

RF-HOI: Recognize Human-Object Interaction with Radio Frequency Signals

LIHAO WANG, Johns Hopkins University, USA

LINLU GAO, Johns Hopkins University, USA

JIACAN YU, Johns Hopkins University, USA

YANYU LIN, Johns Hopkins University, USA

YIFAN YIN, Johns Hopkins University, USA

JIANXIN WANG, Johns Hopkins University, USA

TIANMIN SHU, Johns Hopkins University, USA

RENJIE ZHAO*, Johns Hopkins University, USA

Recognizing Human-Object Interactions (HOI) is essential for intelligent systems, underpinning applications in virtual and augmented reality, embodied AI, and assistive robotics. However, vision-based HOI methods face challenges in privacy concerns and poor light conditions. In this work, we introduce *RF-HOI*, the first framework that only uses radio frequency (RF) signals for HOI recognition. A key challenge of *RF-HOI* is that single-modality RF sensing is insufficient to recognize both actions and the objects being interacted with. *RF-HOI* addresses this through a novel modality fusion that combines mmWave radar and RFID, enabling simultaneous action recognition and target identification. Another challenge is limited training data across diverse setups, which impairs the generalizability of the recognition model. To overcome this, we develop a simulator that synthesizes multimodal RF data for diverse HOIs at scale, allowing us to fine-tune with only a small amount of real-world data. Experiment results show that *RF-HOI* outperforms all baselines, approaching vision model performance, and that our diverse synthetic training data can significantly boost our system's performance on real-world scenarios. These results highlight the potential of multimodal RF sensing for robust and privacy-preserving HOI recognition as well as the effectiveness of our RF data synthesis.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Wireless Sensing, Human-Object Interaction, mmWave, RFID

ACM Reference Format:

Lihao Wang, Linlu Gao, Jiacan Yu, Yanyu Lin, Yifan Yin, Jianxin Wang, Tianmin Shu, and Renjie Zhao. 2026. RF-HOI: Recognize Human-Object Interaction with Radio Frequency Signals. In *Proceedings of Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*. ACM, New York, NY, USA, 36 pages. <https://doi.org/XXXXXXX.XXXXXXX>

*Renjie Zhao is the corresponding author.

Authors' Contact Information: Lihao Wang, Johns Hopkins University, Baltimore, Maryland, USA, lwang231@jhu.edu; Linlu Gao, Johns Hopkins University, Baltimore, Maryland, USA, lgao25@jh.edu; Jiacan Yu, Johns Hopkins University, Baltimore, Maryland, USA, jyu197@jh.edu; Yanyu Lin, Johns Hopkins University, Baltimore, Maryland, USA, ylin185@jh.edu; Yifan Yin, Johns Hopkins University, Baltimore, Maryland, USA, yyin34@jhu.edu; Jianxin Wang, Johns Hopkins University, Baltimore, Maryland, USA, jwang616@jh.edu; Tianmin Shu, Johns Hopkins University, Baltimore, Maryland, USA, tianmin.shu@jhu.edu; Renjie Zhao, Johns Hopkins University, Baltimore, Maryland, USA, rjzhao@jhu.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., Woodstock, NY

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Recognizing Human-Object Interactions (HOI) is foundational to human activity understanding tasks such as action anticipation [81] and goal inference [93, 97, 135, 141]. This capability is crucial for intelligent systems that are designed to interact with humans, including virtual agents in Virtual Reality (VR) [8, 38] and Augmented Reality (AR) [48], and assistive robots [97, 101]. HOI recognition specifically seeks to *simultaneously classify the human action and identify the interaction target*—the object being manipulated [12, 33, 131, 133]. This joint recognition provides more information than the widely studied Human Activity Recognition (HAR), which considers only the action. For example, in Fig. 1, when a user picks up a book, HOI recognizes a full tuple <pick up, book>, whereas HAR outputs only the action <pick up>.

