [16] and Ego-Exo4D [17] can enhance models' ability to understand human actions in greater detail since they provide detailed descriptions of the data. However, these datasets are limited in terms of modality coverage. As shown in Table 1, Ego-4D and Ego-Exo4D include RGB data but lack other essential modalities, such as depth, thermal, infrared, and skeleton. However, to the best of our knowledge, there are currently no multimodal LLMs that can effectively understand modalities such as depth, thermal, infrared, skeleton, IMU, and mmWave. In addition, state-of-the-art (SOTA) captioning models, such as Tarsier [52] and Tarsier2 [61], are specifically designed for the RGB modality, resulting in suboptimal performance in other modalities, such as depth. Specifically, we conducted experiments on the UTD [9], Thermal-IM [49], and PKU-MMD [12] datasets using Tarsier and Tarsier2. As shown in Fig. 3, we observed that the SOTA captioning models often make the following mistakes: providing inaccurate action descriptions, missing actions, or sometimes both. 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.

2.3 Summary

In summary, existing coarse-grained datasets are unsuitable for HAU and HARn tasks due to their lack of descriptive details, while existing fine-grained datasets fail to cover multiple modalities comprehensively. To address these limitations, CUHK-X bridges these gaps by providing a wide variety of modalities (e.g., RGB, depth, thermal, IR, skeleton, IMU, and captions), a diverse range of activities, and a larger cohort of participants. This comprehensive dataset not only supports HAR tasks but also enables reasoning-based tasks like HARn, which require understanding sequential actions and predicting future behaviors, as well as fine-grained HAU tasks that demand a deeper contextual understanding. Furthermore, CUHK-X facilitates the development of robust multimodal models for HAR, HAU, and HARn tasks, while also establishing benchmarks for these tasks.

3 DATA COLLECTION SETUP

In this section, we describe our hardware and environmental configuration and provide a visualization of the collected data.

Table 1: A summary of the related coarse-grained HAR and fine-grained HAU datasets (● means included).

Dataset	Years	# of Samples	# of Subjects	# of Activities	RGB	Depth	Thermal	IR	Skeleton	IMU	mmWave	Caption
USC [64]	2012	840	14	12	○	○	○	○	○	●	○	○
Shoaib [44]	2014	70	10	7	○	○	○	○	○	●	○	○
HHAR [46]	2015	3,240	9	6	○	○	○	○	○	●	○	○
UTD [9]	2015	3,444	8	27	●	●	○	○	●	●	○	○
UCI [41]	2016	180	30	6	○	○	○	○	○	●	○	○
NTU-60 [42]	2016	56,880	40	60	●	●	○	●	●	○	○	○
PKU-MMD [12]	2017	20,000	66	51	●	●	○	●	●	○	○	○
NTU-120 [32]	2019	114,480	106	120	●	●	○	●	●	○	○	○
Cosmo [37]	2022	3,434	30	14	○	●	○	○	○	●	●	○
mRI [1]	2022	300	20	12	●	●	○	○	○	●	●	○
Thermal-IM [50]	2023	783	2	24	●	○	●	○	○	○	○	○
MM-Fi [58]	2023	320,000	40	27	●	●	○	○	○	○	●	○
XRF55 [51]	2024	42,900	39	55	●	●	○	●	○	○	●	○
ActivityNet [7]	2015	9,682	-	203	●	○	○	○	○	○	○	●
Ego-4D [16]	2022	5,831	923	146	●	○	○	○	○	●	○	●
Ego-Exo4D [17]	2024	5,035	740	689	●	○	○	○	○	●	○	●
CUHK-X	2025	58,445	30	40	●	●	●	●	●	●	●	●

3.1 Hardware Setup

3.1.1 Ambient sensors. As shown in Fig. 4a, firstly, we use a Goermicro Vzense NYX 650 camera to capture RGB, depth, and infrared 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.

3.1.2 Wearable sensors. 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 ($\pm 2000°/s$), and magnetic field (±2 Gauss). It supports output frequencies ranging from 0.2 Hz to 200 Hz and provides angular measurements of up to ±180° (X/Z) and ±90° (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. Measuring32.5×23.5×11.6 mm in size and weighing just 9g, each participant was equipped with 5 of these devices, with sensors placed on the wrists, ankles, and waist using adjustable bands, as shown in Fig. 4b.

3.2 Environmental Setup

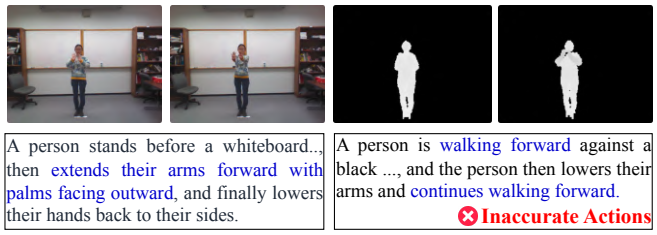
We collected data from two indoor environments, with a focus on four common room settings: the living room, kitchen, bedroom, and bathroom. These rooms were selected to represent a diverse range of daily activities for studying and analyzing human behavior in realistic, everyday scenarios. To ensure 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. 5, 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.

4 GT-FIRST DATA COLLECTION

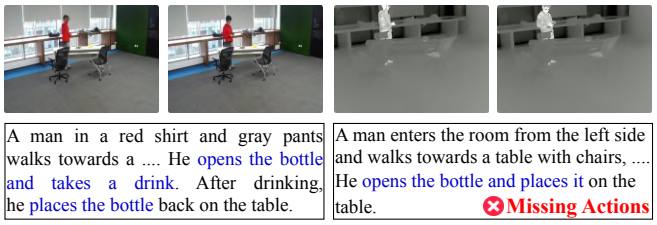
In this section, we discuss different methods of data collection, such as the data-driven and GT-driven methods, and further discussed solutions to the possible biases that may occur in GT-driven methods. Based on these discussions, in Section 5 we will further substantiate our GT-driven data collection method, by ....

