Multimodal Daily-Life Logging in Free-living Environment Using Non-Visual Egocentric Sensors on a Smartphone

KE SUN, University of California, San Diego, USA
CHUNYU XIA, University of California, San Diego, USA
XINYU ZHANG, University of California, San Diego, USA
HAO CHEN, Standard and Mobility Innovation Lab, Samsung Research America, USA
CHARLIE JIANZHONG ZHANG, Standard and Mobility Innovation Lab, Samsung Research America, USA

Egocentric non-intrusive sensing of human activities of daily living (ADL) in free-living environments represents a holy grail in ubiquitous computing. Existing approaches, such as egocentric vision and wearable motion sensors, either can be intrusive or have limitations in capturing non-ambulatory actions. To address these challenges, we propose EgoADL, the first egocentric ADL sensing system that uses an in-pocket smartphone as a multi-modal sensor hub to capture body motion, interactions with the physical environment and daily objects using non-visual sensors (audio, wireless sensing, and motion sensors). We collected a 120-hour multimodal dataset and annotated 20-hour data into 221 ADL, 70 object interactions, and 91 actions. EgoADL proposes multi-modal frame-wise slow-fast encoders to learn the feature representation of multi-sensory data that characterizes the complementary advantages of different modalities and adapt a transformer-based sequence-to-sequence model to decode the time-series sensor signals into a sequence of words that represent ADL. In addition, we introduce a self-supervised learning framework that extracts intrinsic supervisory signals from the multi-modal sensing data to overcome the lack of labeling data and achieve better generalization and extensibility. Our experiments in free-living environments demonstrate that EgoADL can achieve comparable performance with video-based approaches, bringing the vision of ambient intelligence closer to reality.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Daily-life logging, Egocentric non-visual sensors, Multi-modal data

ACM Reference Format:
Ke Sun, Chunyu Xia, Xinyu Zhang, Hao Chen, and Charlie Jianzhong Zhang. 2024. Multimodal Daily-Life Logging in Free-living Environment Using Non-Visual Egocentric Sensors on a Smartphone. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 17 (March 2024), 32 pages. https://doi.org/10.1145/3643553

1 INTRODUCTION

The emerging Internet of Things (IoT) promises to embed a massive population of sensors in the environment to form an *ambient intelligence* [1]. Such omnipresent IoT sensors can generate huge personalized data to enable life-logging and support many activity-aware applications. In particular, they can monitor a subject's activities of daily living (ADL), including not only body motion (*e.g.*, walking, bathing), but also interaction with the physical

Authors' addresses: Ke Sun, kesun@ucsd.edu, University of California, San Diego, La Jolla, California, USA; Chunyu Xia, cxia@ucsd.edu, University of California, San Diego, La Jolla, California, USA; Xinyu Zhang, xyzhang@ucsd.edu, University of California, San Diego, La Jolla, California, USA; Hao Chen, hao.chen1@samsung.com, Standard and Mobility Innovation Lab, Samsung Research America, Plano, Texas, USA; Charlie Jianzhong Zhang, jianzhong.z@samsung.com, Standard and Mobility Innovation Lab, Samsung Research America, Plano, Texas, USA.

ACM 2474-9567/2024/3-ART17
https://doi.org/10.1145/3643553

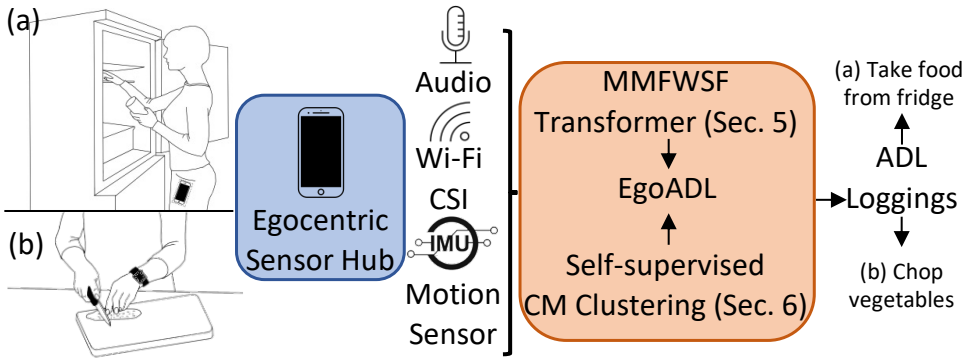


Fig. 1. EgoADL is an egocentric ADL sensing system, leveraging an on-body smartphone as a sensor hub to capture the audio, Wi-Fi CSI, and motion sensor signals simultaneously. EgoADL employs a self-supervised cross-modal (CM) clustering to encode a general feature representation from the large-scale unlabeled data (Sec. 6) and a supervised Multi-Modal Frame-Wise Slow-Fast (MMFWSF) transformer model to recognize the ADLs (Sec. 5).

environment and daily objects (*e.g.*, kitchen appliances, water cups, faucets). They can transform the healthcare domain which, to date, has been relying on laborious monitoring and subjective questionnaires/reports for diagnosis, assessment, and emergent response. Examples include tracking medical compliance, evaluating rehabilitation (*e.g.*, for stroke patients), detecting onset of chronic diseases (*e.g.*, the Alzheimers' disease), *etc*.

To approach the vision of ubiquitous ADL sensing, substantial research has investigated the egocentric sensing scenario, where the sensors are co-located with subjects [2]. In particular, egocentric visual sensing relies on a head-mounted camera or smart glasses to capture the first-person views [3–5]. Due to the limited angle of view and constrained wearing style, they can only partially capture the user's ambulatory activities, leading to low accuracy [4, 6]. Furthermore, these approaches inherit the limitations of camera sensing–They are privacy intrusive, energy hungry, and crippling for continuous sensing. On the other hand, egocentric non-visual sensors such as wearable motion sensors are limited to capturing only ambulatory actions without the interaction with the physical environments and daily objects [7]. Therefore, audio [8, 9], and motion sensor [10, 11] are typically used to assist the egocentric video to improve accuracy.

In this paper, we design EgoADL, a multimodal ADL sensing system that employs ubiquitous non-visual sensors on an egocentric in-pocket smartphone to log ADL in free-living environments. Compared with vision-based approaches, EgoADL is less intrusive and better approximates Mark Weiser's pioneering definition of ubiquitous intelligence [12], *i.e.*, sensing technologies that quietly serve human in the background. As shown in Fig. 1, a user performs daily routines with the sensor hub, *i.e.*, an in-pocket smartphone. EgoADL is designed to log a comprehensive range of basic ADL, which are characterized by open-ended, fine-grained activities encompassing both body motion and human-object interactions. These activities include, but are not limited to, routine physical movements (*e.g.*, walking, sitting, bending down), common household tasks (*e.g.*, cooking, washing dishes, mopping floor), and interactions with various objects in daily life (*e.g.*, using utensils, chopping vegetables, taking food from fridge) (see Fig. 11 and Fig. 12 for detailed ADL list). The focus is on detailing these everyday actions in a manner consistent with the definitions used in computer vision-based ADL recognition [5, 13–15].

To realize EgoADL, we resolve three key challenges:

ADL representation–beyond activity classification. Most of the traditional DNN-based ADL analysis models are designed to perform *classification* [3, 14], which assigns integer IDs to various ADL. However, it requires prescribing a known set of ADL, which falls short of extensibility when new ADL of interest emerge. In contrast,

EgoADL is designed to enable comprehensive daily life logging for humans, covering a wide spectrum of distinct ADLs. Therefore, we propose to use a transformer-based sequence-to-sequence model to decode these feature representations as label name semantics using a sequence of words. Moreover, these ADL representation establishes a connection between sensor data and natural language. This enables us to harness the semantic information inherent in these natural language labels. By seamlessly integrating this information with language models, we significantly elevate the overall performance of EgoADL.

Egocentric multi-modal fusion–overcoming limited resolution of non-visual sensors. The second major challenge is that the non-visual sensors have much lower resolution than camera, for sensing both human body motion and interaction with daily objects [16]. To overcome the challenge, we select and synthesize 3 specific modalities which have already been embedded in commercial devices, *i.e*. motion sensor for ambulatory actions of leg; wireless sensing for full-body motion and interactions with ambient environments; audio recording for motion and human-object interaction with unique sound events. Our empirical analysis of real-world ADL data indicates that each sensing modality is amenable to different ADL patterns, and a judicious combination of them can potentially achieve near-vision resolution. Therefore, we propose a Multi-Modal Frame-Wise Slow-Fast (MMFWSF) encoder to learn the feature representation of multi-sensory data, which characterizes *multi-modal fast-changing motion*, *continuous scene sounds*, and *cross-modal frame-wise alignment*. We then use a transformer-based sequence-to-sequence model to decode these feature representations as label name semantics using a sequence of words.

Self-supervised ADL learning–achieving generalization with limited labeled data. To achieve high sensing accuracy, extensibility for non-frequent ADL and generalization (across different ADL, users and environments), EgoADL needs a massive amount of training data, which entails exorbitant labeling cost due to the high variability and "invisibility" of the sensing records. We observe that capturing the data without labeling is relatively easy for EgoADL, since all the sensors are embedded in an in-pocket smartphone. Thus, we collect large-scale *unlabeled* data and adopt a self-supervised learning (SSL) framework to encode a general feature representation. Specifically, we design a cross-modal deep clustering model that extrapolates two self-supervisory signals from unlabeled data: *i)* Audio captures human-object interaction which can inform the motion sensor and Wi-Fi CSI. *ii)* Correspondence between different modalities when observing the same ADL. In addition, we leverage the vast amount of existing labeled audio datasets [17] to pretrain a feature embedding DNN. These datasets already encompasses the natural logic of ADL, thus further alleviating the training workload of the self-supervised ADL learning.

To validate the design principles behind EgoADL, we implement an Android app to collect the multi-modal data from in-pocket smartphones, when users freely perform daily activities. Our implementation leads to the *first* non-visual multimodal dataset for egocentric ADL. The dataset consists of large-scale unlabeled data, along with a labeled subset for users who are willing to wear a head-mounted camera to capture the ground truth. We implement a labeling software that allows the users to playback the audio/video recordings and annotate the sensor data accordingly.

Using this platform, we have collected 20 hours of labeled data and more than 100 hours of unlabeled data, which includes 221 different types of ADL involving 70 actions and 91 objects. The data was collected from 20 different home environments and 30 users who performed an unrestricted set of activities, encompassing both ambulatory motion and interaction with daily objects. Our evaluation results show that EgoADL achieves 72.5% top-1 and 90.8% top-5 mean Average Precision (mAP) for recognizing 105 frequently-used ADL, which are 21.0% and 14.7% higher than the baseline model using traditional modal-wise sensor fusion. When considering the 35 state-based ADL which typically last more than 5 seconds, EgoADL achieves 85.9% top-1 and 94.5% top-5 mAP. Our results suggest that EgoADL can achieve comparable performance with vision-based egocentric sensing, particularly for non-ambiguous actions and objects using non-visual sensors.

The main contributions of EgoADL are as follows.

• We introduce a new concept of multi-modal egocentric ADL sensing based on non-visual sensors on in-pocket smartphones. We build a platform for EgoADL sensor data collection and labeling, and establish the dataset

	Modality	Scenario	Range	Task	# of Activities	Method	User Study	Receiving Sensors
E-eyes [18]	WiFi CSI	DF	Apartment	HA Classification	9	SM+SL	1 S, 2 E	1 WiFi AP
CARM [19]			Single Room		9	SM+SL	25 S 4 E	1 WiFi AP
EI [20]			Single Room		6	SM+SL	10 S 3 E	1 WiFi AP
WiPose [21]			Static	HPC	16	SM+SL	10 S 1 E	9 WiFi AP
RF-Diary [22]	FMCW radar	DF	Single Room	HA + HOI Captioning	157 Acts, 38 Objs	SM+S	10 S 10 E	FMCW Radar (need floormap)
DeepSense [23]	Motion sensor	Ego	/	HA Classification	6	SM+SL	9 S	Wearable Devices
LIMU [24]		Ego	/		7	SM+SSL	73 S	Wearable Devices
DESED [25]	Audio	DF	/	SEDL	10	SM+SL	/	Mic Array
Ubicoustics [26]		Ego /DF	/	SE Classification	30	SM+SL	7 S	Wearable devices
Cosmo [27]	Ego For motion sensor DF For Depth Cam, Radar		/	HA Classification	14	MM+SSL	30 S 1 E	Wearable devices, Radar, Depth Camera
Wrist-ADL [28]	Audio + Motion sensor	Ego	/	HA Classification	23 Acts	MM+SL	15 S 15 E	Smartwatch
EgoADL	**WiFi CSI + IMU + Audio**	**Ego**	**Whole Apartment**	**HA + HOI Captioning**	**221 Acts 91 Objs**	**MM + SSL**	**30 S 20 E**	**Smartphone**

Table 1. Representative ADL systems using Wi-Fi CSI, IMU and Audio. In "Scenario" column, Ego: Egocentric; DF: Device-free. In "Task" column, HA: Human Activities; HOI: Human-Object Interaction. HPC: Human Pose Construction; SE: Sound Event; SEDL: Sound Event Detection and Localization; In "Method" column, SM: Single-Modal; MM: Multi-Modal; SL: Supervised; SSL: Self-Supervised Learning. In "User Study" column, S: Subjects; E: Environments. This table focuses exclusively on ADL systems employing Wi-Fi CSI, IMU and Audio. There may exist additional sensors applicable in ADL systems, such as PIR sensors, magnetic sensors, sonar sensors, *etc.* [16]

for multi-modal egocentric ADL sensing by using non-visual sensors. Both the platform (https://github.com/Samsonsjarkal/EgoADL) and dataset (https://doi.org/10.5281/zenodo.8248159) are released to facilitate further research.

• We design multi-modal fusion approaches to learn the feature representation of multi-sensory data by leveraging the complementary advantages of audio, motion sensor and wireless sensing.

• We propose an SSL framework that extrapolates single-modal and multi-modal supervisory signals from unlabeled data, in order to boost the model accuracy, generalization and extensibility.

• We propose to leverage the semantic information from the natural language labels by distilling knowledge from external text datasets and refining the labels to fit the sensing capability.

2 RELATED WORK

Egocentric vision-based ADL sensing: The wide availability of head-mounted or body-worn cameras has resulted in massive first-person vision data, and fueled research in egocentric ADL analysis [3–5, 29]. However, egocentric vision approaches still face fundamental deployment barriers. In particular, wearing a camera is inconvenient and invasive [30]. Besides, due to the limited field of view (FoV), the egocentric video data are highly heterogeneous and lack compatibility [3–5]. For instance, the EPIC-KITCHENS [3] captures the users' hands whereas Charades-Ego [4] misses them, causing disparate inference results. Recently, more egocentric vision research uses additional modalities to assist the egocentric vision including audio [8, 9] and motion sensor [10, 11]. However, these approaches inherit the limitations of camera sensing including privacy intrusive, energy hungry, and crippling for

continuous sensing. EgoADL proposes to bring the egocentric ambient intelligence to real life by using non-visual sensors which are less intrusive and insensitive to the FoV problem, yet achieving similar performance as the vision-based approaches.