While advances in computer vision have spurred extensive research on image- and video-based HOI recognition [12, 33, 66, 68, 131, 133], these approaches face significant real-world challenges. Specifically, vision-based methods are constrained by privacy concerns [143], degraded performance in poor or variable lighting [29], and challenges in distinguishing visually similar objects [15]. Such issues hinder their long-term and ubiquitous deployment, particularly in privacy-sensitive environments like bathrooms or bedrooms, or in variable lighting throughout the day. In contrast, recent advances in wireless sensing have demonstrated that Radio Frequency (RF) signals can enable effective user action classification [9, 70, 103, 113, 132, 140, 142] and target identification [30, 72, 96, 112], while remaining robust to lighting conditions and maintaining user privacy.

Further, driven by the adoption of commercial products, wireless sensing systems increasingly have the potential to leverage existing infrastructure, thereby reducing deployment effort and hardware cost. For example, mmWave radars have been widely integrated into off-the-shelf smart home products by companies such as Xiaomi [123] and Logitech [73]. In parallel, many real-world environments already deploy RFID tags as part of existing systems. In retail settings, retailers such as H&M [36] and Decathlon [19] routinely attach RFID tags to individual products to support inventory management, customer assistance, and fast checkout. Beyond retail, RFID-tagged personal items are also common in commercial healthcare and assisted living environments [6, 86], where medication containers, daily necessities, and assistive tools are tagged for tracking and safety purposes. RFID tags have further been adopted in smart home environments [25, 99], where reusable containers or packages can be instrumented to support activity monitoring and context-aware assistance. Moreover, RFID readers can be implemented in lightweight and wearable forms, such as wrist-worn [11] or hand-held [10] devices, which have been adopted in practical retail [110] and healthcare [85] deployments. Building on these existing deployments, our system leverages RFID-tagged objects to enable HOI recognition, reducing the need for additional sensing infrastructure while providing reliable object identification.

In light of the above, we propose *RF-HOI*, the first exploration of HOI recognition using only RF signals. In a privacy-preserving manner, *RF-HOI* focuses on single-user HOI recognition with commercial devices, offering a better user understanding in indoor scenarios. Despite these advances, applying wireless sensing techniques to HOI recognition is non-trivial, since it faces two challenges.

(1) *Single modality design cannot be directly extended for HOI tasks.* As to be detailed in Sec. 2.1, although single modality can address a subtask of HOI like using mmWave radar for HAR or RFID for target identification, none of them can address the tasks at the same time. Furthermore, naively combining the two modalities is insufficient for HOI because it will miss the mapping between the action and objects. To grapple with these challenges, we propose a multimodal system that organically combines the advantages of the mmWave radar and RFID modalities for the HOI tasks. As illustrated in Fig. 2, each candidate object is attached with an RFID tag, while the human does not wear any devices. We use an mmWave radar to collect signal changes introduced by human actions and an RFID reader to collect the signal changes of all RFID tags. We devise a deep neural network to effectively fuse the two modalities. The model first applies modality-specific encoders to extract spatiotemporal features within each modality. Because HOI durations vary across scenarios, we use transformer-based temporal