4.1 Motivation

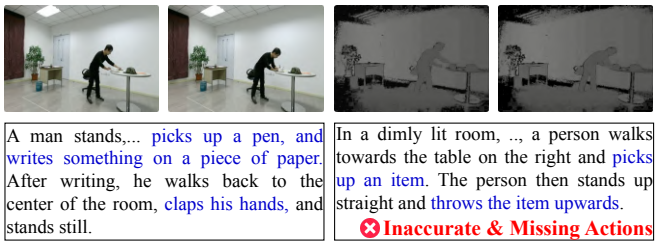
An ideal solution for building a dataset of "⟨data, GT⟩" pairs for human-related actions would be to passively collect large volumes



(a) UTD Dataset



(b) Thermal-IM Dataset



(c) PKU-MMD Dataset

Figure 3: An illustration of the limitations in current LLMs, such as Tarsier [52] and Tarsier2 [61], which face challenges in achieving accurate HAU with depth and thermal modalities, while RGB performs accurate results.

of real-world data and then annotate them with human labels, i.e., a data-first approach. While conceptually appealing, at our target scale this is cost- and labor-prohibitive, raises privacy and safety concerns, and is ultimately impractical and not scalable.

Conversely, GT-first approaches [9, 32], i.e., obtaining ground truth first and then using it to guide data collection, allow us to target rare, safety-critical, or otherwise scarce events, reduce label noise by design, and enforce coverage constraints within a predefined scope. Accordingly, our objective is not to approximate the full real-world distribution, but to curate a high-coverage, well-documented dataset focused on the specific phenomena we explicitly care about. Note that we make no claim of exhaustiveness beyond this scope.

4.2 Bias in GT-First Approaches

GT-first approaches, i.e., those that derive supervision or text directly from a predefined ontology and associated templates/paraphrases, offer clear benefits in controllability, reproducibility, and label faithfulness. In CUHK-X, we define a focused action ontology and employ LLMs to instantiate GT annotations under varying scene conditions. However, it also induces systematic artifacts that reflect the prior encoded by the ontology and the templating process. In particular, they may bias models toward (i) *Coverage and Diversity*, (2) *Discrepancy and Affinity*. In the following, we justify why such biases are acceptable in CUHK-X.

Table 2: Evaluation on discrepancy and affinity.

BertScore-F1	Compare-Cap.	Free-Act.	Instruct-Act.
Task-1	-%	90.40%	93.20%
Task-2	-%	87.20%	86.90%
Task-3	-%	90.00%	89.40%
Task-4	-%	90.90%	89.80%

Coverage. We define coverage as the number of classes included in the dataset. First, with respect to *coverage*, it is hard for a dataset to enumerate the full space of human actions, and real-world taxonomies vary in granularity, exhibit polysemy, and differ across datasets. Even when labels are covered, their realizations may show narrow lexicalization, limited contextual variety, and stereotyped co-occurrence patterns. 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 (§5.1.1 and §5.1.2).write more to be self-contained.

Diversity. We defined diversity as, for each modality, how much different samples in the same class differ from each other. For example, (RGB, image), difference kinds of drinking. Similarly, (IMU data) We mitigate diversity constraints by generating captions from diverse action combinations (§5.2.1) and enriching them via the language diversity to vary lexical choice, syntax, and context (§5.2.2).

Discrepancy is denoted as the mismatch between captions generated by an LLM and those authored by humans for the *same* meaning in CUHK-X. To minimize this discrepancy, we incorporate a human-in-the-loop verification stage that enforces physical plausibility, causal/logical coherence, and contextual appropriateness (§5.2.3), aiming to bring LLM captions close to human parity. We also provide an experiment for further evaluation. In particular, we staged an office-like scene comprising 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 data, we obtained three independent human-authored captions and one LLM-generated caption, with the latter filtered by our verification pipeline. We compare the semantic meaning of these captions using BertScore with the corresponding RGB recordings. As shown in Table 2 (left second column), the two sets of captions of 4 tasks are semantically very similar, as reflected by high BertScore. The primary result is that we focus on human actions and validate LLM outputs against physics-based constraints.

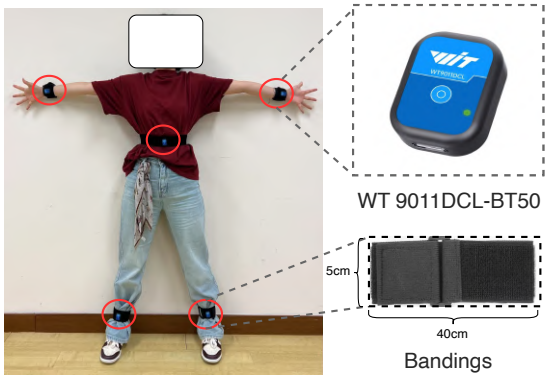
5 SCENE-BASED CAPTION GENERATION

5.1 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. Firstly, a two-stage selection process is employed to select representative actions.

5.1.1 Coarse-grained Action Selections via Predefined Categories and Cross-dataset Frequency. Firstly, based on the ATUS [35] activity hierarchy and ActivityNet [7], 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. CUHK-X adopts a structured semantic framework that leverages hierarchical relationships between

(a) Ambient sensors include the Goermicro Vzense NYX 650 for depth sensing, the Texas Instruments IWR6843ISK for radar sensing, and the Hikvision TB4117 for thermal imaging. TSRV-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 WT 9011DCL-BT50 IMU module.

Figure 4: Visualization on ambient and wearable sensor data.

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 [32, 42], UTD [9], and UCI [41], which are primarily summarized in Table 1. As illustrated in Fig. 6, 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.

5.1.2 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 [40]. The *Working* category (5 actions) was inspired by [2], while the *Socializing and Leisure* category (6 actions) was shaped by [5]. Finally, the *Sports and Exercises* (8 actions) and *Caring and Helping* (3 actions) categories were supported by insights from [15].

As shown in Fig. 7, we select 40 actions which are divided into seven categories include the following: **(1) Personal Care**, which has 6 actions including Washing face, Brushing teeth, Combining 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 the music with headphones; **(6) Sports and Exercises**, which has 9 actions including Walking, Lunges, Siting 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.

5.2 Prompt-based Scene Creation

In this subsection, we propose a prompt-based scene creation approach designed to logically connect the selected actions (§5.1) via constructing various daily living scenes.

5.2.1 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," "Brushing 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.

5.2.2 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 [20] or DeepSeek [30], to add more attributes

Figure 5: 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.