ADL sensing using non-visual sensors: There exist a huge portfolio of non-visual modalities for ADL sensing, including motion sensor [23, 24, 31], audio [17, 25, 32–34], RF signals [19, 21, 35] and others [16].

Within this area, motion sensor-based ADL sensing has mostly focused on classifying a small set of prescribed activities associated with specific body parts [23, 31]. SSL has been introduced recently to train feature representation models based on motion sensor data [24, 36–38]. EgoADL wildly expands this strand of research towards cross-modal SSL, which fuses multiple modalities to log more complex ADL, similar to a human transcriber.

Sound event detection, as one modality to understand ADL, has been extensively studied [17, 25, 32–34]. Existing solutions use an external microphone array, *e.g.*, one equipped on a voice assistant, to capture the sound. Ubicoustics [26] is the only work that has a smartphone-based egocentric sound capturing setup similar to EgoADL, but it only classifies 30 acoustic activities. Unlike traditional sound event classification, EgoADL is more extensible due to its SSL architecture, and evades majority of the labeling burden by SSL and distilling knowledge from existing sound datasets.

Device-free RF sensing has gained major traction, demonstrating abilities to classify a dozen of prescribed activities [19, 21, 35]. However, due to limited antenna aperture and hence spatial resolution, commercial RF signals, like Wi-Fi sensing, cannot capture the nuances of human-object interactions (HOI), unless augmented with dedicated hardware. For example, LiveTag enables HOI by attaching passive touch-sensitive tags on the objects [39]. RF-Diary [22] employs a powerful FMCW radar and a floor map of object locations to detect HOI. Besides the hardware complexity, cost and labeling burdens, the device-free sensing systems bear a few common limitations. First, the coverage area is typically limited to a single-room. Second, without dedicated hardware [22, 39], the sensing performance is highly sensitive to environment and transceiver locations. State-of-the-art device-free WiFi sensing systems can only achieve $< 75\%$ accuracy in recognizing 6 activities, when tested across different environments [20]. In contrast, EgoADL is the first to explore egocentric Wi-Fi sensing by capturing ambient Wi-Fi CSI using a in-pocket smartphone, combined with other sensors to overcome such limitations.

SSL for ADL sensing: One of the major challenges for ADL analysis is lack of labelled data, especially for relatively rare ADL [13]. This challenge exists even for egocentric vision due to the limited FoV and the complexity of ADL. SSL has proven to be a promising solution for vision [40], motion sensor [24, 36–38, 41], and other modalities [27]. In particular, recent work adopted SSL on unlabeled audio-visual data [42–44]. By understanding the video-audio correspondence, such methods achieve the state-of-the-art performance on egocentric ADL recognition. EgoADL introduces new modeling mechanisms (*e.g.*, joining single- and multi-modal SSL) to tackle a disparate set of modalities. Furthermore, EgoADL distills knowledge from external datasets to guide the SSL towards a better cross-modal feature representation.

3 EGOADL SETUP AND DATA COLLECTION

EgoADL employs a commodity smartphone as an egocentric sensor hub, which captures the audio, wireless sensing signals (*i.e.* Wi-Fi), and motion sensor signals continuously. Users can perform *arbitrary* daily routines with the sensor hub, *i.e.* an in-pocket smartphone, in free-living environments. EgoADL will recognize ADLs including both human activity and human-object interaction from the sensor data without human intervention. We will first discuss EgoADL data collection setup and dataset. We include more setup, data collection details in our Methodological Transparency & Reproducibility Appendix (Sec. 10), and release the source code and datasets to facilitate future research. Our study was approved by the IRB before all the data collection.

Smartphone as an egocentric sensor hub: The EgoADL data collection app is implemented by JAVA in Android OS. It simultaneously collects 3 sensing modalities from the user's smartphone. *(i) Audio capturing:* We

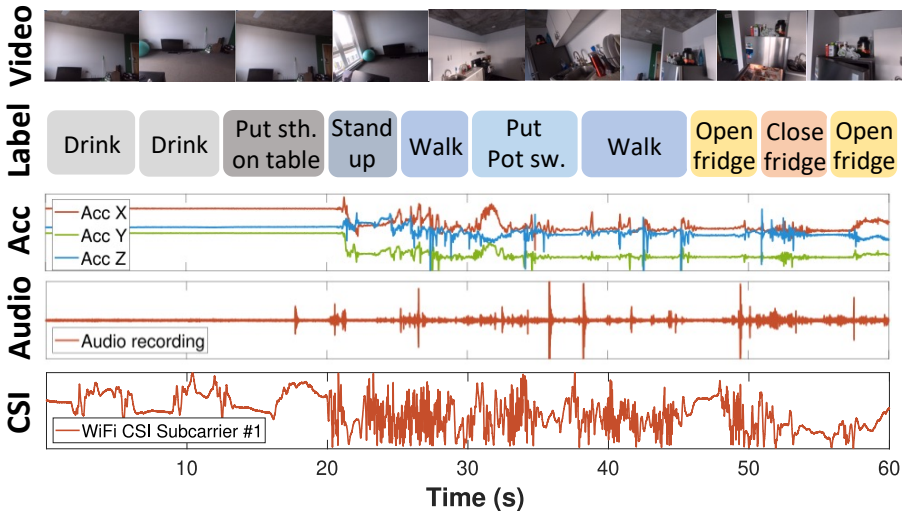


Fig. 2. EgoADL data in the time domain, including Wi-Fi CSI, audio, accelerometer signals, ground truth labels and egocentric video from a head-mounted GoPro for ground truth labeling.

record the monophonic sound at 48 kHz sampling rate through the smartphone's bottom microphone. *(ii) Motion sensors:* We capture the 3-axis accelerometer signals at 200 Hz sampling rate, and log the per-sample timestamp for later uniform resampling. *(iii) Wi-Fi CSI:* We use the Nexmon CSI tool [45, 46] to extract the incoming Wi-Fi packets' CSI. Our EgoADL prototype uses Nexus 5 for its compatibility with Nexmon [45, 46]. During the data collection, we employ a commodity 802.11ac Wi-Fi access point (AP) to transmit data at 400 packets/second. Due to packet losses, the receiving packet rate tends to be lower. Nevertheless, our data collection consistently maintains a minimum reception rate of 200 Wi-Fi packets/second, ensuring data quality (see Sec. 10 for details).

Data collection procedure and setup: For our data collection, we recruited 30 participants, comprising 8 females and 22 males with an average age of 25.9 (refer to Fig. 9(b) for detailed demographics). We clearly communicated the data collection objectives to the participants. Each participant was instructed to record data over a week, aiming for at least 5 hours in total. To guarantee a diverse and sufficient collection of ADL, we advised them to record during routine activities, excluding stationary periods such as working at a desk or sleeping. We did not impose any specific ADL types or scripted activities.

Participants were provided with necessary devices, including a Wi-Fi AP and a smartphone equipped with the EgoADL app, along with instructions for setting up the devices. We asked participants to deploy the Wi-Fi AP at an *arbitrary location* in their home and they are required to simply put the smartphone into their left/right trouser pocket, freely performing daily routines as the data collection app runs in the background. Participants are also allowed to take the smartphone out of the pocket and use the smartphone freely as usual. Upon reaching a user-specified time limit (typically 10 to 30 minutes), the app plays a notification sound and saves the sensor data. For participants who agreed to provide ground-truth labels (7 males and 3 females with an average age of 28.5, Fig. 9(a)), the procedure was identical, with an additional step of wearing a head-mounted GoPro camera. This camera captured egocentric audio and video at a resolution of 2560×1440, 30 FPS, and with a linear field of view.

Data preprocessing: To mitigate the impact of unstable sample timing on Commercial Off-The-Shelf (COTS) smartphones, we first resample the motion sensor data uniformly to 200 Hz. For the Wi-Fi CSI data, we normalize the per-subcarrier magnitude by the corresponding automatic gain control (AGC) values to mitigate the AGC artifacts, and then resample the CSI sequence to 400 Hz. Afterwards, we synchronize the three sensing modalities based on their sampling timestamp. (We include more implementation details in Appendix 10.)

Data labeling: Each ground truth label needs to specify the ADL, *i.e.*, an ambulatory action or human-object interaction event, along with the start and end timestamps. For instance, Fig. 2 shows 1-min labeled data containing a sequence of ADL. We design a labeling tool to allow playback of the GoPro video/audio recordings and annotating the data using a set of ADL labels created by existing state-of-the-art video/audio based ADL sensing systems [5, 13, 15, 47](See more details in our Appendix 10). To ensure accurate labeling, the annotators are exactly the data collectors, compensating for any limitations in egocentric video field of view (FoV) by using their memory of the events. They are equipped to segment and label the video footage, either using the provided predefined ADL labels or by adding new ones as they identify them. This open-ended approach to data labeling and annotation has yielded a comprehensive dictionary encompassing 1,000 words specifically related to ADL. The maximum allowable duration for each labeled segment is 10 seconds, aligning with the typical duration of discrete ADL observed in our studies. We further conduct the user studies with annotators and they empirically separate the ADL into *state-based* and *event-based* ADL [3]. *State-based ADL* usually last more than 5 s each and often periodically and continuously, like "walking" and "chopping meat", *etc*. *Event-based ADL* are one-short, like "opening the door" and "sitting down in chair", *etc*.

Dataset Scale: Following the data labeling process, we streamlined the dataset by reducing the representation of longer-duration event-based ADL, particularly those that span over 10 consecutive minutes. This process yielded an effectively condensed dataset with an average duration of approximately 2 hours per participant. For the data without labels, we selectively refined the data by eliminating segments that lacked significant variation across all three modalities. Finally, we organized the data into three datasets. *(i) Labeled dataset.* The labeled dataset contains 20 hours of records from 10 users across 7 homes, area ranging from 600 to 2000 ft^2 with a variety of layouts. It comprises 7,000 ADL samples, including 221 types of ADL involving 70 actions and 91 objects. A detailed list is in Appendix 10. This dataset serves as a baseline for preliminary experimentation and few-shot fine tuning. *(ii) Balanced dataset with labels.* Remarkably, the uncontrolled user activities manifest an imbalanced long-tail distribution– more than half of the ADL in EgoADL are infrequent and only have less than 15 samples. To establish a baseline supervised learning model, we select a subset from (i), which involves 105 ADL each with 15 to 25 samples (2,500 samples in total), referred to as a *balanced dataset*. *(iii) Unlabeled data.* To facilitate the SSL (Sec. 6), we collected more than 100 hours of unlabeled egocentric data from 30 users in 20 homes.

4 PRELIMINARY STUDY

In this section, we will discuss the advantages of EgoADL design choices, *i.e*. egocentric sensing and sensing modalities selection.

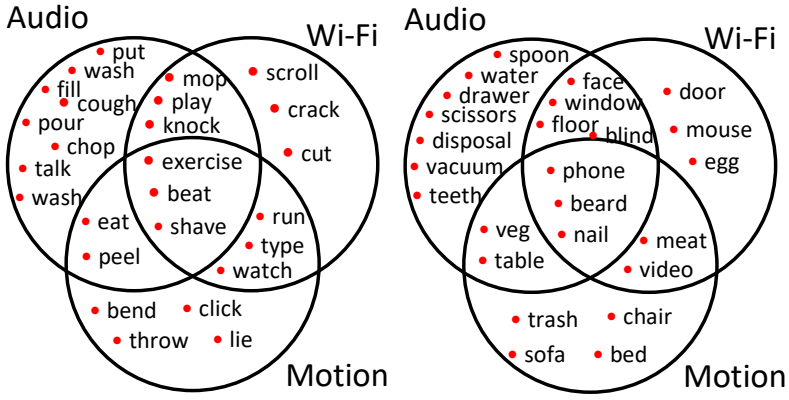
Advantages of Egocentric Sensing: To better understand the advantages of the egocentric sensing, we conduct controlled experiments in a 1400 ft^2 apartment, and compare EgoADL against the conventional device-free setup [48] which captures users' ADL through off-body non-visual sensors, *e.g*., voice assistant [26, 33, 34] and Wi-Fi AP [19, 20]. The details of our experiments are in the Appendix (Sec. 10). We summarize the insights as follows:

i). Sensing Space Coverage: The signal strength of device-free sensing suffers from severe attenuation, diffraction, and scattering effects and drops dramatically as the user moves away from the Tx/Rx or behind the wall. In comparison, in the egocentric setup, the sensor hub accompanies the user, so it achieves whole-home coverage with consistent signal quality for both audio, Wi-Fi CSI and motion sensor.

ii). Resilience to interference: Wi-Fi CSI and audio suffer from interference since they can not distinguish the motion from targeted subject. In contrast, the egocentric sensing setup makes the presence of an interfering user impact significant only when it is in close proximity of the target user.

Sensing Modality Selection: There are mainly two reasons why EgoADL combines 3 modalities, *i.e*. the motion sensor signals, wireless sensing signals and audio, to sense users' ADLs.

i). Availability on existing commodity devices: These 3 non-visual sensors are widely equipped on mobile/wearable devices, including smartphones and smartwatches, which can be easily repurposed as an egocentric sensor hub



(a) Action top-1 mAP. (b) Object top-1 mAP.

Fig. 3. Venn diagram visualizes the advantages of each modality. For each action/object class, if a modality achieves comparable top-1 mAP (within 15%) with the best modality, we plot this class in the intersection of two circles representing these two modalities. Since the results are trained on single-modality data, there may be some overfitting in this visualization (*e.g.*, Wi-Fi is good at recognizing "mouse" and "egg").

to log ADL. In particular, the Wi-Fi CSI can be collected on such devices [46] but is heavily underutilized as a potential sensing modality.

ii). Complementary advantages of each modality: To understand the *complementary advantages* of different modalities, we conduct experiments to recognize the ADL by using each single modality data and the DNN model proposed in Sec 5. We use the balanced dataset with labels (Sec. 3), with a $7 : 1.5 : 1.5$ split among training, validation, and testing set. Fig. 3 visualizes the *action* and *object* categories with $> 60\%$ mAP for at least one single modality. We summarize our insights as follows:

• In-pocket motion sensor easily captures ambulatory actions of the leg, *e.g.*, "sitting in the chair", "lying on the bed", *etc.*, but cannot easily capture whole-body motion or object interaction.

• Wireless sensing signals, *i.e.* Wi-Fi CSI, can recognize certain full-body motion and interactions with ambient environment, like "opening the door", "opening the window", *etc.*, but it falls short in discriminating fine-grained activities.

• Audio sensing can easily recognize ADL with unique sound events, like "coughing", "operating vacuum", "brushing teeth", *etc.*, but can hardly identify those with weak or similar sounds.

EgoADL aims to approach near-vision sensing resolution by synergizing the complementary advantages of the non-visual modalities.

5 EGOADL SUPERVISED LEARNING

5.1 Problem Formulation

Most of the traditional DNN-based ADL analysis models are designed to perform *classification* [3, 14]. The key limitation is that they have to prescribe a known set of ADL, which falls short of extensibility when new ADL of interest emerge. Besides, such models only classify the ADL as integer IDs which do not fully utilize the label semantics from the natural language level [49].