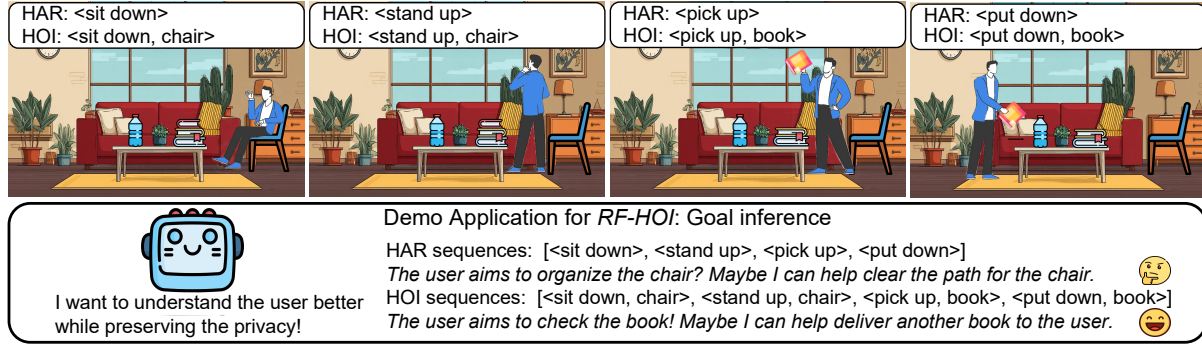


Fig. 1. Using RF signals, *RF-HOI* can understand the user better in a privacy-preserving way by recognizing the user's interaction with objects. Compared with HAR, HOI recognition considers both human actions and target objects. As is shown in the figure, one of the applications of HOI recognition are the goal inference task. The agent can infer a more appropriate goal of the user based on HOI sequences, and then provide assistance correspondingly.

blocks to handle variable-length sequences. Then the model performs cross-modal fusion and decouples the output of actions and objects with two separate branches for better generalizability and interpretability.

(2) *HOI recognition introduces much more costs on data collection.* To date, there is no dataset with object annotations for RF-only HOI recognition. Building such a dataset is costly for two reasons. First, the number of HOI categories far exceeds those in HAR [70, 103, 113], inflating annotation costs [66]. Second, for model generalizability, wireless sensing datasets have to cover diverse setups (e.g., user orientation, distance, body shape, and environment), which is especially important for HOI recognition since both actions and targets are affected by setups. Consequently, it is prohibitively expensive to assemble a sufficiently diverse, well-annotated real-world dataset for RF-based HOI recognition. To address this challenge, we first analyze the impacts of various factors to guide future large-scale data collection. To improve model performance and generalizability with limited real data, we develop an RF signal simulator that generates synthetic HOI data using well-annotated 3D-mesh traces from recent works in computer vision[49–51]. This enables the generation of diverse synthetic data, including 7,518 well-annotated multimodal samples. We highlight that the simulator enables the generation of synthetic data with expanded object categories. When these categories exhibit interaction patterns similar to those seen during training, a recognition model pre-trained on such synthetic data can generalize to them, even though they are not included in real-world fine-tuning. Finally, for evaluation purposes, we build a data collection platform and collect 3,615 real-world samples across 78 distinct setups. We pre-train the proposed modality-fusion model on synthetic data and fine-tune it on a small subset of real-world samples to bridge the sim-to-real gap.

We conduct comprehensive evaluations using the collected real-world data. Results show that *RF-HOI* achieves an average accuracy of 96.97%, with only a 1.96% performance gap compared to the vision baselines. *RF-HOI* outperforms its non-fusion variant by 21.29%, RFID-only by 32.90%, and mmWave-only by 73.20%. Moreover, incorporating synthetic data boosts performance by an average of 12.00% across various real-world training set sizes and by 12.09% across different numbers of seen setups.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to address HOI recognition only using RF signals, providing a privacy-preserving alternative to vision-based methods.
- We introduce a novel multimodal approach that combines mmWave radar and RFID sensing for HOI recognition. We design a modality fusion network that processes spatiotemporal features and outputs decoupled HOI recognition.