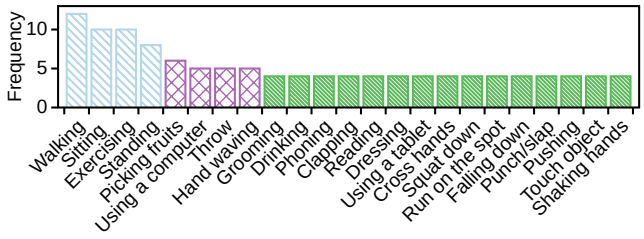


Figure 6: Frequency of the actions among USC [64], Shoaib [44], HHAR [46], UTD [9] ActivityNet [60], UCI [41], NTU [32, 42], PKU-MMD [12], Cosmo [37], mRI [1], Thermal-IM [49] Datasets. We present the high-frequency occurrences (#frequency > 4) in these datasets.

and adverbials. For instance, we can enrich the description of the example in §5.2.1 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.

5.2.3 Human-driven Checking via Physical Knowledge and Logical Coherence. To reduce discrepancy and hallucination in LLM-generated captions (§5.2.1, §5.2.2), we implement a human-in-the-loop verification stage that enforces physical plausibility, scene logic, and dataset conventions before acceptance as ground truth. Graduate-level raters, trained on our environment layouts (Fig.5), 40-action ontology (Fig.7), and sensor capabilities, conduct a two-pass review (independent dual rating followed by adjudication) using a structured checklist and edit protocol.

We validate captions against the following criteria: *(1) Physical feasibility and kinematics.* Body-pose transitions must be anatomically plausible; no teleportation, discontinuous trajectories, or impossible joint motions. Object-state transitions must obey continuity (e.g., "cup is empty" → "pouring" → "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 field of view (FOV). *(3) Temporal and causal coherence.* Event order must be logically progressive (e.g., "grabbing utensil" precedes "eating"; "undressing" before "getting dressed" is flagged unless explicitly motivated). *(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. 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, prior to pairing with multimodal data for HAR, HAU, and HARn benchmarks.

6 PUT ALL THINGS TOGETHER

6.1 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

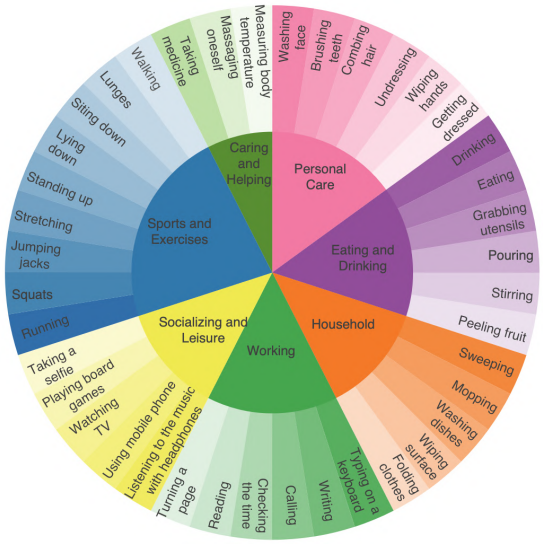


Figure 7: Overview of 40 actions based on the seven categories.

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 processing.

6.2 Data Statistics

Here, we provide a statistical description of our dataset. As shown in Fig. 8, 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, we each participant contributes over 30 minutes of footage with more than 100 samples. For example, vision modalities include 4,029 clips, comprising with a total duration of 19 hours and 29 minutes.

6.3 Data Visualizations

We also provide a visualization of our multimodal data. Fig. 9 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.

6.4 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 §5 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.

7 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.

7.1 Tasks Descriptions

HAR Task. HAR is a task focused on identifying and classifying human activities such as walking, running, sitting, and standing

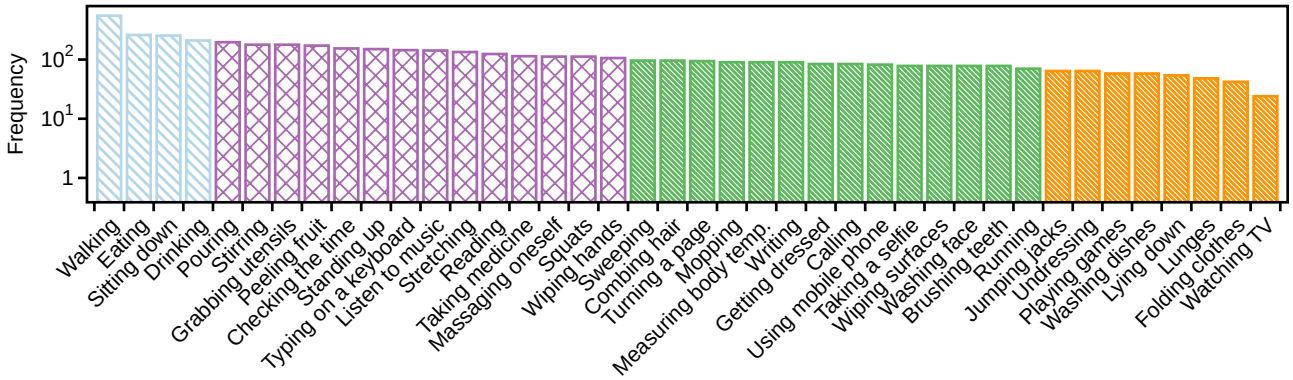


Figure 8: Data Statistics of CUHK-X.

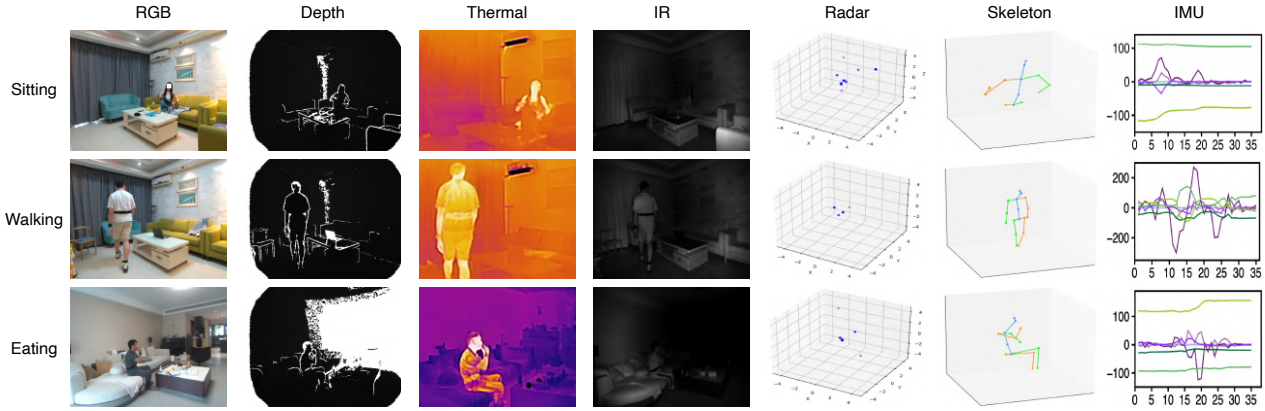


Figure 9: Visualization on ambient and wearable sensor data.

from sensor data. In particular, we define 40 classes across seven categories for recognition as our HAR tasks.