In contrast, we propose a solution that formulates the problem as a sequence-to-sequence (seq-2-seq) task. Our approach encodes multi-modal sensory features and decodes them as the label name semantics using a sequence of

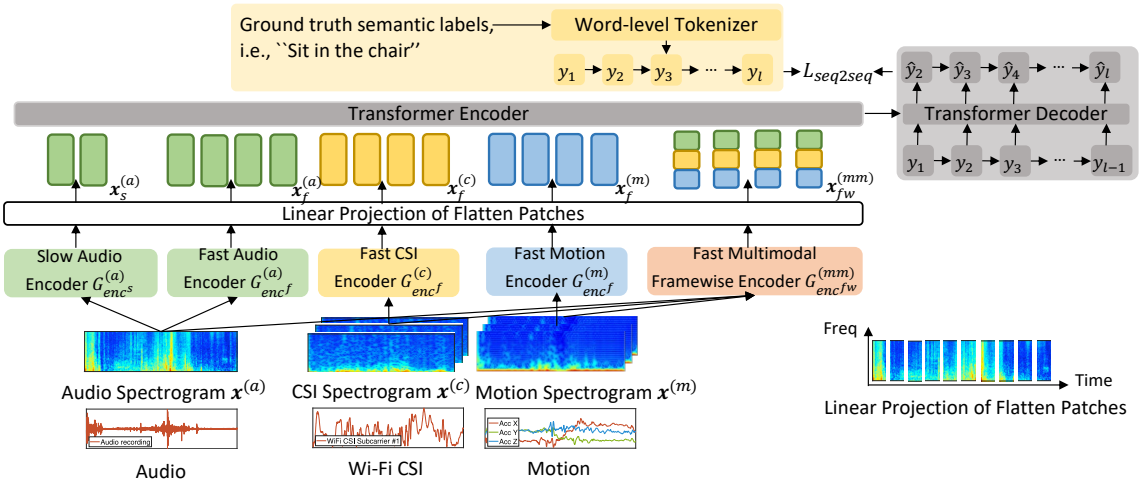


Fig. 4. EgoADL Multi-Modal Frame-Wise Slow-Fast (MMFWSF) transformer design with modality-specific encoders and a transformer-based seq-2-seq model for translating sensory feature into natural language.

words, rather than simple classification labels. This enables us to capture more nuanced and precise information about ADL. Specifically, our goal is to translate synchronized signals $\mathbf{x} = \{\mathbf{x}^{(a)}, \mathbf{x}^{(c)}, \mathbf{x}^{(m)}\}$, where $\mathbf{x}^{(a)}$, $\mathbf{x}^{(c)}$, and $\mathbf{x}^{(m)}$ correspond to audio recordings, Wi-Fi CSI, and motion sensor signals, respectively, into semantic labels $\mathbf{y}$ for ADL in the form of a sequence of words.

5.2 Input and Output Design

As shown in Fig. 4, we first discuss the input feature design.

Audio Recordings: We extract the T-F log mel spectrograms with a sampling rate of 32 kHz, Hamming window size of 31.25 ms, hop size of 10 ms and 64 mel filter banks [32].

Wi-Fi CSI: We first normalize each CSI subcarrier by mean-std normalization, and then extract the Doppler spectrogram applying STFT on the time-domain sequence of CSI values across subcarriers. The STFT uses a Hamming window size of 200 ms, a hop size of 10 ms and FFT size of 128 at 400 Hz sampling rate. The window size is much larger than audio because the Doppler feature caused by human motion exhibits lower frequency.

Motion Sensor: We utilize the linear acceleration (excluding the effect of gravitational force) as the raw input data [50], apply mean-std Z-normalization [37, 41], and then extract the motion sensor spectrogram via STFT on the time domain output of each sensor channel [23, 50, 51]. Such a preprocessing pipeline, as recommended by existing studies in human activity recognition using motion sensors [23, 37, 41, 50, 51], not only enriches the time-frequency domain representation but also helps to reduces the impact of variations in device orientation. Note that we use the *same hop size* of 10 ms for all 3 modalities to ensure *feature alignment* in the time domain. For each modality, the input of the DNN model is a sequence of frequency spectrogram, $\mathbf{x} = \{x_1, x_2, .., x_t\}$, where t is the length of the spectrogram in the time domain. For each timestamp i, $\mathbf{x}_i = \{x_i^{(a)}, x_i^{(c)}, x_i^{(m)}\}$ represent the audio log-mel spectrogram, Wi-Fi CSI Doppler spectrogram, and motion sensor spectrogram, respectively.

EgoADL outputs a natural language text description of the ADL. As shown in Fig. 4, the output can be represented as a sequence of tokens. Unlike conventional ASR which uses the subword-level or character-level tokenizer corresponding to the phonological units, we adopt a *word-level tokenizer*, corresponding to the basic unit in ADL semantics. For example, for the "chop vegetables" activity, we expect EgoADL to recognize the action

"chop" and the object "vegetables", and generate a sequence of tokens "chop vegetables". Our tokenizer dictionary contains $1,000$ frequently-used words for describing ADL from EgoADL dataset labeling (see the data labeling discussion in Sec. 3). Finally, the output sequence is $\mathbf{Y} = \{y_1, y_2, ..., y_l\}$, where l denotes the number of words in the output text.

5.3 MMFWSF Transformer

Design Principles: Next, we introduce our Multi-Modal Frame-Wise Slow-Fast (MMFWSF) transformer. Compared to existing methods that fuse traditional modalities such as audio, video, and text [52], the multi-modal fusion in EgoADL possesses the following unique properties which we can leverage:

• Audio, Wi-Fi CSI, and motion sensor data all have complementary advantages when it comes to capturing *fast-changing motion* [53], including both one-shot event-based ADL such as "sit down" and "stand up", and periodical state-based ADL such as "walking" and "clapping".

• In contrast to Wi-Fi CSI and motion sensor data, which mainly capture *fast-changing motion*, audio data has a distinct advantage in capturing *continuous scene sounds* with specific frequencies, such as "operating a vacuum," "shaving bread," and "brushing teeth".

• Another essential aspect of our approach is the use of *framewise alignment* between multiple modalities. Since no single modality can fully capture all aspects of ADL, there may always exist modality missing for different behaviors. For instance, in Fig. 2, the "Drinking" behavior is only captured by Wi-Fi CSI. However, the missing modality can provide additional supervision, helping us determine whether a specific modality can capture a specific behavior and what the cross-modal cues are at the frame level.

MMFWSF Transformer design: We design our multi-modal fusion and transformer-based seq-to-seq model to fully utilize the aforementioned insights.

First, to leverage the complementary advantages of each modality for capturing *fast-changing motion*, we design *fast pathway* CNN-based encoder $G^{(a)}_{enc^f}$, $G^{(c)}_{enc^f}$ and $G^{(m)}_{enc^f}$ for audio, Wi-Fi CSI and motion sensor, respectively, to achieve a fine feature representation along the temporal dimension. The basic idea is to design CNN encoders with a small temporal stride of τ, resulting in the length of frequency spectrogram t/τ, to guarantee high temporal resolution. The default value is $\tau = 4$ in our experiments. The other parameters of our CNN-based encoders are shown in Sec. 10. Finally, with the *fast pathway* CNN-based encoder, the feature representations of audio ($j = a$), Wi-Fi CSI ($j = c$) and motion sensor ($j = m$) are

$$\mathbf{x}^{(j)}_f = \{x^{(j)}_{1^f}, x^{(j)}_{2^f}, ..., x^{(j)}_{t/\tau^f}\} = G^{(j)}_{enc^f}(\mathbf{x}^{(j)}) \tag{1}$$

where $x^{(j)}_{i^f} \in \mathbb{R}^{C\times F}$ ($i = 1 \sim t/\tau$), C is the number of channels, and F is the number of frequency bins after *fast pathway* CNN-based encoder. The input of the transformer is $\mathbf{x}^{(mm)}_f = (\mathbf{x}^{(a)}_f, \mathbf{x}^{(c)}_f, \mathbf{x}^{(s)}_f)$.

Second, we design an additional audio *slow pathway* with a CNN-based encoder $G^{(a)}_{enc^s}$ to learn the feature representation for *continuous scene sounds*. Basically, we use a large temporal stride of $\alpha\tau$, where $\alpha > 1$ to focus on learning frequency semantics [54]. We set $\alpha = 16$ as default. And the feature representation after the audio *slow pathway* CNN-based encoder is

$$\mathbf{x}^{(a)}_s = \{x^{(a)}_{1^s}, x^{(a)}_{2^s}, ..., x^{(a)}_{\frac{t}{\alpha\tau}{}^s}\} = G^{(a)}_{enc^s}(\mathbf{x}^{(a)}) \tag{2}$$

where $x^{(a)}_{i^s} \in \mathbb{R}^{C\times F}$ ($i = 1 \sim \frac{t}{\alpha\tau}$).

Third, to learn the feature representation of *framewise alignment* between multiple modalities, we propose to further fuse the multi-modal sensory data at the frame level. Note that to make sure the final input sequence of feature representation can be the input of the transformer-based seq-2-seq model, we need to make sure that each feature representation is of the same size. Therefore, we train another 3 *fast pathway* CNN-based encoders

$G^{(a)}_{enc^{fw}}$, $G^{(c)}_{enc^{fw}}$ and $G^{(m)}_{enc^{fw}}$ for each modality with $C/3$ channels, so that the feature representation of each modality $x^{(j)}_{i^{fw}} \in \mathbb{R}^{\frac{C}{3}\times F}$, where $i = 1 \sim t/\tau$. Finally, the feature representation of *frame-wise* fusion is

$$x^{(mm)}_{i^{fw}} = concat(G^{(a)}_{enc^{fw}}(\mathbf{x}^{(a)}), G^{(c)}_{enc^{fw}}(\mathbf{x}^{(c)}), G^{(m)}_{enc^{fw}}(\mathbf{x}^{(m)}), dim = i) \tag{3}$$

where $x^{(mm)}_{i^{fw}} \in \mathbb{R}^{C\times F}$, and $\mathbf{x}^{(mm)}_{f^w} = \{x^{(mm)}_{1^{fw}}, x^{(mm)}_{2^{fw}}, ..., x^{(mm)}_{t/\tau^{fw}}\}$.

Note that to make sure all these representations can fit into the transformer, we intentionally enforce the feature representation of a single frame as $\mathbb{R}^{C\times F}$. Thus, inspired by vision transformer [55] and audio spectrogram transformer [56], our MMFWSF transformer model can concatenate the sequence of $\mathbf{x}^{(a)}_s, \mathbf{x}^{(mm)}_f, \mathbf{x}^{(mm)}_{fw}$ along the temporal dimension and then use the linear projection of flatten patches along the channel and frequency dimension to fit into the transformer model as shown in Fig. 4. Our transformer model contains 12 encoder layers and 6 decoder layers. All the detailed DNN layer designs are shown in Methodological Transparency & Reproducibility Appendix (META) (Sec. 10).

5.4 Training Strategy

Training loss design: We use the seq-to-seq loss based on the autoregressive decoder, where the previous output is fed back into the input, to decode the input sequence to the output token sequence. In the testing phase, the predicted word label $\hat{y}_i$ and the hidden state h_i of the decoder at step i can be updated as $\hat{y}_i, h_i = \text{Decoder}(h_{i-1}, \hat{y}_{i-1}, c_i)$, where c_i is the context vector generate by the encoder. In the training phase, we use teacher forcing methods to train the model, which means $\hat{y}_i, h_i = \text{Decoder}(h_{i-1}, y_{i-1}, c_i)$, where y_{i-1} is the last token of the ground truth label. The objective is to minimize the corresponding cross entropy loss l_{seq2seq}.

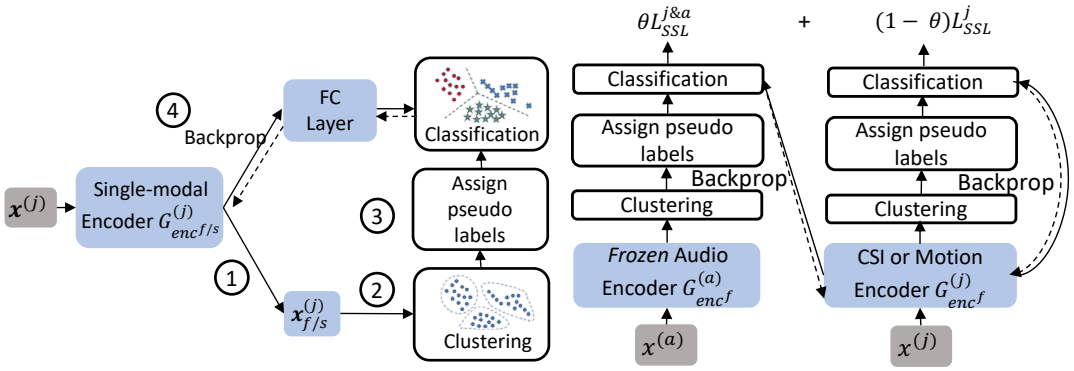
Beam search during testing: In the testing phase, we use beam search, a widely adopted method in NLP and ASR [57], to search for the top-K candidate sequences. For each step, we predict the K most promising next tokens, and then feed these K alternatives into the decoder to select the best K hypothesis at the next step iteratively.

6 EGOADL SELF-SUPERVISED LEARNING

We have conducted experiments on the supervised model (Sec. 5) and identified limitations and potential pathways towards a self-supervised EgoADL model design. We found that while the supervised model performed well on a balanced labeled dataset, it struggled with *overfitting* and *lack of generalization and extensibility*, especially for infrequent ADL, due to limited ground truth labels. Scaling up the dataset requires an enormous amount of labeling effort, especially for non-visual sensors. On the other hand, in contrast to vision-based setups that require users to wear multiple devices (*e.g.*, head-mounted or chest-mounted cameras [3, 5]), collecting large-scale unlabeled data is much easier with EgoADL, as it is non-intrusive and only requires users to carry a smartphone in pocket. We harness this unique advantage through an SSL model that can improve *(i)* accuracy, *(ii)* extensibility for few-shot ADL with limited labels, and *(iii)* generalization across different ADL, users, and environments. *Our SSL model trains more generic encoders (see Fig. 4) by leveraging intrinsic supervisory signals within unlabeled data, and by distilling knowledge of human behavioral logic from external audio datasets.*

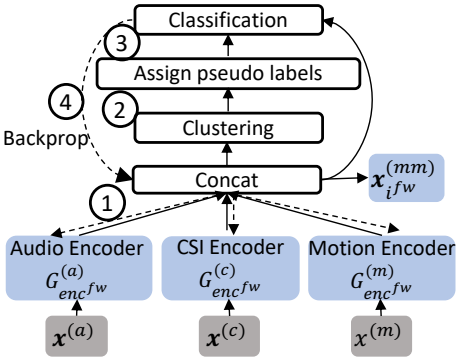
6.1 Single-modal Self-Supervised Deep Clustering

We first introduce a self-supervised clustering method to learn the single-modal encoders from unlabeled data. Inspired by vision-based SSL [58], our deep clustering method takes the input feature $\mathbf{x}^{(j)}$ from a single modality (j) as input. For each epoch, the training procedure of the deep clustering method follows the steps ①–④ in Fig. 5(a). First, the single-modal encoder generates the feature representation $G_{enc^{(j)}}(\mathbf{x}^{(j)})$. Second, an unsuperivsed clustering method is applied to all the feature representations from EgoADL's unlabeled training data. Then, we assign *pseudo labels*, *i.e.*, the cluster index of each resulting cluster, to the corresponding data. Finally, we append



(a) Single-modal SSL. EgoADL uses the deep clustering to assign pseudo labels for unlabeled training data to train the encoders $G^{(j)}_{encf/s}$, where "j" here represents to audio, motion sensor or Wi-Fi CSI.