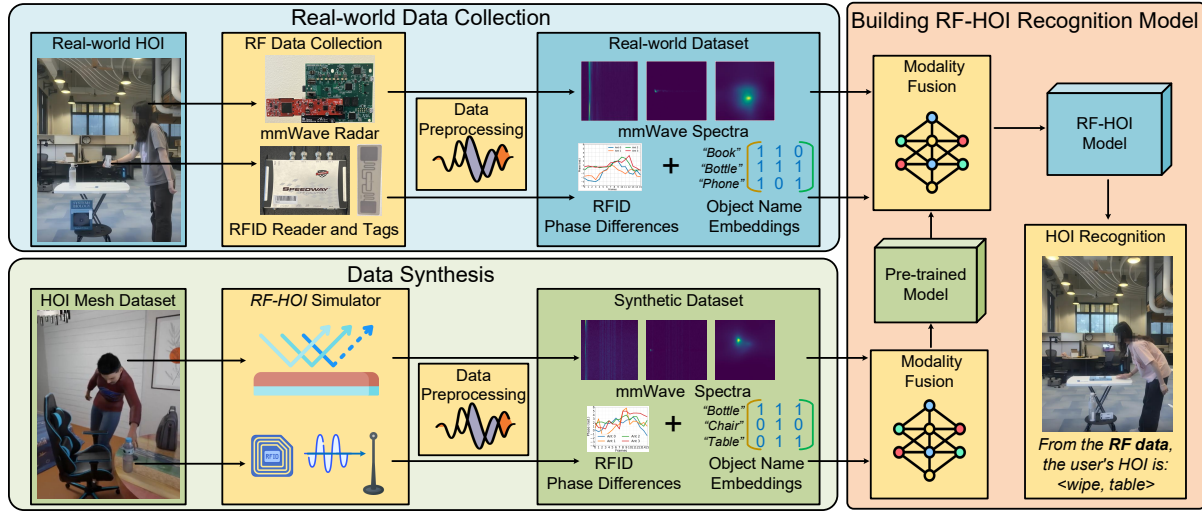
Fig. 2. Overview of *RF-HOI* Framework.

Table 1. Compare different sensing settings. RF-HOI fuses both mmWave and RFID modalities.

Modality	Device assistance	Fine-grained action capture	Target identification	User convenience	HOI recognition
mmWave	Device-free HAR [9, 70, 103, 117, 142]	Yes	Low	High	Low
	Tags on objects [4, 5, 64, 75, 105]	No	Medium	High	Low
	Tags as background [41, 108, 113, 119]	No	Low	High	Low
RFID	Tags on objects [35, 60, 106, 139]	No	Medium	High	Low
	Tags on humans [52, 115, 120, 130]	Yes	Low	Low	Low
RF-HOI	Tags on objects	Yes	High	High	High

- We develop a multimodal simulator that can synthesize diverse, high-quality RF data, enabling improved HOI recognition performance and supporting generalization to unseen object categories with similar interaction patterns, especially when real-world RF data is limited.
- We build a real-world prototype using commercial devices for experiments across different distances, orientations, environments, and users. Extensive evaluations demonstrate the effectiveness and generalizability of *RF-HOI* in the real world.

2 Challenges and Motivation

2.1 Deficiency in single modality

For effective HOI recognition, a sensing system must *capture fine-grained action features*, *identify the target object*, and *offer user convenience*. However, as summarized in Table 1, widely used single-modality approaches—such as mmWave-only or RFID-only—fall short of meeting all these requirements.

mmWave suffers for target identification. While mmWave radar, like vision-based approaches, can capture fine-grained action features in a convenient device-free manner, it is largely limited to HAR tasks [9, 69, 70, 83, 103, 117, 142]. Target identification is challenging for mmWave in device-free settings, as commercial mmWave radars lack the resolution to directly distinguish objects [1]. Even advanced solutions with larger radar arrays