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 must identify the correct emotion exhibited by the participants. For example, if a participant performs actions in a hurried manner, it can be inferred that they might be experiencing anxiety.
- **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.

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

7.2 Baseline and Metrics

7.2.1 HAR Task. We use ResNet-18 [18] for its visual recognition effectiveness. Radar data is processed with PointNet [38], enhanced by feature engineering to capture spatial characteristics. Skeleton data employs MotionBERT [66], 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 [48] with three convolutional layers and a linear classification head for temporal dynamics. HAR task evaluation uses Accuracy (Acc.), Precision (Pre.), Recall (Rec.), and F1-score metrics.

7.2.2 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 (QwenVL-3B) [3], QwenVL2.5-7B (QwenVL-7B) [3], VideoLLaVA-7B (VLLaVA-7B) [29], and VideoChatR1-7B (VChatR1-7B) [28]. For the caption comparison task in HAU, we use metrics including BERT-Score, ROUGE-1, ROUGE-L, and BLEU-1 scores. Additionally, we use accuracy to evaluate the emotion analysis, sequential action reordering, activity selection, and HARn task.

7.3 Configurations

7.3.1 Processioning of Depth Data. 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

Table 3: Overall Performance of HAR Task in ActScene.

Modality	Accuracy	Precision	Recall	F1-Score
RGB	89.96%	90.96%	89.69%	90.41%
Depth	87.92%	89.56%	87.91%	88.53%
IR	90.33%	91.00%	90.32%	90.18%
Thermal	97.01%	97.37%	96.95%	96.56%
mmWave	32.35%	33.68%	29.87%	13.08%
IMU	34.80%	30.96%	26.48%	27.16%
Skelton	87.76%	87.74%	87.62%	84.31%

in Figure 5. 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. 5, 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. 5, 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.

7.3.2 Implementation Details of HAR, HAU, and HARn. For both training schemes, we set the learning rate to 0.001, use a batch size of 32, and update parameters with the Adam [26] optimizer. For HAR tasks, since each action is repeated three times for each subject, we use the first two repetitions used for training and the final repetition reserved for testing. For RGB, IR, thermal, and depth modalities, models are initialized with ResNet-18 pre-trained weights, while for IMU and radar modalities, baseline models are initialized using Kaiming initialization method [18] since no general pre-trained models are available for these modalities. Besides, for the skeleton modality, we initialize MotionBERT [66] with weights pre-trained on the NTU RGB+D dataset [42]. 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) emotion analysis- *"What emotion does the person experience while performing the activities?"*; (3) sequential action reordering- *"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."*; (4) action selection- *"Please sort the following activity lists in chronological order based on the video content."*.; (5) HARn- *"What activity is the person likely to do next?"*.

8 BENCHMARKS

We present three benchmarks of CUHK-X, including HAR, HAU, and HARn. Note that in HAU and HARn, we only compare the vision-based modality, as there is currently less widely accepted LLM that demonstrates strong performance in other sensor-based modalities, such as IMU or mmWave.

8.1 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?*

8.1.1 Overall Performance. As shown in Table 3, modalities exhibit varying performance across experimental settings. Under conventional supervised training, vision-based modalities demonstrate superior performance, with IR achieving the highest F1-score with 90.32%, followed by RGB with 89.69% and depth with 87.91%. This demonstrates that our dataset contains knowledge that supports HAR tasks, as its performance significantly exceeds random guessing. In addition, other sensor-based modalities, such as mmWave, IMU, show considerably lower performance, with radar with 32.35% and IMU with 34.80% on accuracy. This gap might arise from the lower spatial resolution and signal-to-noise ratio of mmWave reflections, sensitivity of IMU signals to sensor placement and orientation, and greater domain shift across trials for these sensors.[2]

8.1.2 Long-tailed Class Performance. As shown in Fig.8, CUHK-Xexhibits a long-tailed class-frequency distribution, which may affect the results. The imbalance ratio is approximately 10 (i.e., the most frequent class appears about ten times as often as the rarest), indicating a moderate, though not extreme, level of imbalance [45]. This issue can be alleviated with standard techniques such as data resampling [59], data augmentation[21, 22], or balanced-loss objectives [45]. In particular, as shown in Fig. 10a, applying class-balanced resampling on RGB, IMU and Skeleton modalities in CUHK-X yields a measurable improvement in accuracy. In particular, resampling consistently improves accuracy across modalities, with the largest gain for RGB from 88.60% to 97.00%, and smaller but noticeable gains for IMU.

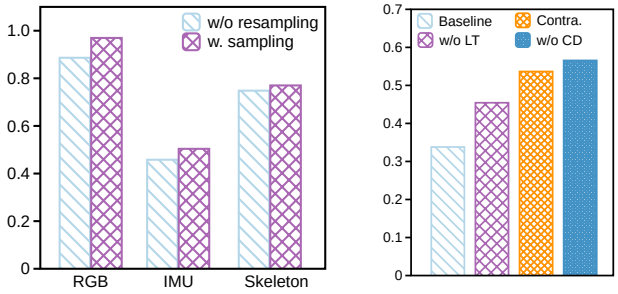
8.1.3 Cross-subject Performance. We evaluate cross-subject performance of CUHK-Xusing a leave-one-subject-out (LOSO) protocol: in each fold, one subject is held out for testing and the remaining subjects are used for training; results are averaged over five folds. As shown in Fig. 10b, the Baseline (LOSO on RGB only) exhibits a substantial accuracy drop due to subject shift and the long-tailed label distribution. Performance improves as we progressively mitigate these factors: removing long-tailed classes (w/o LT) yields a clear gain; adding contrastive learning (Contra.) further strengthens subject-invariant representations; and excluding cross-domain data (w/o CD) achieves the best result by eliminating domain shift, reaching 56.56%. Compared with the perfomance of conventional HAR, accuracy drops markedly in the cross-domain setting, which is intrinsically challenging due to domain shift; even state-of-the-art methods report only $\sim$60% accuracy in this setup[31].