(b) Single-modal deep clustering via cross-modal self-supervision. EgoADL utilizes the pseudo labels generated by pretrained audio encoder $G^{(a)}_{encf}$, to train the encoders $G^{(j)}_{encf}$, where "j" represents motion sensor or Wi-Fi CSI.



(c) Cross-modal deep clustering. EgoADL fuses the modalities and learns their correspondence by training the concatenated representations of the 3 modalities encoders ($G^{(a)}_{encfw}$, $G^{(c)}_{encfw}$ and $G^{(m)}_{encfw}$)

Fig. 5. EgoADL SSL methods.

fully-connected layers to the encoders to classify the feature representations to their corresponding pseudo labels and use the back propagation training algorithm to update the parameters in the encoders. The trained single-modal feature representations will be used in our later cross-modal SSL stage.

In our implementation, we use the entire 100-hour unlabeled EgoADL dataset to train the encoders for each modality separately, *i.e.* $G^{(a)}_{encf}$, $G^{(a)}_{encs}$, $G^{(c)}_{encf}$ and $G^{(m)}_{encf}$. We choose K-means and 3 fully-connected layers with ReLU activation functions as the default clustering and classification method, respectively. To optimize the number of clusters k for K-means, we conducted the end-to-end experiments by varying k on a logarithmic scale during the hyperparameter tuning phase. These experiments, conducted within the context of the EgoADL dataset, which encompasses 221 distinct ADL, indicate that a k value within the range of 10^2 to 10^3 yielded near-optimal performance. Consequently, we selected $k = 200$ as our default configuration.

6.2 Cross-Modal Self-Supervised Deep Clustering

Second, we also leverage cross-modal self-supervisory to enhance the single-modal deep clustering. More specifically, we leverage the external audio dataset, *i.e.* AudioSet [17], a massive dataset with more than 5,000 hours of labeled audio recordings for 527 event classes from YouTube videos, to first train generic audio encoders, *i.e.* $G^{(a)}_{enc^f}$ and $G^{(a)}_{enc_s}$ [32]. Then, we use the audio pseudo labels from the same frame to train the Wi-Fi CSI $G^{(c)}_{enc^f}$ and motion sensor encoders $G^{(m)}_{enc^f}$. The use of audio pseudo labels offers several benefits. First, such labels from the pre-trained audio feature embeddings help prevent the Wi-Fi CSI and motion sensor models from overfitting to particular user characteristics, such as Wi-Fi transmission locations and sensor orientations. Second, they harness the broad sensing capabilities of audio, capturing both the event-based sound and continuous scene sounds for comprehensive supervision (see Sec. 5.3). Lastly, these labels can accelerate the training of deep clustering models for Wi-Fi CSI and motion sensor data. Fig. 5(b) shows the training procedure of this mechanism. $L^{j\&a}_{SSL}$ represents the loss when we use the audio pseudo labels to classify the Wi-Fi CSI or motion sensor feature representations. The final loss $L^{j_{all}}_{SSL} = \theta L^{j\&a}_{SSL} + (1-\theta)L^{j}_{SSL}$. Our evaluation results show that without using the pre-trained audio feature embeddings will result in reduction in both accuracy and generalization performance (see Sec. 8.2).

After training the single-modal encoders for each of the 3 modalities, we employ cross-modal deep clustering to fuse the modalities and learn their *correspondence*. As shown in Fig. 5(c), we concatenate the representations of the 3 modalities encoders ($G^{(a)}_{enc^{fw}}$, $G^{(c)}_{enc^{fw}}$ and $G^{(m)}_{enc^{fw}}$) from the same frame to perform the training of the cross-modal deep clustering. The training procedure of the cross-modal deep clustering follows the same steps as the single-modal case as discussed in Sec. 6.1. After finishing cross-modal deep clustering training, the resulting fast multimodal framewise encoder generates the multi-modal feature representation $\mathbf{x}^{(mm)}_{if_w}$. In contrast to training single-modal deep clustering and directly concatenating the feature representations of 3 modalities, cross-modal deep clustering tries to automatically learn the co-occurring features from different modalities, which helps the DNN model understand the correspondence from different modalities.

Finally, after the SSL training, we use the cross-modal SSL models to replace the supervised encoders (see Fig. 4). We then use the small labeled EgoADL dataset to train the entire model end to end.

7 KNOWLEDGE DISTILLATION FROM NATURAL LANGUAGE LABELS

Given that the output of EgoADL materializes in the form of natural language text, it opens up opportunities for leveraging the inherent semantics of natural language labels to enhance performance even further. In this section, we embark on two approaches. First, we introduce a label refinement mechanism aimed at ensuring that the granularity of labels remains congruent with the capabilities of the sensors. Subsequently, we put forth the idea of utilizing pre-existing natural language text to cultivate contextual reasoning for sequences of ADLs.

7.1 Label Refinement

To understand the limits of non-visual sensors, we compare EgoADL with egocentric vision-based methods (see Sec. 8.4). While most ADL show reasonable accuracy, we observed that several ADL had significantly lower accuracy. This is because we annotated and labeled the EgoADL dataset by manually observing the egocentric video and audio. Nevertheless, as we discussed in the previous sections, the non-visual sensors (audio, motion sensor, and wireless sensing) have limited resolution compared to visual sensors, which may have impacted the labeling accuracy. To better understand the limits of EgoADL, we propose to refine the labeling by merging ADL that are difficult to distinguish using EgoADL non-visual sensors. Our label refinement involves three steps: *i)* ranking ADL based on mean average precision (mAP), *ii)* visualizing the confusion matrix of ADL with less than ϵ, and *iii)* merging actions or objects in ADL based on both the confusion matrix and our knowledge. We have empirically set ϵ to 30%, predicated on the observation that EgoADL's overall system mAP is approximately 60%,

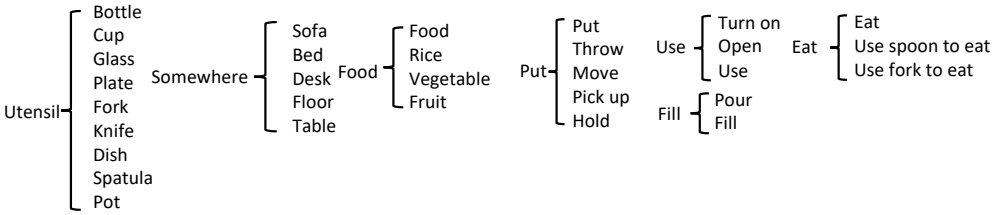


Fig. 6. Representative label refinement.

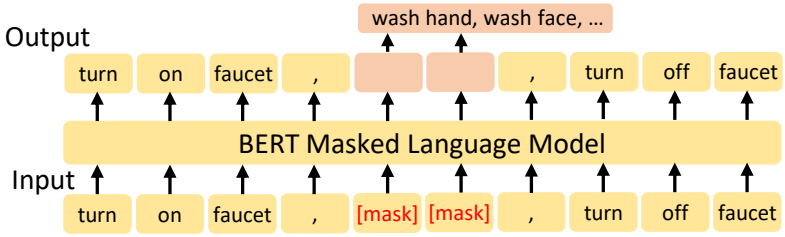


Fig. 7. BERT-based EgoLM design, which learn the contextual information of ADL.

and its discriminative power is significantly reduced for ADL with an mAP below 30%—a scenario applicable to roughly 30/105 of the ADL. We summarize the representative label refinements in Fig. 6. The refined labels consist of 75 frequently-used ADL, 38 object interactions, 41 actions, 29 state-based ADL, and 46 event-based ADL (see Sec. 10). When employing these labels to assess the performance of EgoADL, the achieved results notably surpass those attained from labels derived from egocentric video and audio sources. This disparity is attributed to the fact that the former set of labels encapsulates the intrinsic capacities of the sensors. Note that our current method requires to refine the labels manually. To further understand the boundary of EgoADL, we need to design an automatic label refinement solution for each modality and multimodal fusion, as well as fine-tuning the hyperparameter ϵ. This will be left as our future work.

7.2 Distilling Contextual Information from Text

The above models only process the short segments of sensor data for single human behavior (with each segment varying in length but not exceeding 10 seconds, as detailed in Sec. 3). Longer segments of input may provide contextual information to boost the performance. But it will dramatically increase model complexity, and lead to severe overfitting due to the limited multi-modal dataset. In EgoADL, we harness existing natural language text to learn the contextual reasoning instead.

Inspired by the language model in NLP [59], our idea is to learn a "contextual language model" for ADL, referred to as *EgoLM*. Unlike traditional NLP which calculates the probability distribution over sequences of words, EgoLM outputs the probability of a given sequence of ADL via natural language text. As shown in Fig. 7, EgoLM is fine tuned from the celebrated Bidirectional Encoder Representations from Transformers (BERT), a transformer-based NLP model pre-trained by Google [59]. In the training phase, EgoLM takes the text description of a sequence of 30 s ADL as input, and uses a comma to mark the end of the last ADL. During the training phase of EgoLM, we experimented with various masking strategies, including masking a word-level token or an entire ADL, and then we utilize the contextual information from surrounding ADLs to predict the masked elements. Unlike the traditional BERT approach of word-level masking, we choose to mask an entire ADL (as shown in Fig. 7), the strategy that resulted in the most optimal performance (see Sec. 8.5).

To make EgoLM more general, we collect the training text corpus from not only the EgoADL dataset but also existing video-based domestic ADL datasets, including Charades [13], CharadesEgo [4], EPIC-KITCHENS [3]

and EGTEA-GAZE [60]. We extract the text corresponding to the sequence of ADL within a 30 s period from these datasets, which typically contains $3 \sim 8$ ADL. Note that most of the datasets segment the ADL in 10 s units. However, the state-based ADL (see Sec. 3) typically last for more than 10 s, and the same ADL labels may appear consecutively. We thus merge these ADL to prevent replication of state-based ADL in the sequence. Overall, our EgoLM text corpus has $3,000$ and $34,000$ sequences of ADL from EgoADL dataset and 4 external datasets. We first use the whole EgoLM text corpus to train the original BERT model [59], and then fine tune it with the EgoADL dataset.

EgoLM can be applied to any EgoADL models, that we discussed previously, in the testing phase. We first use EgoADL model to generate the potential prediction of each single ADL via beam search, and save the score of the loss function from the EgoADL DNN model. And then, we apply a second round of beam search to combine the EgoADL loss with the language model loss $l_{\text{all}} = \gamma l_{\text{EgoADL}} + (1-\gamma) l_{\text{EgoLM}}$, where l_{EgoADL} learns the information from the raw data in the current segment, and l_{EgoLM} learns the contextual information in a long period (30 s) before current segment. Finally, we select the ADL with the lowest score of l_{all} to output the top-K prediction.

8 IMPLEMENTATION AND EXPERIMENTAL EVALUATION

8.1 EgoADL Implementation and Evaluation Metrics

DNN Implementation: The EgoADL DNN model is implemented in PyTorch. For training, the self-supervised feature embedding DNN models for 3 modalities are first trained separately using single-modal SSL and then jointly using cross-modal SSL, as discussed in Sec 6. Next, we freeze these models and train the end-to-end seq2seq model as discussed in Sec. 4. We use the Adam optimizer with a $1e-4$ initial learning rate followed by annealing. The current EgoADL implementation has 421.6 M parameters in total.

Evaluation Metrics: We evaluate EgoADL using classwise mAP metrics and captioning metrics which are adopted in egocentric video-based ADL recognition and captioning [61]. We measure the mAP for the aforementioned "action" and "object" categories, along with "state-based ADL" and "event-based ADL" (Sec. 3), and an "overall ADL" (aka. "ADL") which is the superset of the 4 categories. Our EgoADL dataset contains 35 state-based and 186 event-based ADL classes, the former typically have more labeled data samples because state-based ADL last longer. We use two captioning metrics to measure the similarity between predicted and reference captions [62], *i.e.* BLEU and SPICE, which are based on n-gram overlapping and scene graph similarity, respectively.

8.2 Micro Benchmark Analysis of EgoADL Supervised Learning Model

We conduct an ablation study to compare the EgoADL design across different modality fusions. As shown in Tab. 2, we evaluate EgoADL in 7 settings by using a single modality and combining multiple modalities by using supervised learning models. 5 different methods are evaluated: *(i)* "Fast-only" for single modality, *(ii)* "Modalwise" where the multi-modal features are fused along the modality dimension *(iii)* "Framewise" where the multi-modal features are fused along the frame dimension, *(iv)* "MMFW" where the multi-modal features are fused along both modality and frame dimensions without audio slow pathway (see Sec. 5.3), *(v)* "MMFWSF" represents to "MMFW" with audio slow pathway (Sec. 5.3). For a fair comparison, all the methods are trained using a balanced dataset with $2,500$ labeled samples. We employed an $8:2$ split between training and validation with 5-fold cross validation. For testing, we use the remaining unbalanced $2,800$ samples. It is important to note that this unbalanced testing set does not skew our final results, as all outcomes are reported on a classwise basis (average acorss different ADL). The distribution of samples from all 10 users is balanced across the training, validation, and testing sets.