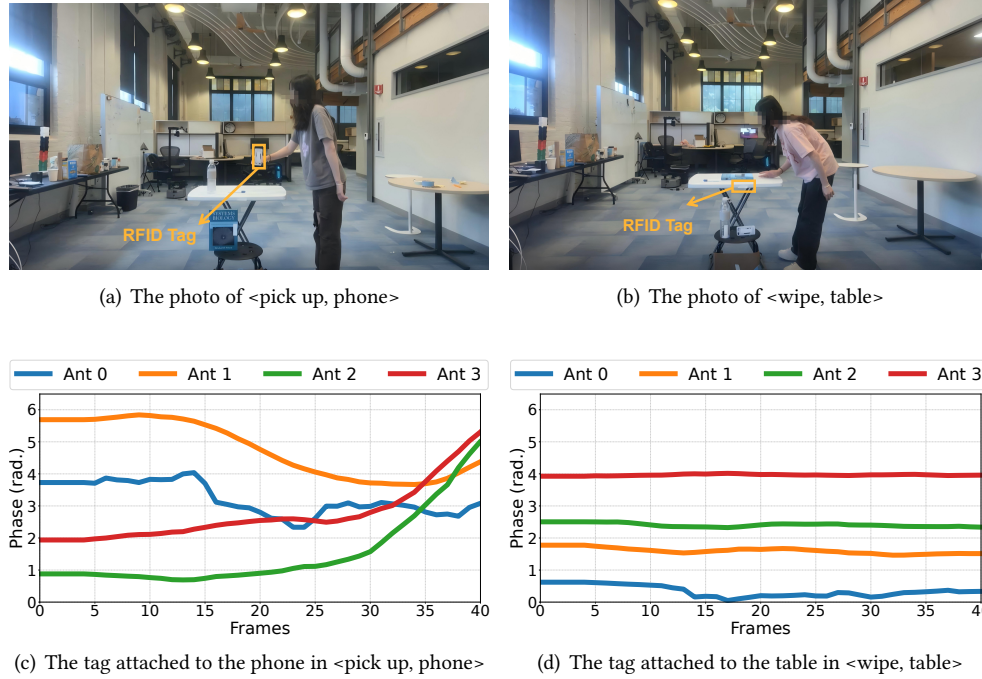


Fig. 3. Photos and corresponding RFID signal phases of the tag attached to the target object in two different interactions.

[137] or metasurfaces [118] typically infer object identity from contours and can only handle one object at a time [13, 34], making them unsuitable for HOI scenarios with multiple candidate objects. Attaching mmWave tags to objects [4, 5, 64, 75, 105] may support object identification via tag ID and localization, and could be an option in the future with hardware improvements and commercialization. At present, however, these solutions are impractical because they require specialized hardware design and their tags may be expensive or require external energy.

RFID suffers for capturing action features. While attaching RFID tags enables target identification, RFID-based methods face significant limitations in action recognition. First, placing large arrays of RFID tags in the background [28, 41, 108, 113, 119] introduces a high deployment cost, has a limited update rate due to the limited RFID reading capacity, and needs to constrain the activity between reader and tags. Second, attaching tags to target objects [35, 60, 106, 139] enables target identification only when the object is moved, but fails to capture action features or recognize interactions where the user does not touch the tag or move the object—situations common in HOI tasks. For example, as is shown in Fig. 3, using a four-antenna RFID reader, we observe distinct phase changes during the <pick up, phone> interaction when the tagged object was moved, but minimal phase variation during <wipe, table> when the tag (affixed to a stationary table) was not directly manipulated. Without discernible patterns in RFID signals, accurate action classification becomes infeasible for RFID-only approaches. Finally, although directly attaching tags to the human body [52, 114, 115, 120, 122, 130] can capture fine-grained actions, this approach is impractical for real-world deployment as it is inconvenient for users.

Therefore, we argue that it is essential for *RF-HOI* to integrate both mmWave and RFID sensing for HOI recognition—leveraging mmWave for robust action classification, RFID for accurate target identification, and

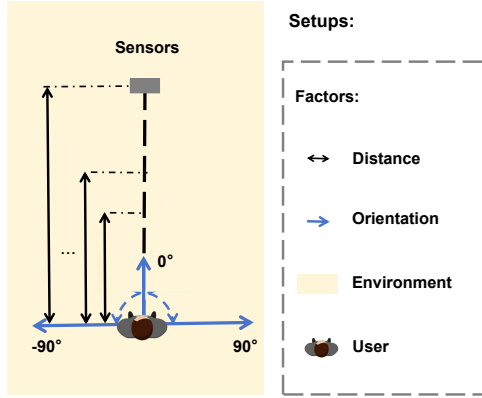


Fig. 4. Factors that may affect model performance. A setup is a combination of the four factors.