8.2 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.

8.2.1 Results of Caption Comparison. As shown in Table 4, QwenVL-3B and VLLaVA-7B demonstrate superior performance across multiple modalities including RGB, IR, depth, and thermal, achieving higher scores in BERT-Score precision, recall. QwenVL-3B particularly excels in generating accurate and coherent captions, as reflected in its recall and BLEU-1 (B.-1) scores. Interestingly, InternVL-2B and InternVL-8B models demonstrate relatively BERT-Scores

[2]Note that we want to indicate that each modality carries some task-relevant information, not which modality is superior.



(a) Long-tailed Class. **(b) Cross-subject on RGB.**

Figure 10: Long-tail and cross-subject performance. *w/o LT* means long-tail classes are removed; *Contra.* means contrastive learning is added; *w/o CD* means cross-domain data are excluded.

across all modalities but underperform in ROUGE and BLEU metrics. InternVL-2B/8B shows consistent performance but slightly lags in recall and B-1. VR1Chat-7B performs competitively but is outpaced by larger models in fluency. Overall, larger models (e.g., 7B) generally outperform smaller ones, emphasizing the importance of scale for multimodal reasoning in CUHK-X. However, we observe that the BERT-Score is not an effective metric in this context (see §9 for details). Therefore, we propose three sub-tasks to further evaluate the LLMs' capability in HAU.

Table 4: Results of caption comparison. We evaluate the precision (Pre.), recall (Rec.), F1-score (F1-S.) of BERT-Score as well as ROUGE-1 (R.-1), ROUGE-L (R.-L), and BLEU-1 (B.-1), respectively.

Model	BERT-Score			R.-1	R.-L	B.-1
	Pre.	Rec.	F1-S.			
RGB						
InternVL-2B	86.18%	82.17%	84.11%	4.32%	3.51%	0.76%
InternVL-8B	85.81%	81.99%	83.84%	2.61%	2.28%	0.58%
QwenVL-3B	86.22%	86.19%	86.19%	16.67%	14.07%	21.95%
QwenVL-7B	84.74%	86.59%	85.65%	18.77%	12.35%	55.97%
VLLaVA-7B	85.82%	86.41%	86.10%	13.09%	14.40%	22.32%
VR1Chat-7B	84.76%	87.18%	85.94%	14.42%	11.21%	21.51%
IR						
InternVL-2B	85.84%	82.12%	83.92%	4.16%	3.53%	0.50%
InternVL-8B	86.15%	82.08%	84.05%	2.87%	2.46%	0.59%
QwenVL-3B	86.74%	86.37%	86.54%	14.75%	14.10%	22.19%
QwenVL-7B	84.74%	86.15%	85.45%	19.58%	11.67%	19.58%
VLLaVA-7B	86.27%	85.59%	85.92%	19.73%	13.67%	19.73%
VR1Chat-7B	85.36%	86.72%	86.03%	15.55%	12.14%	21.31%
Depth						
InternVL-2B	85.67%	81.99%	83.77%	4.58%	3.93%	0.48%
InternVL-8B	85.47%	82.04%	83.71%	2.84%	1.88%	0.67%
QwenVL-3B	84.77%	85.34%	85.05%	13.96%	11.70%	18.45%
QwenVL-7B	84.76%	85.49%	85.04%	20.29%	11.70%	20.29%
VLLaVA-7B	86.04%	86.29%	86.04%	4.95%	11.11%	4.95%
VR1Chat-7B	84.28%	85.63%	84.94%	19.86%	11.59%	19.86%
Thermal						
InternVL-2B	85.94%	81.48%	83.54%	3.20%	3.20%	0.27%
InternVL-8B	85.56%	81.84%	83.71%	2.94%	2.94%	0.72%
QwenVL-3B	84.59%	85.48%	85.03%	12.15%	12.09%	17.29%
QwenVL-7B	85.64%	84.48%	85.04%	15.41%	12.09%	15.41%
VLLaVA-7B	85.64%	84.83%	85.02%	16.14%	14.23%	17.14%
VR1Chat-7B	83.79%	85.07%	84.42%	15.44%	12.88%	17.09%

8.2.2 Results of Emotion Analysis. As shown in Fig. 11a, VLLaVA-7B and QwenVL-3B demonstrate superior performances, achieving the highest accuracies across all modalities, demonstrating their strong capabilities in interpreting emotional cues from diverse sensory inputs. The InternVL models (2B/8B) show consistently lower emotion accuracy (24.24%-35.35%) across modalities, suggesting that while their architecture may be optimized for general captioning tasks, they lack specialized emotional understanding capabilities. VR1Chat-7B demonstrates moderate performance (42.09%-49.49%) in standard RGB and Depth modalities but shows decreased effectiveness in IR and Thermal domains. Notably, thermal imaging shows the highest variance (29.29%-77.77%), potentially due to the unique challenges of interpreting emotional states from thermal patterns.