Performance gain due to multiple modalities: As shown in Tab. 2, "Audio" is the most informative modality among these three modalities. Compared to audio-only solution, the multi-modal fusion achieves an overall 13.5% and 16.6% improvement for top-1 and top-5 overall ADL mAP, respectively. Fig. 8 demonstrates the performance gain introduced by the EgoADL multi-modal fusion design. With the motion sensor located at users' trouser pocket,

Modal	Methods	Top-1 mAP (%)					Top-5 mAP					Captioning	
		ADL	A	O	S	E	ADL	A	O	S	E	BLEU	SPICE
Audio	Fast-only	44.1	56.0	54.3	71.3	31.7	62.6	75.3	67.1	82.4	53.5	0.42	0.40
Wi-Fi	Fast-only	25.9	37.0	36.1	49.7	15.0	51.6	71.9	57.8	80.4	38.3	0.29	0.28
Motion	Fast-only	23.0	32.8	30.0	41.6	14.5	44.5	53.9	51.2	61.6	36.6	0.28	0.23
Audio + Motion	Modalwise	50.7	61.1	61.0	73.0	40.5	67.2	80.8	81.7	90.5	56.5	0.48	0.46
	Framewise	46.4	59.1	55.8	71.7	34.8	67.1	79.2	77.1	87.0	57.9	0.46	0.44
	MMFW	51.5	63.8	61.9	73.8	41.3	67.4	80.7	78.3	89.4	57.3	0.48	0.45
	MMFWSF	52.1	64.0	62.0	75.0	41.6	67.8	81.7	78.5	90.2	57.5	0.49	0.45
Audio + Wi-Fi	Modalwise	47.4	58.5	59.2	71.3	36.4	67.8	78.5	81.3	90.5	57.3	0.46	0.43
	Framewise	47.9	57.8	56.3	72.8	36.5	64.0	74.5	78.6	86.5	53.7	0.44	0.43
	MMFW	48.6	62.6	62.9	73.1	37.4	70.1	82.3	81.6	90.2	60.8	0.49	0.46
	MMFWSF	49.2	62.9	63.5	74.5	37.5	71.0	82.5	81.9	90.5	62.1	0.49	0.46
Wi-Fi + Motion	Modalwise	40.2	51.9	46.9	66.2	28.2	64.4	78.1	72.1	84.2	55.2	0.43	0.41
	Framewise	41.1	50.6	46.9	66.8	29.3	64.6	78.9	72.6	84.5	55.5	0.43	0.41
	MMFW	41.3	53.1	48.0	66.8	29.5	65.0	78.2	73.1	83.9	56.1	0.43	0.41
Audio+ Motion+ Wi-Fi	Modalwise	51.5	63.4	62.7	73.8	41.2	76.1	85.2	86.9	90.8	68.3	0.52	0.51
	Framewise	53.8	66.7	64.1	76.7	43.2	74.7	86.5	87.1	90.6	67.4	0.53	0.52
	MMFW	54.6 ± 2.1	67.6 ± 1.8	**66.2** ± 1.7	76.0 ± 1.5	44.8 ± 2.8	76.6 ± 1.9	86.6 ± 1.2	88.0 ± 1.5	90.7 ± 1.2	70.1 ± 2.5	0.54 ± 0.02	0.52 ± 0.02
	MMFWSF	**55.3** ± 2.0	**68.1** ± 1.5	65.8 ± 1.9	**77.0** ± 1.3	**45.3** ± 2.4	**78.1** ± 1.2	**87.8** ± 0.9	**88.4** ± 1.1	**92.5** ± 1.0	**71.5** ± 1.4	**0.56** ± 0.02	**0.52** ± 0.01
	Refine	**68.2** ± 1.5	**70.4** ± 1.6	**72.2** ± 1.4	**83.2** ± 0.8	**55.9** ± 2.0	**87.3** ± 1.3	**90.1** ± 1.9	**92.3** ± 1.4	**94.8** ± 1.0	**82.7** ± 2.2	**0.65** ± 0.01	**0.63** ± 0.01

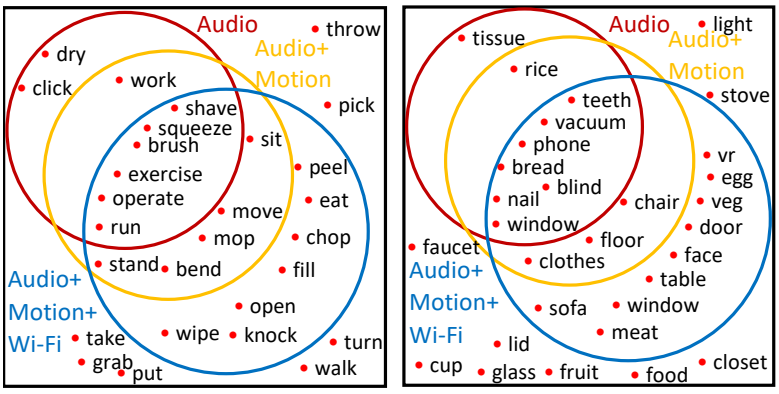
Table 2. EgoADL micro benchmark. We benchmark *5 categories* of ADLs. “ADL”, “A”, “O”, “S”, “E” represent to “Overall ADL”, “Action”, “Object”, “State” and “Event”, respectively (see Sec. 3). “Refine” represents the case where the non-ambiguous labels for non-visual sensors, as discussed in Sec. 8.4. The number below the mAP result is the 95% confidence interval for the mAP.

EgoADL is able to recognize more actions and objects related to lower body, including “Sit in a chair”, “Bend down”, “Mop the floor”, etc. Wi-Fi CSI further characterizes the full-body motions with large environment changes. ADL, like “Open/close the door”, “Open/close the window”, etc., are easily recognized through EgoADL’s multi-modal fusion. Besides, we also find that, compared to the “Audio” which is effective in identifying *event-based ADL*, both “Motion” and “Wi-Fi” have advantages in recognizing *state-based ADL*, which may involve periodic motions. As shown in Fig. 8, the multi-modal fusion only incurs slightly lower accuracy for few ADL, like “Dry hand”, “Grab tissue”, etc.

Performance gain due to MMFWSF fusion design: As shown in Tab. 2, compared to traditional modalwise/framewise fusion algorithms, our MMFWSF fusion achieves better performance for all kinds of ADL. By utilizing the complementary modalities of different modalities, EgoADL further pushes the limits to achieve 55.3% and 78.1% top-1 and top-5 overall ADL mAP. Further, the 95% confidence intervals for the mAP across all categories are within ±2.5%, indicating the consistent performance of EgoADL.

8.3 Accuracy, Generalization and Extensibility of EgoADL SSL Model

In this section, we evaluate the EgoADL models using various SSL training methods (Sec. 6), focusing on improvements in accuracy, generalization, and extensibility. The EgoADL models are trained by 3 different training approaches: *i).*“W/o” for without SSL, *ii).*“SM” for single-modal SSL (Sec. 6.1), *iii).* “CM” for cross-modal SSL (Sec. 6.2). Initially, models are pretrained using 100 hours of unlabeled data from 20 additional subjects. This is followed by fine-tuning the models using a balanced dataset comprising 2,500 samples for training, and an



(a) Actions. (b) Objects.

Fig. 8. Venn diagram visualizes the advantages and limitations of EgoADL multi-modal fusion. Actions/ objects with $> 80\%$ top-1 mAP are in the circle for different modality fusions

unbalanced dataset with 2,800 samples for testing, with the same 5-fold cross validation setup described in the previous section.

Accuracy gain due to SSL: Table 6 shows that leveraging SSL into EgoADL yields a marginal Top-1 mAP enhancement across various ADL categories when compared to the non-SSL EgoADL models. Notably, the adoption of cross-modal SSL (Sec. 6.2) allows EgoADL to capitalize on audio pseudo labels derived from pre-trained audio feature embeddings, culminating in a further 2.9% mAP increment for the overall ADL.

Generalization gain due to SSL: For generalization evaluation, the models are trained and tested using a leave-one-out cross-validation approach, to evaluate on unseen users (UU) and unseen environments (UE). Table 4 shows the top-1 mAP. The baseline model shows poor generalization, with 35.3%, 27.7% top-1 "ADL" mAP and large 95% confidence interval for UU and UU+UE respectively. In contrast, with the "Cross-modal SSL" design, EgoADL achieves 47.7% and 47.1% in the same metrics with a smaller 95% confidence interval. Additionally, when we fine-tune the EgoADL model with 200 labeled samples for a specific user ("Personalized fine tuning" in Tab. 4), the model reaches 77.1% and 91.5% top-1 mAP for "ADL" and "state" respectively.

Extensibility gain due to SSL: For extensibility assessment, we focus on 116 event-based "tail" ADL, which refers to the ADL with less than 15 data samples in our labeled dataset. Training and validation utilize a balanced set of 2,500 labeled samples, supplemented by 1,000 "tail" ADL samples. In total, 3,500 out of 7,000 labeled samples form the training set while the remaining 3,500 unbalanced labeled samples form the test set. As shown in Tab. 5, our SSL approach yields substantial increases in mAP for these "tail" classes: 39.8% top-1 and 62.3% top-5. Further fine-tuning with an additional 3 samples per "tail" class significantly boosts these metrics to 60.2% top-1 and 74.1% top-5.

8.4 Evaluating the Limits of Non-Visual Sensors

To understand the limits of non-visual sensors, we compare EgoADL with egocentric vision-based methods. We reimplemented the SOTA methods, *i.e.* Ego-exo [63], on two egocentric vision datasets, *i.e.* Charades-Ego [4] and GTEA-GAZE [6]. The ADL sets differ across datasets. For a fair comparison, we only consider the classes which have the same labels in both the egocentric vision and EgoADL datasets (see detailed ADL sets in Fig. 19). For EgoADL, we use the model trained in the previous section. For Ego-exo [63], we employed models fine-tuned with the Charades-Ego [4] and GTEA-GAZE [6] datasets. The models are evaluated not only on the egocentric video

	Top-1 (%)					Top-5 (%)				
	ADL	A	O	S	E	ADL	A	O	S	E
W/o	55.3 ± 2.0	68.1 ± 1.5	65.8 ± 1.9	77.0 ± 1.3	45.3 ± 2.4	78.1 ± 1.2	87.8 ± 0.9	88.4 ± 1.1	92.5 ± 1.0	71.5 ± 1.4
SM	56.3 ± 1.3	67.5 ± 1.0	66.9 ± 1.2	77.5 ± 0.9	46.6 ± 1.6	78.5 ± 0.8	87.1 ± 1.0	89.5 ± 1.3	93.0 ± 0.7	71.8 ± 1.3
CM	59.2 ± 1.0	68.5 ± 1.2	68.9 ± 1.4	79.5 ± 0.8	49.8 ± 1.1	79.8 ± 1.0	87.6 ± 1.1	90.6 ± 1.2	92.8 ± 0.6	73.8 ± 1.3

Table 3. Accuracy of EgoADL SSL. “W/o”, “SM” , “CM” represent to “W/o SSL”, “Single-modal SSL”, “Cross-modal SSL”, respectively. “ADL”, “A”, “O”, “S”, “E” represent to “Overall ADL”, “Action”, “Object”, “State” and “Event”, respectively. The number below the mAP result is the 95% confidence interval for the mAP.

	UU (Top-1 mAP %)					UU + UE (Top-1 mAP %)				
	ADL	A	O	S	E	ADL	A	O	S	E
W/o	35.3 ± 6.3	46.3 ± 5.5	36.6 ± 6.9	48.5 ± 3.3	29.3 ± 7.4	27.7 ± 6.6	36.6 ± 5.8	34.5 ± 7.0	41.3 ± 4.1	21.5 ± 8.0
SM	41.9 ± 4.1	52.6 ± 3.7	54.5 ± 4.0	60.3 ± 2.8	33.4 ± 4.8	42.8 ± 4.4	48.5 ± 3.4	52.1 ± 4.0	65.5 ± 3.5	32.2 ± 5.0
CM	47.7 ± 3.5	56.1 ± 3.3	62.5 ± 4.0	69.7 ± 2.9	37.6 ± 4.5	47.1 ± 3.8	55.8 ± 3.2	60.9 ± 3.9	68.5 ± 2.0	37.2 ± 4.0
P	77.3 ± 2.3	88.2 ± 2.1	87.3 ± 1.8	93.0 ± 1.3	70.1 ± 2.7	77.1 ± 2.5	87.9 ± 2.0	87.5 ± 2.4	91.5 ± 1.9	70.5 ± 3.0

Table 4. Generalization of EgoADL SSL. “UU” and “UE” represent to “Unseen user” and “Unseen environment”. “W/o”, “SM” , “CM” represent to “W/o SSL”, “Single-modal SSL”, “Cross-modal SSL”, respectively. “P” represents to “Personalized fine tuning”. “ADL”, “A”, “O”, “S”, “E” represent to “Overall ADL”, “Action”, “Object”, “State” and “Event”, respectively. The number below the mAP result are the 95% confidence interval for the mAP.

	Tail Classes					
	Top-1 (%)			Top-5 (%)		
	E	A	O	E	A	O
W/o	22.8	38.5	41.2	43.2	50.2	53.5
SM	35.6	51.9	55.0	58.2	62.3	65.5
CM	39.8	53.5	58.3	62.3	70.9	72.3
P	60.2	69.3	71.2	74.1	80.2	83.1

Table 5. Extensibility of EgoADL SSL. “W/o”, “SM” , “CM” represent to “W/o SSL”, “Single-modal SSL”, “Cross-modal SSL”, respectively. “P” represents to “Personalized fine tuning”. “E”, “A”, “O” represent to “Event”, “Action”, “Object”.

	# of ADL	Top-1 mAP (%)				
		ADL	A	O	S	E
EgoADL v.s.	40 (105)	48.8	56.3	57.5	46.4	50.4
Charades-Ego v.s.	40 (157)	39.9	41.2	34.5	25.0	46.3
EgoADL vision	40 (105)	31.1	40.5	31.3	21.1	35.3
EgoADL v.s.	40 (105)	47.3	54.8	50.0	65.3	48.3
EGTEA-GAZE v.s.	40 (106)	46.9	57.9	56.4	67.3	42.2
EgoADL vision	40 (105)	40.0	42.6	41.8	40.3	39.8

Table 6. Comparison between EgoADL and egocentric vision. “ADL”, “A”, “O”, “S”, “E” represent to “Overall ADL”, “Action”, “Object”, “State” and “Event”, respectively. “EgoADL vision” means that we evaluate egocentric vision models using the testing video data collected by ourselves, which is originally used for labeling.

dataset but also on the egocentric video data from the EgoADL dataset, which is originally used for labeling. For a fair comparison, all evaluations are conducted in the "UU + UE" scenario.

Methods	Top-1 mAP (%)					Top-5 mAP (%)				
	ADL	A	O	S	E	ADL	A	O	S	E
EgoADL w/o LM	68.2	70.4	72.2	83.2	55.9	87.3	90.1	92.3	94.8	82.7
EgoADL w/ LM, Word Masking	69.9	70.4	72.2	83.8	59.6	88.5	90.9	92.9	94.6	83.2
EgoADL w/ LM, ADL Masking	72.5	75.3	76.5	85.9	65.8	90.8	92.1	93.4	94.5	83.9

Table 7. Performance gain due to EgoLM. “ADL”, “A”, “O”, “S”, “E” represent to “Overall ADL”, “Action”, “Object”, “State” and “Event”, respectively.

Tab. 6 shows the results compared with egocentric vision. We found that testing on the egocentric vision data collected by ourselves (see “EgoADL vision” in Tab. 6) has much worse performance than testing on existing egocentric vision datasets. This is because that egocentric vision datasets require the users to adjust the camera FoV or use specialized cameras with a larger FoV to capture the subject hand and interaction objects. In contrast, our collected egocentric vision data will only be used by data labeling. Therefore, we use a commodity camera with limited FoV. Part of our egocentric vision data is not able to capture the hand motion and interaction object of the subject. This does not affect the data labeling as users can remember what they are doing and label the ADLs based on their memory. However, directly using such data will unfairly degrade the performance of other datasets. We also evaluate egocentric vision using their dataset and comprare the results with EgoADL. *We found that EgoADL achieves comparable performance with vision-based methods for the overlapped classes between the egocentric vision datasets and the EgoADL datasets.* The performance of both highly depends on the label type and granularity (Tab. 6), because they both have unique advantages and limitations, which we summarize as follows:

Adv: EgoADL shows remarkable advantages in recognizing state-based ADL with unique motion patterns or sound events. It achieves a top-1 "state" mAP improvement of 21.4% over Charades-Ego [4], as most state-based ADL cannot be entirely captured by egocentric vision with limited field of view.