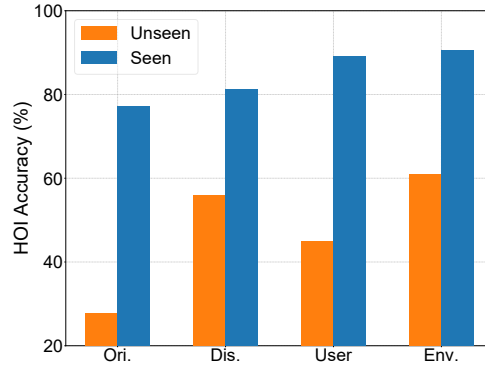


Fig. 5. Impact of unseen factors (Ori: orientation, Dis: distance, Env: environment).

their complementary contextual information to enhance overall performance. We will present our multimodal data preprocessing pipeline in Sec. 3.2 and our fusion network design in Sec. 3.3.

2.2 High data collection cost for HOI

RF-HOI faces three key challenges in applying empirical recognition models:

The novel RF modalities require customized data and annotation. Unlike prior works on HAR [18, 62], existing datasets cannot be repurposed, and there is no RF signal-based HOI dataset available. Most HOI datasets are designed for vision-based approaches and contain only images or videos. Meanwhile, existing RF-based HAR datasets typically lack object interaction scenarios and do not provide the necessary HOI annotation.

Data collection needs to cover diverse setups. For comprehensive evaluations, data collection must include a wide range of *setups* — combinations of user orientation, distance to the device, background environment, and body shape [14, 140], as illustrated in Fig. 4. Inadequate coverage of these factors in the training set can significantly impact HOI recognition performance. To demonstrate this, we conducted a preliminary study using data that varies across two environments, two users, three distances, and two orientations. For each factor, we reserved all samples with one specific value as the unseen factor from the training set and used them for testing, while the baseline included all factors in training. As shown in Fig. 5, HOI accuracy drops by 49.41% for unseen orientation, 25.32% for unseen distance, 44.06% for unseen user, and 29.69% for unseen environment, highlighting the importance of diverse setup coverage.

HOI tasks introduce many categories with the combination of actions and objects. However, despite previous works on HAR [17, 113, 140], collecting sufficient and diverse data for HOI recognition is even more challenging. Considering the same number of actions, HOI recognition usually has more categories than HAR, since one action may have multiple candidate target objects. As a consequence, we have to collect and annotate more samples if we want to cover every possible setup for each HOI category [66]. Suppose we want to cover three environments, five users, four distances, and seven orientations for seven actions with three candidate target objects on average. Then, if we repeat each HOI category five times in each setup, the total number of samples we need to collect will be the product of the above numbers, *i.e.*, $3 \times 5 \times 4 \times 7 \times 7 \times 3 \times 5 = 44,100$. In fact, this number has exceeded the size of many large-scale benchmarks for HAR, with only a few factors considered.

Given the high cost of real-world data collection, and inspired by numerous HAR studies leveraging synthetic data [2, 9, 14, 22, 23, 61], *RF-HOI* adopts synthetic data to enhance model performance with limited real-world