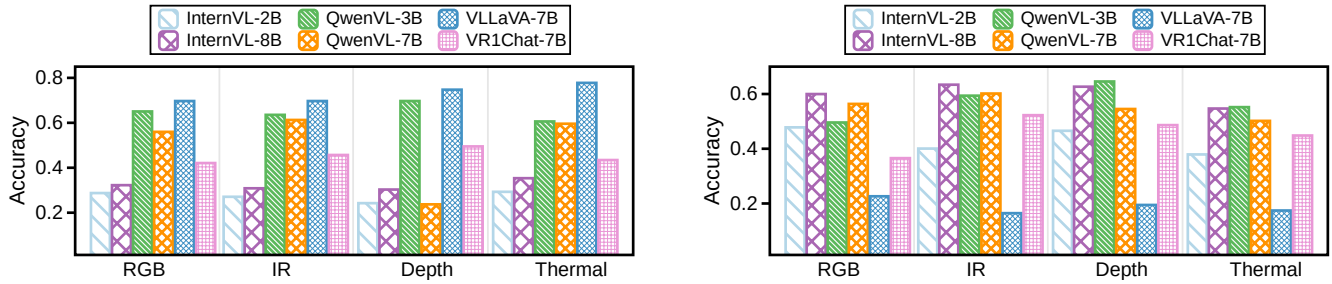
8.2.3 Results of Sequential Action Reordering. As shown in Fig. 11b, we notice that QwenVL-7B consistently achieves the highest accuracy across all modalities, showcasing its superior capability in sequential action ordering tasks. VLLaVA-7B performs closely, particularly excelling in depth and IR modalities, where its accuracies nearly match those of QwenVL-7B. In contrast, InternVL-2B exhibits the weakest performance across all modalities, highlighting the limitations of smaller-scale models in this task. Notably, Depth and IR modalities yield higher overall performance compared to RGB and Thermal, suggesting their greater relevance for sequential action reordering. These results emphasize the impact of model scale and modality choice in complex action understanding.

8.2.4 Results of Action Selection . As shown in Fig. 12, QwenVL-7B consistently achieves the highest or near-highest scores across all metrics and modalities, demonstrating its strong capability in action selection tasks. VLLaVA-7B performs competitively, particularly excelling in IR and thermal modalities. InternVL-8B shows moderate performance, especially in RGB, while InternVL-2B consistently lags behind other models, reflecting its limitations as a smaller-scale model. Notably, performance in IR and Thermal modalities is generally higher than in RGB and Depth, suggesting these modalities provide richer information for action selection. These results highlight the importance of model scale and modality in achieving robust action selection capabilities.

8.3 Benchmark of HARn

In HARn, our goal is to evaluate the performance of captioning and reasoning models across different modalities in inferring intentions and causal relationships within human action sequences.

8.3.1 Results of HARn. As shown in Fig. 13, QwenVL-7B achieves the highest performance across all three modalities, demonstrating its superior ability to understand and reason about human activities. VLLaVA-7B follows closely, particularly excelling in Depth, where its performance approaches that of QwenVL-7B. InternVL-8B and QwenVL-3B exhibit moderate results, with noticeable improvements in RGB and IR compared to InternVL-2B, which consistently underperforms due to its smaller scale. Depth and IR modalities yield higher overall accuracy compared to RGB, indicating their greater informativeness for HARn tasks. These results highlight the importance of leveraging larger models and specific modalities for achieving robust human activity reasoning.



(a) Context Analysis (b) Sequential Action Reordering

Figure 11: Results of context analysis and sequential action reordering in CUHK-X. The metric is accuracy.

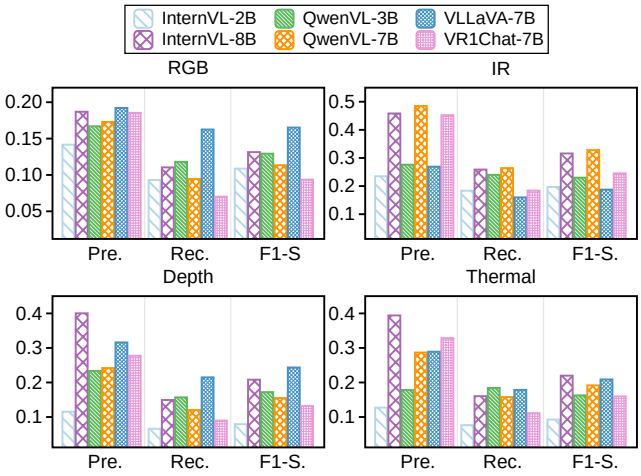


Figure 12: Accuracy results of action selection tasks in HAU.

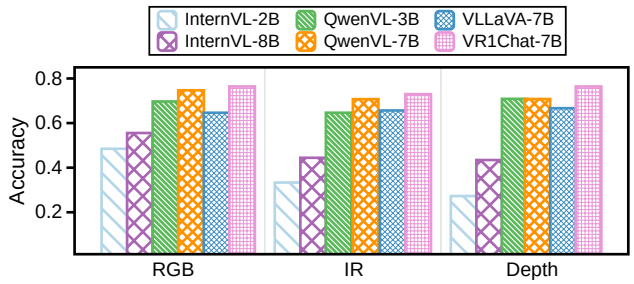


Figure 13: Results of HARn.

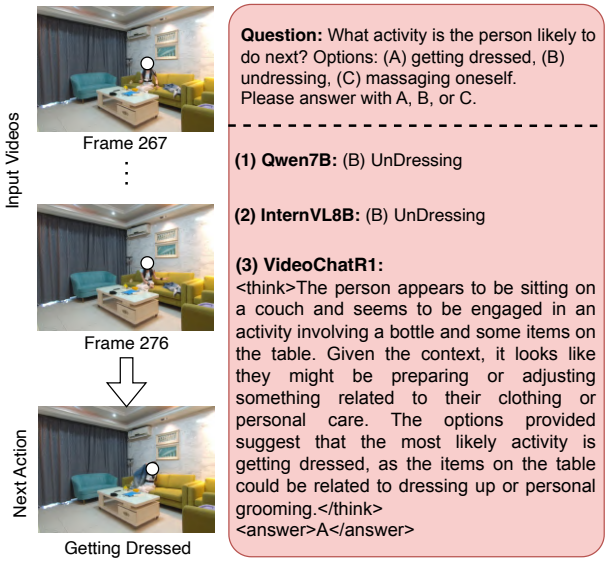


Figure 14: An illustration example of analysis why the reasoning model performs well than the captioning model.

8.3.2 Why reasoning model works in HARn? Fig. 14 illustrates the superiority of reasoning-based models in HARn tasks. Unlike captioning models, e.g., Qwen-7B, InternVL-8B, that misinterpret superficial cues, the reasoning model, i.e., VideoChat-R1, leverages contextual understanding and logical inference. It associates observed actions, such as interacting with items on the table, with the most likely next action ("Getting Dressed"). Additionally, the reasoning model excels in handling ambiguity and provides transparent explanations, enhancing interpretability. This capability to integrate temporal reasoning and contextual synthesis makes reasoning models more reliable for HARn tasks, where understanding intent and action progression is critical.