Limit1: EgoADL is limited in recognizing ambiguous actions, i.e., actions that are similar to non-visual sensors but can be described by natural language in different ways. For example, human actions, *i.e.* “grab”, “put”, “take”, “hold”, “pick”, “throw”, all involve humans using their hands to fetch something. Our classwise detailed experiments in Fig. 8(a) and Fig. 17 show that EgoADL can only achieve $< 30\%$ mAP on average to distinguish these few actions. Further such actions are not only hard to recognize for non-visual sensors, but also for vision-based methods without detailed contextual information [6].

Limit2: Without vision information, EgoADL is limited to recognize detailed objects. Although audio can capture the specific sound of human-object interaction to distinguish different objects, without vision information, non-visual sensors can only recognize the object with coarse granularity. For instance, when subjects are chopping something in the kitchen, the vision-based methods will be able to recognize the detailed type of objects, like “carrot”, “potato”, “watermelon”, “beef”, *etc.*. However, EgoADL can only recognize the “chop” action but not the type of objects. In EgoADL, we do not label the objects with such granularity. Thus, in Tab. 6, EgoADL even achieves higher performance for “object” mAP as it can recognize many objects outside camera FoV.

8.5 Evaluation on Knowledge Distillation from Natural Language

Label refinement for non-visual sensors: We follow the steps discussed in Sec. 7.1. The refined labels consist of 75 frequently-used ADL, 38 object interactions, 41 actions, 29 state-based ADL, and 46 event-based ADL. As shown in Tab. 2 and Fig. 18, with the label refinement, EgoADL achieves an overall mAP of 68.2%, with 83.2% and 55.9% for state-based and event-based ADL, respectively. This is significantly higher than the labels obtained from egocentric video and audio.

Performance gain due to EgoLM: To evaluate the performance gain due to EgoLM, we need to use continuous testing samples in the time domain since EgoLM takes the prediction text of EgoADL in a long period (30 s) as the input to learn the contextual information. So we use 24 continuous recordings, each lasting 5-min, as the EgoLM

testing dataset (2 hours in total). Tab. 7 summarizes the results. We evaluate the EgoLM with two different masking strategies, masking *i).* a word-level token, and *ii).* an entire ADL. Tab. 7 indicates that masking an entire ADL led to a notably improved performance compared to merely masking word-level tokens. This outcome suggests that masking complete ADL is more effective in enabling the EgoLM model to grasp the contextual relationships integral to ADL. EgoLM gains an additional 4.3% (from 68.2% to 72.5%) top-1 overall ADL mAP for EgoADL, matching the intuition that EgoLM can better understand the contextual information when the original model performance is sufficiently high. Besides, EgoLM is proficient in enhancing the mAP of event-based ADL which tend to have more contextual information.

8.6 Energy Consumption

In this section, we evaluate the energy consumption associated with the sensing capabilities of EgoADL, while a detailed discussion on computational resource consumption is provided in Sec. 9. We conduct a preliminary profiling of the EgoADL sensor data capturing app by using Android's native battery usage measurement. During the measurement, EgoADL collects the audio recordings, Wi-Fi CSI and motion sensor signals in the background with the display off. We found that *EgoADL only consumes less than* 60 *mAh per hour on a Nexus 5 smartphone.* That means EgoADL can work on a Nexus 5 with 2300 mAh battery for about 7.6 days if it continuously records the multi-modal sensing data for 5 hours a day. All the 3 sensor modalities are significantly more energy efficient than a camera (more than 600 mAh per hour) [64], making it a promising potential in practical scenarios.

9 DISCUSSION AND LIMITATIONS

Privacy consideration: EgoADL requires capturing the egocentric audio signals, which may inadvertently include users' daily conversations. However, thanks to the EgoADL DNN design, we can separate the audio branch from the whole model, and calculate the audio feature embedding on-device without uploading raw audio data to the edge/ cloud devices to protect users' privacy. To this end, we can employ the model designed to be deployed on smartphones or edge device [65] as the basic feature embedding network. Another potential solution is to selectively anonymize or mask the speech data [66] from audio recording. These privacy enhancement mechanisms are left for our future exploration.

Generalization of EgoADL dataset: Due to the availability of Wi-Fi sensing, one of the limitations of EgoADL is that the dataset is only collected by a single type of commodity smartphone (*i.e.* Nexus 5). However, it will not significantly affect the generalizability of the dataset because of the following reasons: *i).* There is no limitation imposed on the placement of Wi-Fi AP nor on the manner in which users carry the smartphones in their trouser pockets when collecting the data. Therefore, this leads to greater variability in the data than the type of device used, owing to the variability in Wi-Fi AP locations, which can vary by several meters, and the differences in smartphone Wi-Fi antenna positions, which can vary by several centimeters. *ii).* We focus on the Wi-Fi CSI Doppler shift induced by human motion and environment factors. Given that the Wi-Fi signal's wavelength at a frequency of 5 GHz is approximately 6 cm, the resultant Doppler shift predominantly reflects motions with displacements on the order of tens of centimeters. Therefore, such features are not significantly influenced by variations between different smartphone models. Another limitation of EgoADL dataset is the demographics of participants (see Sec. 10). We hope that, by open-sourcing EgoADL, we can encourage a broader spectrum of participants and researchers to contribute to the EgoADL data collection, thereby enhancing the demographic diversity of our dataset.

Potential Missing Data: Generally, smartphones are capable of continuously capturing both audio and motion sensor data with minimal data loss. However, due to packet losses, the receiving packet rate of Wi-Fi CSI tends to be lower, especially when there is significant distance between the subject and Wi-Fi AP. In our data collection within apartments up to 2000 ft^2, and where the distance between the Wi-Fi AP and the smartphone is impeded by fewer than two solid walls, we observed negligible loss of Wi-Fi packets. Conversely, in scenarios where the

distance exceeds 15 meters or involves more than three solid walls, we noted a notable decrease in packet reception, resulting in a lower packet rate for Wi-Fi CSI. To ensure the data quality, we ensured a minimum packet reception rate of 200 packets per second on smartphones during data collection, corresponding to a 50 Hz Doppler shift akin to daily human motion maximum speed (about 3 m/s). In practical scenarios, if reception rates drop below this threshold, we can alternatively use EgoADL models that operate without Wi-Fi CSI data requirement.

System resource consumption of EgoADL: EgoADL focuses on improve the performance of ADL sensing performance, rather than optimizing the system resource usage. Currently, the computational resource requirement is relatively high. We notice that most of the parameters (275.5 M) are contributed by the self-supervised feature embedding VGG-like DNN models (Sec. 6). We plan to replace them using more efficient DNN models, like MobileNet [65], without losing significant accuracy. Further, a full-fledged implementation of EgoADL needs to carefully split the on-device vs. in-cloud processing, and strikes a balance between computation and communication energy cost. This is left for our future work.

Applicability for EgoADL device: Currently, smartphones serve as the device for EgoADL, primarily chosen for their availability to capture Wi-Fi CSI. However, it is noted that smartphone is not always carried by users, particularly among the elderly population. We recognize that wearable devices, such as smartwatches, may present a more suitable option for EgoADL. One promising direction to explore in future work is to include a more diverse set of wireless sensors, *i.e*. low-cost ultrasound sonar, UWB radar or mmWave radar, on wearable devices [67].

Integrating Large Language Model (LLM) into ADL sensing: EgoADL fine-tunes a language model, *i.e*. BERT [59], to extract and distill contextual information pertaining to human behaviors, as detailed in Section 7.2. While the current implementation deals with computational complexities by fine-tuning a more manageable model size, the approach can be scaled to accommodate the fine-tuning of a larger language model in the future. Moreover, given that large language models are designed for general natural language processing tasks, it may be feasible for EgoADL to bypass fine-tuning altogether. Instead, EgoADL could provide its generated sequence of words and the corresponding probability distribution and organize them as the input prompt directly to the LLM. This would allow the LLM to employ its robust contextual capabilities to refine and correct the word sequence autonomously. We posit that EgoADL paves the way for a novel integration of sensory data with natural language processing. Moving forward, our research will explore to leverage LLMs to enhance the perception and understanding of ADL through different sensing technologies.

10 CONCLUSION

This paper presents the first study that uses a commodity smartphone as an egocentric multi-modal sensor hub to recognize unrestricted user behaviors in free living environment. Although the absolute sensing accuracy of the proposed EgoADL system still leaves room for improvement, its performance is already comparable to state-of-the-art egocentric vision-based solutions. EgoADL verifies several promising mechanisms, such as a joint design of self-supervised single-modal and multi-modal clustering, and context distillation from generic data, which can overcome the fundamental barriers–particularly the generalization and labeling–in ubiquitous sensor-based behavior analysis. Our EgoADL dataset will be released as open source to promote research in both ubiquitous computing and machine learning.

ACKNOWLEDGMENT

We would like to thank the anonymous editors and reviewers for their valuable comments. This work is partially supported by NSF CNS-1901048, CNS-1925767, CNS-2128588, NIH NIA-P30AG073105, Google Ph.D. Fellowship, and Samsung collaboration grant.

REFERENCES

[1] Albert Haque, Arnold Milstein, and Li Fei-Fei. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 2020.

[2] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. In *Proceedings of IEEE CVPR*, 2017.

[3] Dima Damen, Hazel Doughty, and et.al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of ECCV*, 2018.

[4] Gunnar A Sigurdsson, Abhinav Gupta, and et.al. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of CVPR*, 2018.

[5] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of IEEE/CVF CVPR*, 2022.

[6] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of IEEE WACV*, 2021.

[7] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 2021.

[8] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of IEEE/CVF ICCV*, 2019.

[9] Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. Seeing and hearing egocentric actions: How much can we learn? In *Proceedings of IEEE/CVF ICCV Workshop*, 2019.

[10] Linfeng Xu, Qingbo Wu, Lili Pan, Fanman Meng, Hongliang Li, Chiyuan He, Hanxin Wang, Shaoxu Cheng, and Yu Dai. Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning. *arXiv preprint arXiv:2301.10931*, 2023.

[11] Katsuyuki Nakamura, Hiroki Ohashi, and Mitsuhiro Okada. Sensor-augmented egocentric-video captioning with dynamic modal attention. In *Proceedings of ACM MM*, 2021.

[12] Mark Weiser. Some computer science issues in ubiquitous computing. *Communications of the ACM*, 1993.

[13] Gunnar A Sigurdsson, Gül Varol, and et.al. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of ECCV*, 2016.

[14] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[15] EPIC-KITCHENS-100- 2021 Challenges Report, 2022. https://epic-kitchens.github.io/Reports/EPIC-KITCHENS-Challenges-2021-Report.pdf.

[16] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE TPAMI*, 2022.

[17] Jort F Gemmeke, Daniel PW Ellis, and et.al. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of IEEE ICASSP*, 2017.

[18] Yan Wang and et.al. E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In *Proceedings of ACM MobiCom*, 2014.

[19] Wei Wang, Alex X Liu, and et.al. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of ACM MobiCom*, 2015.

[20] Wenjun Jiang, Chenglin Miao, and et.al. Towards environment independent device free human activity recognition. In *Proceedings of ACM MobiCom*, 2018.

[21] Wenjun Jiang, Hongfei Xue, and et.al. Towards 3d human pose construction using wifi. In *Proceedings of ACM MobiCom*, 2020.

[22] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. In-home daily-life captioning using radio signals. In *Proceedings of ECCV*, 2020.

[23] Shuochao Yao and et.al. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of WWW*, 2017.

[24] Huatao Xu, Pengfei Zhou, and et.al. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of ACM SenSys*, 2021.

[25] Nicolas Turpault, Romain Serizel, and et.al. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. 2019.

[26] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of ACM UIST*, 2018.

[27] Xiaomin Ouyang and et.al. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of ACM MobiCom*, 2022.

[28] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. Leveraging sound and wrist motion to detect activities of daily living with commodity smartwatches. *Proceedings of ACM IMWUT*, 2022.

[29] Yin Li, Miao Liu, and Jame Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE TPAMI*, 2021.

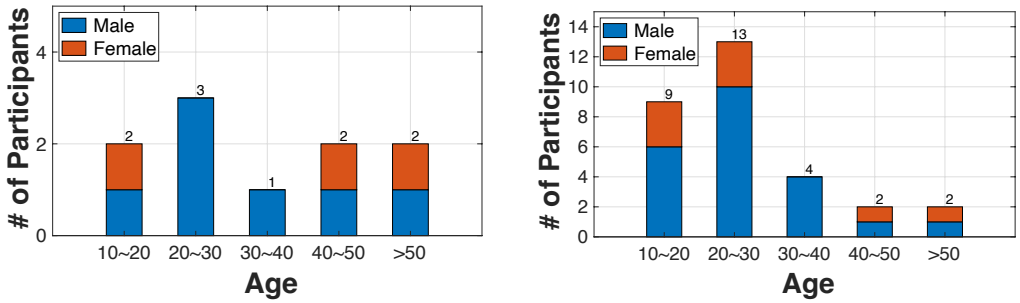
[30] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. *Proceedings of ACM IMWUT*, 2018.
[31] Nirupam Roy, He Wang, and Romit Roy Choudhury. I am a smartphone and i can tell my user's walking direction. In *Proceedings of MobiSys*, 2014.
[32] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM TASLP*, 2020.
[33] Jason Wu and et.al. Automated class discovery and one-shot interactions for acoustic activity recognition. In *Proceedings of ACM CHI*, 2020.
[34] Rebecca Adaimi, Howard Yong, and Edison Thomaz. Ok google, what am i doing? acoustic activity recognition bounded by conversational assistant interactions. *Proceedings of ACM IMWUT*, 2021.
[35] Yan Wang, Jian Liu, and et.al. E-eyes: In-home device-free activity identification using fine-grained WiFi signatures. In *Proceedings of ACM MobiCom*, 2014.
[36] Harish Haresamudram, Irfan Essa, and Thomas Plötz. Contrastive predictive coding for human activity recognition. *Proceedings of ACM IMWUT*, 2021.
[37] Chi Ian Tang and et.al. Selfhar: Improving human activity recognition through self-training with unlabeled data. *Proceedings of ACM IMWUT*, 2021.
[38] Haojie Ma, Zhijie Zhang, and et.al. Unsupervised human activity representation learning with multi-task deep clustering. *Proceedings of ACM IMWUT*, 2021.
[39] Chuhan Gao and et.al. Livetag: Sensing human-object interaction through passive chipless wifi tags. In *Proceedings of USENIX NSDI*, 2018.
[40] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE TPAMI*, 2020.
[41] Harish Haresamudram, Irfan Essa, and Thomas Plötz. Assessing the state of self-supervised human activity recognition using wearables. *Proceedings of ACM IMWUT*, 2022.
[42] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Proceedings of NeurIPS*, 2020.
[43] Hassan Akbari and et.al. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Proceedings of NeurIPS*, 2021.
[44] Brian Chen, Andrew Rouditchenko, and et.al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of ICCV*, 2021.
[45] Francesco Gringoli and et.al. Free your csi: A channel state information extraction platform for modern wi-fi chipsets. In *Proceedings of ACM WiNTECH*, 2019.
[46] Matthias Schulz, Daniel Wegemer, and Matthias Hollick. Nexmon: The c-based firmware patching framework, 2017. https://nexmon.org.
[47] Gierad Laput and Chris Harrison. Sensing fine-grained hand activity with smartwatches. In *Proceedings of ACM CHI*, 2019.
[48] Dan Wu and et.al. Device-free wifi human sensing: From pattern-based to model-based approaches. *IEEE Communications Magazine*, 2017.
[49] Xiyuan Zhang, Ranak Roy Chowdhury, Dezhi Hong, Rajesh K Gupta, and Jingbo Shang. Modeling label semantics improves activity recognition. *arXiv preprint arXiv:2301.03462*, 2023.
[50] Yunus Emre Ustev, Ozlem Durmaz Incel, and Cem Ersoy. User, device and orientation independent human activity recognition on mobile phones: Challenges and a proposal. In *Proceedings of ACM UbiComp*, 2013.
[51] Shuochao Yao, Ailing Piao, Wenjun Jiang, Yiran Zhao, Huajie Shao, Shengzhong Liu, Dongxin Liu, Jinyang Li, Tianshi Wang, Shaohan Hu, et al. Stfnets: Learning sensing signals from the time-frequency perspective with short-time fourier neural networks. In *Proceedings of WWW*, 2019.
[52] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of ECCV*, 2020.
[53] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of IEEE/CVF ICCV*, 2019.
[54] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *Proceedings of IEEE ICASSP*, 2021.
[55] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
[56] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *Proceedings of Interspeech*, 2021.
[57] Ashwin K Vijayakumar and et.al. Diverse beam search: Decoding diverse solutions from neural sequence models. In *Proceedings of AAAI*, 2016.
[58] Mathilde Caron, Piotr Bojanowski, and et.al. Deep clustering for unsupervised learning of visual features. In *Proceedings of ECCV*, 2018.