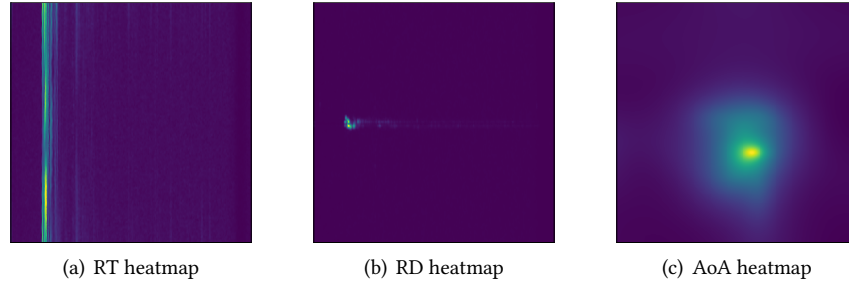


Fig. 6. Examples for mmWave heatmaps.

samples, thereby reducing data collection effort. We detail the simulator design and the integration of synthetic data in Sec. 3.4.

3 Methods

3.1 Overview

As shown in Fig. 2, *RF-HOI* comprises three main components: real-world data collection, data synthesis, and HOI recognition modeling. We first collect mmWave and RFID data in real-world settings, followed by RF signal preprocessing through our **data preprocessing module**. Subsequently, a **modality fusion model** is designed to integrate the complementary strengths of both sensing modalities. To address the limitation of real-world data, we use our proposed ***RF-HOI* simulator** to generate large-scale synthetic data. The model is initially pre-trained on this synthetic data to enhance generalizability, then fine-tuned using a small subset of real-world samples, and finally evaluated in real-world scenarios.

3.2 Data Preprocessing

RF-HOI leverages both mmWave and RFID signals to identify human actions and target objects. However, these raw signals are often noisy, high-dimensional, and complex-valued, which complicates the design of suitable neural networks [142]. To address this, we introduce a dedicated data preprocessing module in *RF-HOI* that processes the raw signals, enabling the effective application of established model architectures such as convolutional neural networks (CNNs) and transformers [111].

3.2.1 Preprocessing mmWave data. We adopt spectrum-based mmWave representations as a task-driven design for RF-HOI, using heatmaps to capture dense temporal measurements of motion-related range and velocity responses. We acknowledge that mmWave point clouds [103, 128, 136] provide an alternative representation and remain a promising direction for future work. Specifically, the raw mmWave radar data consists of Intermediate Frequency (IF) signals with dimensions $(N_{frame}^M, N_{chirp}, N_{Tx}, N_{Rx}, N_R)$, representing the number of frames, chirps, transmitting antennas, receiving antennas, and samples per chirp, respectively. To begin, we apply a static clutter removal algorithm [128] to mitigate background interference. Following standard processing pipelines [47], we then compute a 2D Range-Time (RT) heatmap for distance features using Range-FFT, and a 2D Range-Doppler (RD) heatmap for velocity features from each radar frame using Doppler-FFT. Finally, we extract the Angle of Arrival (AoA) heatmap using the MUSIC algorithm [102]. As illustrated in Fig. 6, this process yields three 2D heatmaps per radar frame, which can be directly fed into the proposed neural network model.