9 DISCUSSIONS

Scalability of More Actions, Modalities, and Participants. A key future direction for CUHK-X is addressing its scalability across actions, modalities, and participants. 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 and more diverse pool of individuals is critical for improving generalizability. Collecting data that better represents different demographics, environments, and cultural contexts would make the dataset more applicable across a wide range of scenarios.

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 makes it a valuable educational resource, serving as a standard dataset for teaching essential topics such as sensor fusion, data annotation, and multimodal reasoning.

10 RELATED WORKS

Human Action Recognition Datasets. Human Action Recognition (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 [42] provides 56,880 videos of daily and health-related actions, while UTD [9] records 27 activities from 8 subjects for classification. Similarly, PKU-MMD [12] and NTU-120 [32] leverage RGB, depth, and skeleton data, supporting 66 and 120 actions, respectively. Sensor-based datasets use wearable or environmental sensors like IMUs, gyroscopes, or radar. UMAFall [8] 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 [64], Shoaib [44], HHAR [46], and UCI [41], 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 [49] employs thermal imaging and multimodal data to address challenges such as lighting variations and occlusion, enabling effective long-term tracking. MM-Fi [58] 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. Human Action Understanding (HAU) involves comprehending actions through perceptual, contextual, and experiential integration, covering recognition, intention, and narrative understanding. Multimodal datasets like PKU-MMD [12] and Ego-Exo4D [17] support the evaluation of algorithms for understanding complex activities. Captioning-based datasets such as Ego-4D [16] and Ego-Exo4D include 3,670 hours of videos with narrations to enrich activity understanding, while Tarsier2 [62] uses large language models for detailed descriptions. Reasoning-based datasets like ActivityNet-QA [60] and Next-QA [55] focus on spatio-temporal and causal reasoning, with annotated question-answer pairs to enhance deeper video content understanding. DailySTR [39] 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.

11 CONCLUSION

In this paper, we present CUHK-X, a large-scale multimodal dataset and benchmark designed to address critical gaps in human activity recognition, understanding, and reasoning. By offering 36,414 samples across seven modalities and two environments, CUHK-X provides a diverse and realistic foundation for advancing HAR-related tasks. Additionally, our novel prompt-based scene creation framework enhances the logical and spatiotemporal representation of activities, enabling more effective evaluation of complex reasoning tasks. With three carefully designed benchmarks encompassing eight tasks, our results demonstrate the robustness of CUHK-X in validating state-of-the-art models across HAR, HAU, and HARn.

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 Nqwayibana, and Syden Mishi. 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, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai 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] Maciej Besta, Lorenzo Paleari, Jia Hao Andrea Jiang, Robert Gerstenberger, You Wu, Patrick Iff, Ales Kubicek, Piotr Nyczyk, Diana Khimey, Jón Gunnar Hannesson, et al. 2025. Affordable AI Assistants with Knowledge Graph of Thoughts. *arXiv preprint arXiv:2504.02670* (2025).

[5] 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.

[6] Massimo Bosetti, Shibingfeng Zhang, Bendetta Liberatori, Giacomo Zara, Elisa Ricci, and Paolo Rota. 2025. Text-Enhanced Zero-Shot Action Recognition: A Training-Free Approach. In *International Conference on Pattern Recognition*. Springer, 327–342.

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 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.

[8] Eduardo 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.

[9] 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.

[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] Leah R Gerber, Zachary Reeves-Blurton, Nika Gueci, Gwenllian D Iacona, JA Beaudette, and Teri Pipe. 2023. Practicing mindfulness in addressing the biodiversity crisis. *Conservation Science and Practice* 5, 7 (2023), e12945.

[16] 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.

[17] 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.

[18] 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.

[19] Qiangqiang He, Shuwei Qian, Jie Zhang, and Chongjun Wang. 2025. Inference retrieval-augmented multi-modal chain-of-thoughts reasoning for language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[20] 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).

[21] Siyang Jiang, Wei Ding, Hsi-Wen Chen, and Ming-Syan Chen. 2022. PGADA: Perturbation-guided adversarial alignment for few-shot learning under the support-query shift. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 3–15.

[22] Siyang Jiang, Rui Fang, Hsi-Wen Chen, Wei Ding, and Ming-Syan Chen. 2023. Dual adversarial alignment for realistic support-query shift few-shot learning. *arXiv preprint arXiv:2309.02088* (2023).

[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] Kumara Kahatapitiya, Anurag Arnab, Arsha Nagrani, and Michael S Ryoo. 2024. Victr: Video-conditioned text representations for activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18547–18558.

[26] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[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, Yinan He, Yi Wang, Yizhuo 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 united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).

[30] Aixin 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] Hanchao Liu, Yujiang Li, Tai-Jiang Mu, and Shi-Min Hu. 2024. Recovering complete actions for cross-dataset skeleton action recognition. *Advances in Neural Information Processing Systems* 37 (2024), 92055–92081.

[32] 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.

[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] 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.

[40] TJ Quinn, K McArthur, G Ellis, and DJ Stott. 2011. Functional assessment in older people. *Bmj* 343 (2011).

[41] 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.

[42] 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.

[43] 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*.

[44] 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.

[45] Xian Shuai, Yulin Shen, Siyang Jiang, Zhihe Zhao, Zhenyu Yan, and Guoliang Xing. 2022. BalanceFL: Addressing class imbalance in long-tail federated learning. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 271–284.

[46] 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 mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.

[47] 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.

[48] 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).

[49] Zitian Tang, Wenjie Ye, Wei-Chiu Ma, and Hang Zhao. 2023. What Happened 3 Seconds Ago? Inferring the Past with Thermal Imaging. In *CVPR*.

[50] 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.

[51] 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.

[52] 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).

[53] Haoyuan Wu, Xueyi Chen, Rui Ming, Jilong Gao, Shoubo Hu, Zhuolun He, and Bei Yu. 2025. ToTRL: Unlock LLM Tree-of-Thoughts Reasoning Potential through Puzzles Solving. *arXiv preprint arXiv:2505.12717* (2025).