[59] Jacob Devlin and et.al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.
[60] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of ECCV*, 2018.
[61] Dima Damen, Hazel Doughty, and et.al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE TPAMI*, 2021.
[62] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Proceedings of IEEE/CVF CVPR*, 2018.
[63] Yanghao Li and et.al. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of IEEE/CVF CVPR*, 2021.
[64] Swaminathan Vasanth Rajaraman and et.al. Energy consumption anatomy of live video streaming from a smartphone. In *Proceedings of IEEE PIMRC*, 2014.
[65] Mark Sandler and et.al. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of IEEE CVPR*, 2018.
[66] Jianwei Qian and et.al. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *Proceedings of ACM SenSys*, 2018.
[67] Wei Wang, Alex X. Liu, and Ke Sun. Device-free gesture tracking using acoustic signals. In *Proceedings of ACM MobiCom*, 2016.
[68] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. Spearphone: a lightweight speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. In *Proceedings of ACM WiSec*, 2021.
[69] Angularjs typeahead, 2022. https://www.npmjs.com/package/ngx-typeahead.
[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS*, 2017.

METHODOLOGICAL TRANSPARENCY & REPRODUCIBILITY APPENDIX (META)

Demographics of Participants in Data Collection

For our data collection, we engaged 30 participants, comprising 8 females and 22 males, as shown in Fig. 9. The participants had an average age of 25.9 years. Among them, 10 participants (3 females and 7 males) consented to provide ground-truth labels for our study. One notable limitation of our current dataset is the underrepresentation of female and elderly participants. To address this and to support future research, we released our dataset (https://doi.org/10.5281/zenodo.8248159), data collection and labeling platform, processing source code (https://github.com/Samsonsjarkal/EgoADL) to facilitate further research. We hope that by sharing these resources, we can encourage a broader spectrum of participants and research institutes to contribute to the EgoADL data collection, thereby enhancing the demographic diversity of our dataset.



(a) Participants in labeled dataset.

(b) Participants in unlabeled dataset.

Fig. 9. Demographics of Participants.

Data Preprocessing and Labeling

Fig. 10 summarizes the preprocessing pipeline. To mitigate the impact of unstable sample timing on COTS smartphones [68], we first resample the motion sensor data uniformly to 200 Hz. For the Wi-Fi CSI data, we preprocess it as follows: 1) Discard the frames without < 80 MHz bandwidth, and only keep the data frames with 80 MHz; 2) Compensate the automatic gain control (AGC) to guarantee the stable amplitude of Wi-Fi CSI signals in the time domain; 3) Discard the subcarriers without Wi-Fi CSI; 4) Resample the Wi-Fi CSI uniformly to 400 Hz sampling rate. Afterwards, we synchronize the three sensing modalities based on their sampling timestamp.

We use egocentric video and audio to assist ground truth labeling. To achieve accurate timestamp labeling, we first synchronize the data collected by smartphone with the egocentric video and audio collected by GoPro. We perform the cross correlation between existing audio recordings from the smartphone and GoPro to achieve the synchronization between data from smartphone and GoPro, as shown in Fig. 10.

Each ground truth label needs to specify the human behavior, *i.e.*, an ambulatory action or human-object interaction event, along with the start and end timestamps. Fig. 11 illustrates the UI of our labeling tool. It allows playback of the GoPro video/audio, and annotating the data using a set of human behavior labels created by two state-of-the-art video/audio based ADL sensing systems [5, 13, 15, 47]. To ensure accurate labeling, the annotators are exactly the data collectors, compensating for any limitations in egocentric video field of view (FoV) by using their memory of the events. They are equipped to segment and label the video footage, either using the provided predefined ADL labels or by adding new ones as they identify them. The maximum allowable duration for each labeled segment is 10 seconds, aligning with the typical duration of discrete ADL observed in our studies. To

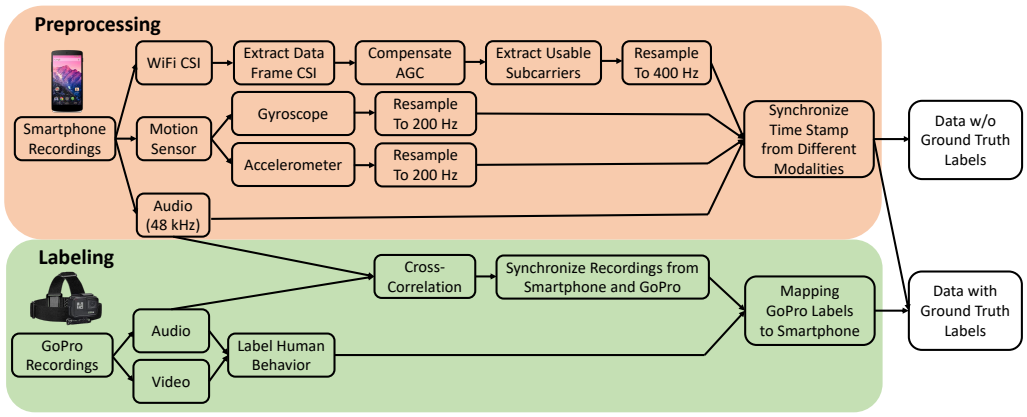


Fig. 10. Implementation pipeline of the EgoADL preprocessing and labeling.

accelerate the labeling, we adopt Angular Typeahead [69]–an input field API allowing a user to quickly type and select from a list of matching labels, or create their own. If the volunteer can not find their preferred ADL labels, they can manually add new ADL labels via the UI of our tool. Note that we only annotate a single most relevant label when multiple behaviors are involved simultaneously.

The labeled dataset comprises 7,000 human behavior samples, including 221 types of human behaviors (Fig. 16(a)) with 70 actions (Fig. 16(b)) and 91 objects (Fig. 16(c)). We also separate the human behavior set into 35 *state-based behaviors* (Fig. 12) and 190 *event-based behaviors* (Fig. 13), The former typically last more than 5 s each and often periodically and continuously, like "walking" and "chopping meat", *etc*. The latter are one-short behaviors, like "opening the door" and "sitting down in chair", *etc*. Fig. 12 and Fig. 13 visualize the frequency of state-based behaviors and event-based behaviors, respectively.

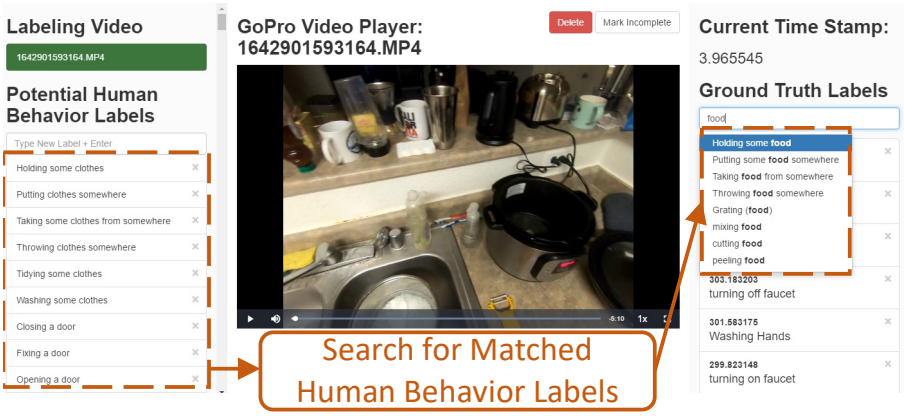


Fig. 11. EgoADL labeling tool

Details of Neural Network Design

In our main paper, we omit the detailed design of the neural network design.

Tab. 8 shows the design of different CNN-based encoder parameters. To make sure all the representations after encoders can fit into the transformer, we intentionally enforce the feature representation of a single frame to be the same size. Therefore, the "Slow-pathway" encoders will have more channels than the "Fast-pathway" encoders while the "Modalwise" encoders will have more channels than "Framewise" encoders. Our transformer network architecture is based on [70]. It comprises 12 encoder layers and 6 decoder layers. Positional encoding is employed to capture temporal dynamics in sensory time-series data. Within the multi-head attention mechanism, we have configured 8 heads, and the dimensionality of the feedforward network is set at 3072 as default.

We released DNN source code (https://github.com/Samsonsjarkal/EgoADL) to facilitate further research.

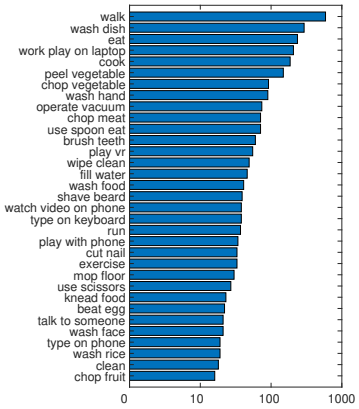


Fig. 12. EgoADL State-based human behaviors

	Slow-pathway	Fast-pathway			Fast-pathway (Framewise)		
Modality	Audio	Audio	Acc	CSI	Audio	Acc	CSI
Spectrogram	(t, 64) C:1	(t, 64) C:1	(t, 64) C:3	(t, 64) C:208	(t, 64) C:1	(t, 64) C:3	(t, 64) C:208
Stride	(8,1)	(2,1)	(2,1)	(2,1)	(2,1)	(2,1)	(2,1)
CNN	3*3, C:120	3*3, C:30	3*3, C:30	3*3, C:30	3*3, C:10	3*3, C:10	3*3, C:10
Activation	BatchNorm2d+LeakyReLU + Dropout(0.1)						
Stride	(2,1)	(2,1)	(2,1)	(2,1)	(2,1)	(2,1)	(2,1)
CNN	3*3, C:240	3*3, C:60	3*3, C:60	3*3, C:60	3*3, C:20	3*3, C:20	3*3, C:20
Activation	BatchNorm2d+LeakyReLU + Dropout(0.1)						
Stride	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)
CNN	3*3, C:480	3*3, C:120	3*3, C:120	3*3, C:120	3*3, C:40	3*3, C:40	3*3, C:40
Concat	/	/	/	/	Framewise Concat		
Concat	Modalwise Concat						

Table 8. CNN-based encoders parameters (C: Channel).

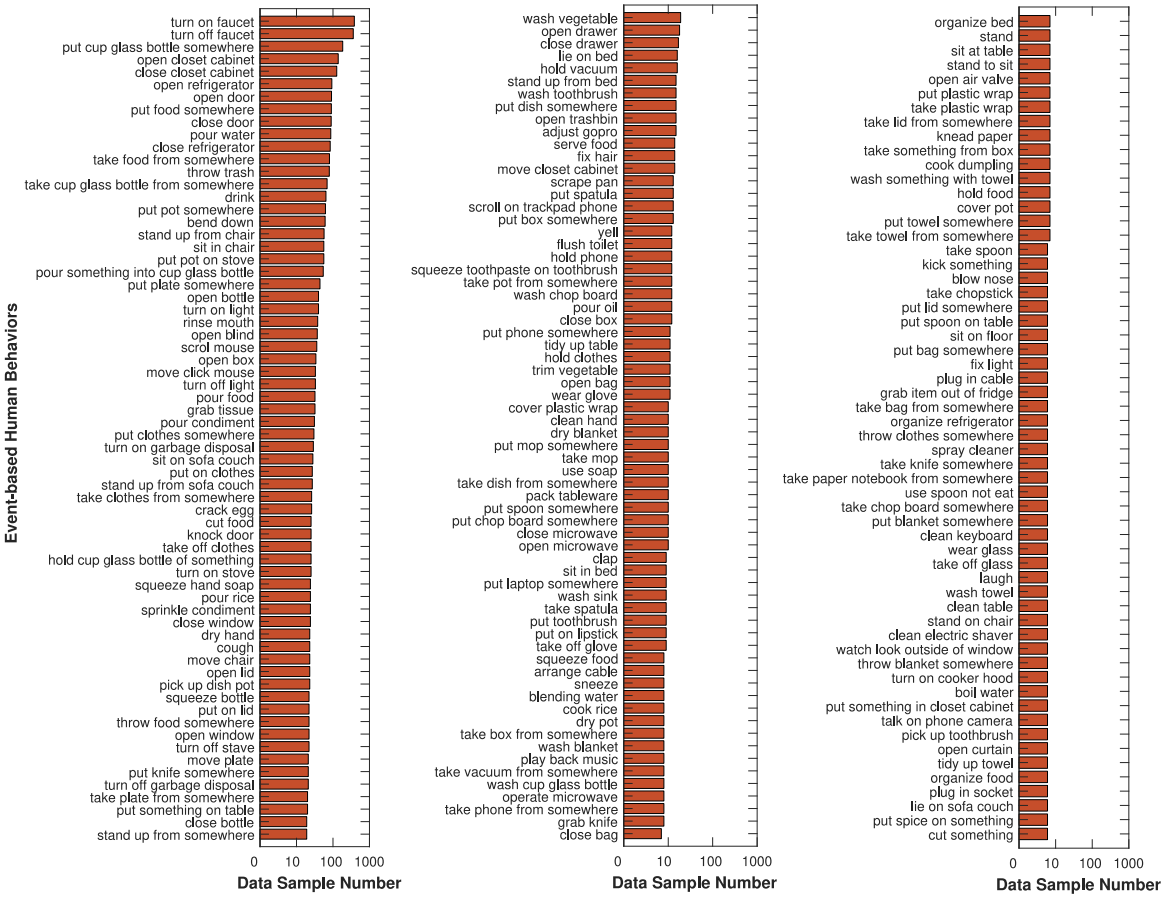


Fig. 13. EgoADL Event-based human behaviors

(a) Device-free audio SNR (b) Device-free Wi-Fi SNR. (c) Egocentric Wi-Fi SNR. (d) WiFi Sensing Model.