[54] Haiyang Wu, Kaiwei Liu, Siyang Jiang, Zhihe Zhao, Zhenyu Yan, and Guoliang Xing. 2024. Demo abstract: Caringfm: An interactive in-home healthcare system empowered by large foundation models. In *2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 255–256.

[55] 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.

[56] Zongxing Xie, Bing Zhou, Xi Cheng, Elinor Schoenfeld, and Fan Ye. 2021. Vitalhub: Robust, non-touch multi-user vital signs monitoring using depth camera-aided uwb. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. IEEE, 320–329.

[57] 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.

[58] 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.

[59] Rongguang Ye, Yantong Guo, Xian Shuai, Rongye Ye, Siyang Jiang, and Hui Jiang. 2023. Licam: Long-tailed instance segmentation with real-time classification accuracy monitoring. *Journal of Circuits, Systems and Computers* 32, 02 (2023), 2350032.

[60] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9127–9134.

[61] 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).

[62] 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:2501.07888 [cs.CV] https://arxiv.org/abs/2501.07888

[63] Liyu Zhang, Yizhen Wang, Wenjie Du, Kwun Ho Liu, and Xiaomin Ouyang. 2025. Demo Abstract: An LLM-Powered Multimodal Mobile Sensing System for Personalized and Interactive Health Behavior Analysis. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. 720–721.

[64] 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.

[65] Yingjie Zhou, Renzhi Chen, Xinyu Li, Jingkai Wang, Zhigang Fang, Bowei Wang, Wenqiang Bai, Qilin Cao, and Lei Wang. 2025. VToT: Automatic Verilog Generation via LLMs with Tree of Thoughts Prompting. In *2025 Design, Automation & Test in Europe Conference (DATE)*. IEEE, 1–2.

[66] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15085–15099.

CLARIFICATION LETTER

A RESPONSE TO SHEPHERD

B RESPONSE TO REVIEW A

***A-C1**: performance in cross-subject and cross-domain evaluations drops sharply, suggesting that either the dataset still lacks sufficient diversity or existing models are not robust enough for deployment in real-world scenarios. Additionally, the dataset exhibits a long-tailed action distribution, which, while realistic, could hinder effective learning and model evaluation, particularly for reasoning tasks involving rare actions.*

A-A1: Thanks for your comment. We conduct cross-subject and cross-domain analyses and introduce several approaches to mitigate this phenomenon in our dataset; these approaches effectively improve the results. More details can be found in §8.1.3.

***A-C2:** Another concern is the reliance on LLM-generated captions before data collection. While this ensures consistency, it may lead to idealized behavior descriptions that do not capture the full variability of spontaneous human activity.*

A-A2: Thanks for your comment. We recognize the risk of idealization in LLM-generated captions. To minimize its impact, we treat captions as auxiliary metadata and focus our modeling and metrics on action-level signals (what is done, not how it is worded). Under this design, residual caption bias is tolerable because it does not alter the action content or the action-centric outcomes we report (§4.2). In addition, for caption generation we employ human-driven verification to mitigate hallucinations in LLM-generated captions(see §5.2.3).

C RESPONSE TO REVIEW B

***B-C1:** The dataset documentation can be improved by providing the age range and demographics of the participants, how many captions each participant collected data for, the average length of each data segment, etc. Also include the sampling frequencies of sensors.*

B-A1: Thanks for your comment. We have added these information which is presented in §6.1 and §6.2. In particular, xx

***B-C2:** Why were the data statistics (Fig 7) skewed? Since the data is collected based on captions that were pre-generated, the activity distribution can be controlled.*

***B-Q5:** For the sequential action rendering and action selection tasks, how many tasks were considered in each instance? How does performance scale with the number of options presented to the models?*

B-A5: Thanks for your comment. In the sequential action reordering and action selection tasks, each instance contains 3–7 actions. We observe that performance tends to decline as the number of actions increases, reflecting the growing combinatorial difficulty in both tasks. We have added the analysis in §8.2.3 and §8.2.4.

***B-Q6:** There is no evaluation on unseen tasks, which raises questions about the generalizability of models trained on this dataset to new scenarios. For example, the authors claim that the dataset can help build systems that could detect cognitive decline by identifying when users appear to forget to complete daily routines or are repeating tasks, but such scenarios are not represented in the dataset.*

B-A6: Thanks for your comment. We also want to

***B-Q7:** In 7.2.1, all the models evaluated seem to be unimodal. If so, the authors should not make claims about multimodal reasoning.*

B-A7: Thank you for the helpful comment. Our intent is to highlight the role of multiple modalities in understanding and reasoning. CUHK-X supports multimodal experiments because all modalities are temporally and semantically aligned, enabling cross-modal fusion and evaluation.

.

D RESPONSE TO REVIEW C

***C-Q1:** The paper should provide a clearer justification for why the LLM-based generation is superior to manual annotation. A comparison of the naturalness and quality between LLM-generated and human-created captions would strengthen this claim.*

C-A1:

***C-Q4:** Figure error: The labeling/annotations in Figure 3 are incorrect.*

C-A4: Thanks for your comment. We have revised the layout of Vzense NYX 650 and Texas Instrument.

E RESPONSE TO REVIEW D

F RESPONSE TO REVIEW E

***E-Q1:** A major concern lies in the synthetic nature of the captions used in ActScene. Instead of organically collected annotations, the captions are generated by LLMs and subsequently acted out by participants, which risks introducing artificial biases and overly "clean" activity descriptions that may not reflect the messiness and ambiguity of real-world behavior. This design choice undermines the ecological validity of the dataset, as the collected samples may align more with LLM-generated narratives than with natural human activities. Moreover, the paper assumes logical consistency in these captions without reporting any independent validation or annotation process, leaving the reliability of the ground truth highly questionable.*

E-A1: Thanks for your comment. We understand the concerns of reviews of , reflecting the messiness and ambiguity of real-world behavior. However, to the best of our knowledge, before CUHK-X, there are less datasets present aligned mutiple modalites with captions in HAR. Therefore, CUHK-X is a first attempt to for help the community to understand HAR with multiple sensors. As for the messiness and ambiguity of real-world behavior, it could be regarded as a hard examples during . We have added this in our disccusion part. In addition, we has proved that the LLMs-generated caption could bring the same qualtiies with better . (Details could be refers to xx) Lastly, we have added the human-in-the-loop §5.2.3.