Fig. 14. Egocentric v.s. Device-free Sensing SNR for audio and Wi-Fi. "User squeezing plastic bootle" sound and "Sit down in chair" motion are used as benchmarks sound event and human activity for audio SNR and Wi-Fi CSI SNR measurements. The heatmap in (a) represents the SNR as we vary sound source locations while fixing the mic. The heatmap in (b) and (c) represents the Wi-Fi CSI sensing SNR of specific location.

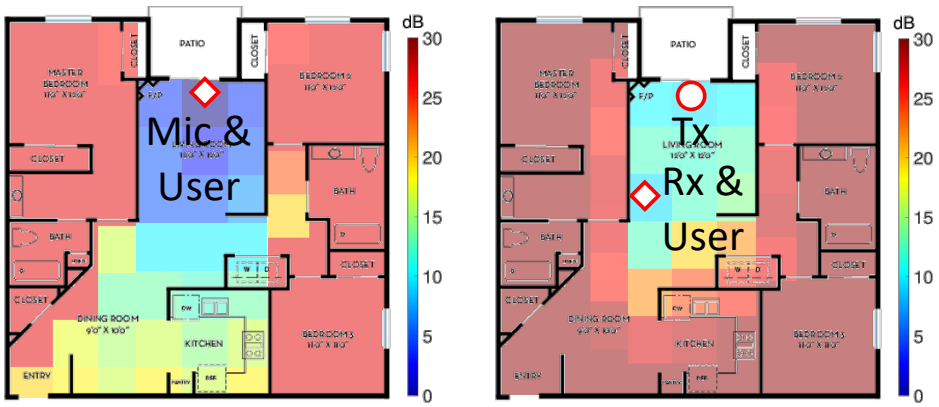
APPENDIX

Preliminary Study on EgoADL

EgoADL employs a commercial smartphone as an egocentric sensor hub, capturing audio, wireless sensing signals (*i.e.* Wi-Fi), and motion sensor signals continuously. Users can perform *arbitrary* daily routines with the sensor hub, *e.g.*, an in-pocket smartphone, in free-living environment. To better understand the advantages of the egocentric sensing, we conduct controlled experiments in a 1400 ft^2 apartment, and compare EgoADL against the conventional device-free setup [48] which captures users' behaviors through off-body sensors, *e.g.*, voice assistant [26, 33, 34] and Wi-Fi AP [19, 20]. We only examine the audio and Wi-Fi CSI modalities here since motion sensors are already widely used in egocentric setup [23, 24], which characterize a specific body part motion without any interference from other users.

Sensing Space Coverage: To control the audio sensing setup, we use a loudspeaker to replay a benchmark sound of "user squeezing plastic bottle" at a constant 68 dBA SPL, to emulate the corresponding user activity. For Wi-Fi CSI sensing, we use "sit down in chair" as the benchmark activity. In the device-free scenario, we vary the sound source location and the location of human activity while fixing the sensor hub at a specific location (◇ in Fig. 14(a) and Fig. 14(b)). In the egocentric scenario, the users put the sensor hub, *i.e.* smartphone, in their trouser pocket. As shown in Fig. 14(a), in the device-free setup, the microphone is sensitive to wall blockage, and can only sense the activity sound at single-room coverage. Meanwhile, Fig. 14(b) shows that, the signal strength of device-free Wi-Fi sensing drops dramatically as the user moves away from the Tx/Rx or behind the wall. This is because it relies on the NLoS signals bouncing off the target user's body, which suffers from severe attenuation, diffraction, and scattering effects. *In contrast, in the egocentric setup, the sensor hub accompanies the user, so it achieves whole-home coverage with consistently high SNR (> 25 dB) for both audio and Wi-Fi CSI.*

Resilience to interference: Since the device-free setup achieves low whole-home SNR even without the interference source, we only examine the resilience to interference under the egocentric setup. Here the targeted user stays at a fixed location, while another (interfering) user performs the benchmark activity at arbitrary locations. We measure the SINR, where the desired signal power equals to the variance of egocentric signals caused by the targeted user's activities, whereas the interference is that from the interfering user. As shown in Fig. 15, for both audio and Wi-Fi CSI sensing, *the presence of an interfering user will noticeably impact the egocentric SINR (from 25 dB to 10 dB) only when it is in close proximity (< 2 m) of the target user or blocking the LoS path between the Tx and Rx for Wi-Fi CSI sensing.*



(a) Audio Egocentric SINR. (b) Wi-Fi Egocentric SINR.

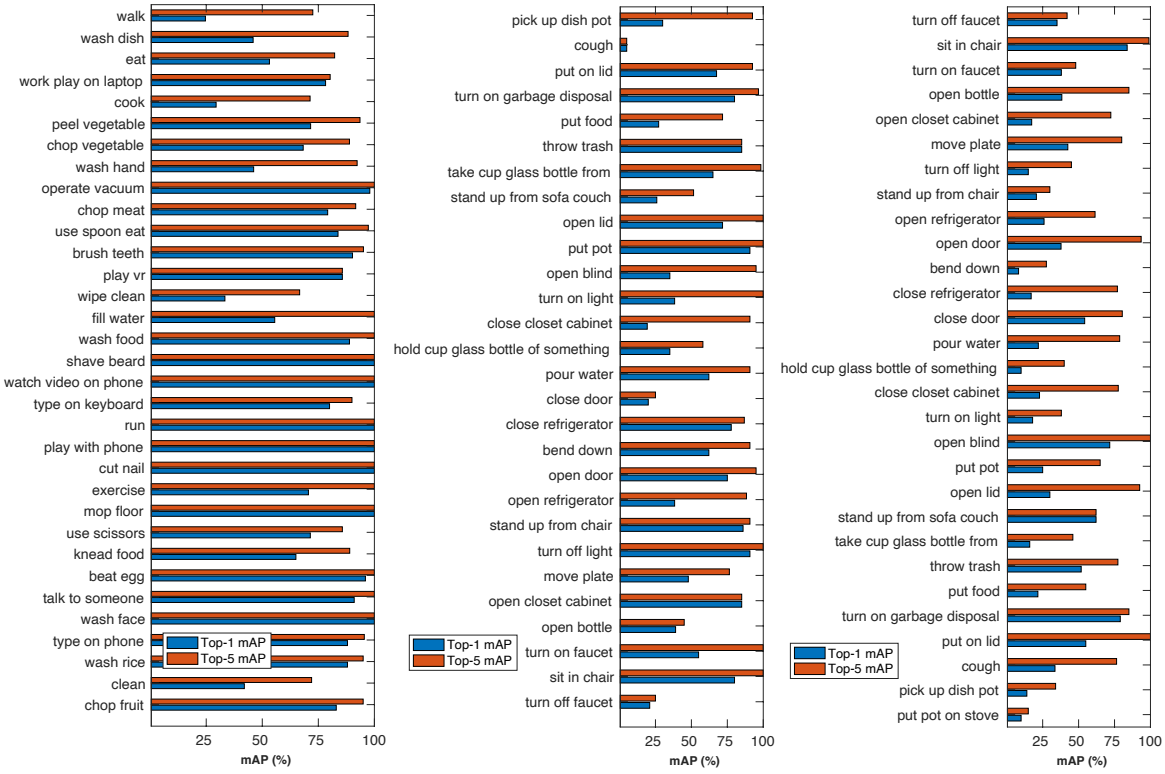
Fig. 15. Egocentric SINR for the same benchmark sound event and human activity as in Fig. 16. The targeted user with egocentric sensor hub is fixed at the location of ◇. The heatmap in (a) and (b) represents the SINR of audio and Wi-Fi CSI of the targeted user when there is an interfering subject performing the benchmark activity at each location.

(a) Behavior Words (b) Action Words (c) Object Words

Fig. 16. EgoADL Dataset Labels Word Cloud

To sum up, compared to conventional device-free setup, egocentric sensing significantly improves the operating range and anti-interference ability for both Wi-Fi CSI and audio sensor. Although close-by interferers may still impact the SINR, we anticipate the motion sensor along with deep modality fusion can neutralize such impacts. Thus, in EgoADL, we do not restrict the presence of interference–*all the data are collected in daily living settings with multiple coresidents*.

The labeled dataset comprises 7,000 human behavior samples, including 221 types of human behaviors (Fig. 16(a)) with 70 actions (Fig. 16(b)) and 91 objects (Fig. 16(c)). We also separate the human behavior set into 35 *state-based behaviors* (Fig. 12) and 190 *event-based behaviors* (Fig. 13), The former typically last more than 5 s each and often periodically and continuously, like "walking" and "chopping meat", *etc*. The latter are one-short behaviors, like "opening the door" and "sitting down in chair", *etc*. Fig. 12 and Fig. 13 visualize the frequency of state-based behaviors and event-based behaviors, respectively.



(a) State-based human behaviors

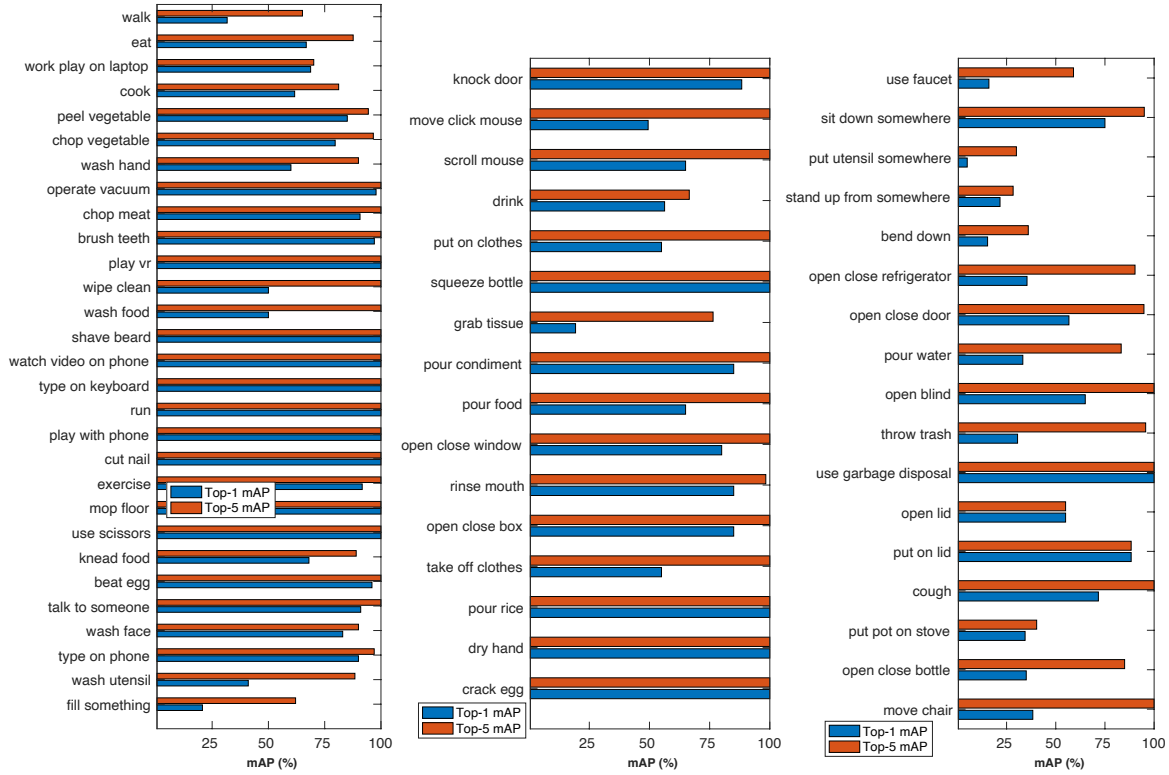
(b) Event-based human behaviors

Fig. 17. EgoADL classwise mAP

Detailed Experimental Results

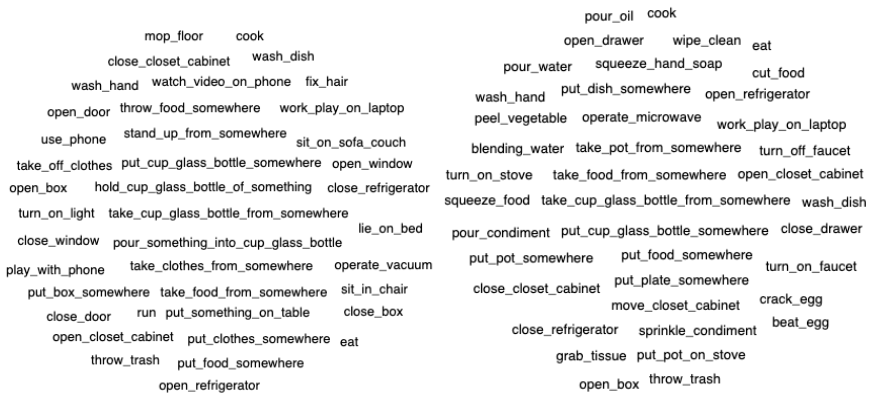
In this section, we present the classwise performance of each human behavior to provide a comprehensive understanding of the strengths and limitations of EgoADL. Our evaluation methodology is as follows: We first train the encoders with the cross-modal self-supervised learning approach described in Sec. 6, using 100 hours of unlabeled data. This is followed by fine-tuning the "MMFWSF" sequence-to-sequence model using a balanced dataset of 2,500 labeled samples. The classwise mean Average Precision (mAP), both top-1 and top-5, is evaluated using an unbalanced set of 2,800 labeled samples, incorporating 5-fold cross-validation. This evaluation ensures a balanced distribution of samples from all 10 users across training, validation, and testing sets. Figure 17 shows the top-1 and top-5 mean Average Precision (mAP) for each ADL. Our results show that EgoADL achieves an overall mAP of 59.2%, with 79.5% and 49.8% for state-based and event-based human behaviors, respectively. For Fig. 18, the only difference in the training, validation, and testing settings is the application of refined labels (Sec. 7.1). After label refinement, we achieve an top-1 overall mAP of 68.2%, with 83.2% and 55.9% for state-based and event-based ADL, respectively.

We have also provided the detailed labels that both appeared in both the egocentric vision and EgoADL datasets. We use these labels to compare the performance between EgoADL and egocentric vision in Sec. 8.4. Fig. 19(a) and Fig. 19(b) shows EgoADL overlapped ADL labels with Charades-Ego [4] and EGTEA-GAZE [6], respectively.



(a) State-based human behaviors (b) Event-based human behaviors

Fig. 18. EgoADL classwise mAP with label refinement



(a) Overlapping between EgoADL and Charades-Ego [4] (b) Overlapping between EgoADL and EGTEA-GAZE [6]

Fig. 19. Overlapped ADL labels between EgoADL and egocentric vision