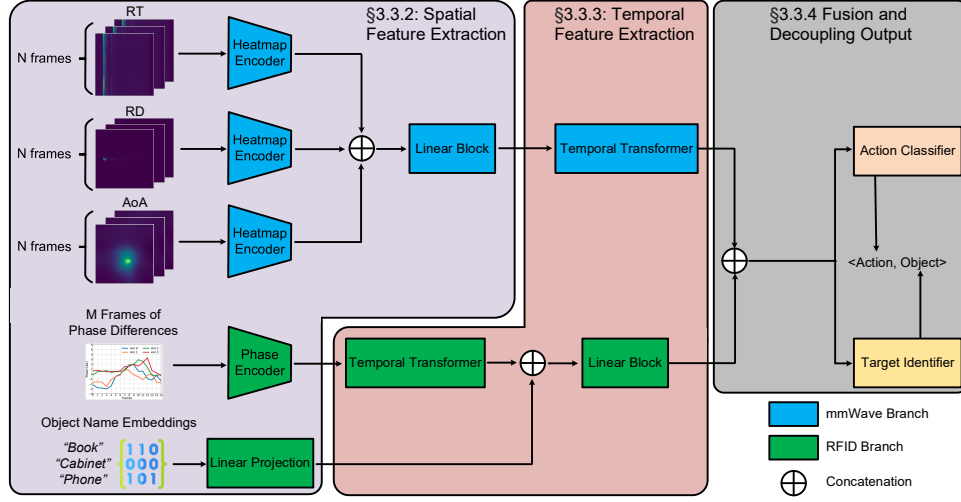


Fig. 7. Structure of the modality fusion network. We carefully designs spatial and temporal feature extraction for each modality, and then fuses the two asynchronous modalities at the feature level. The outputs are decoupled by two branches to improve generalizability and interpretability.

3.2.2 Preprocessing RFID data. Suppose each candidate object is equipped with an RFID tag. The RFID reader generates a table where each entry records the timestamp, tag ID, antenna ID, signal amplitude, and signal phase.

From this table, we first extract phase frames for each antenna–tag pair at each time step. *RF-HOI* primarily utilizes the signal phase, as it is more fine-grained and reliable than amplitude [7, 80]. To account for the asynchronous nature of RFID readings—resulting in different frame counts across pairs—we use nearest neighbor interpolation for alignment [121], followed by a moving average for smoothing. Note that modeling the RFID phase $\varphi(t)$ is challenging due to unknown multipath conditions, and accurate RFID localization typically requires expensive hardware [67]. However, since our focus is on kinematic features of the target objects, we use the phase difference between consecutive frames, $\Delta\varphi(t)$, as it effectively indicates movement direction by its sign.

Additionally, to leverage semantic information from RFID, we extract a tag table mapping tag IDs to object names, and convert these names to word embeddings using spaCy [39]. This allows us to input meaningful object semantics into the neural network.

3.3 Modality Fusion Network

3.3.1 Challenges. For each sample, after preprocessing, the data includes a matrix of $(N_{frame}^M, 3, N_{map}, N_{map})$ for mmWave heatmaps, a matrix of $(N_{frame}^R, N_{tag}, N_{ant})$ for RFID phase differences, and object name embeddings with a shape of (N_{tag}, N_{embed}) , where N_{map} is the heatmap size, N_{frame}^R is the number of RFID phase difference frames, N_{tag} is the number of tags, N_{ant} is the antenna number, and N_{embed} is the embedding dimension.

Fusing these two modalities is challenging due to their heterogeneity in both spatial and temporal dimensions. In the spatial dimension, mmWave data provides three heatmaps, while RFID data offers phase differences $\Delta\varphi$ for each tag-antenna pair, along with object name embeddings. The input dimensions and underlying physical meanings differ significantly, which raises challenges in how to design the model to encode the information of each modality.

In the temporal dimension, mmWave radar has a fixed frame rate, but RFID readings are asynchronous and vary over time [121]. As a result, each radar frame may correspond to a different number of RFID frames, and real-world HOI samples may have varying lengths and frame counts. Therefore, it is difficult to process and fuse both modalities' temporal features at the frame level.

Moreover, directly outputting HOI categories with a single branch would limit the generalizability of the model and cannot support unseen HOI categories (unseen action-object pair) when the constituent actions and objects are seen. For example, the model cannot recognize <pick up, book> even when it has seen <pick up, bottle> and <open, book> in the training data. In addition, coupling the action and object in one output leads to poor interpretability since it is difficult to analyze the performance on action recognition and target identification separately [66, 144].

To address these challenges, we propose a fusion network, as illustrated in Fig. 7. We first extract spatial and temporal features independently for each modality, and then fuse these representations for joint action classification and target identification. Further, the proposed network decouples the output of actions and objects with two separate branches for a better generalizability and interpretability.

3.3.2 Spatial features extraction. After preprocessing, each radar frame contains three heatmaps of size $N_{map} \times N_{map}$. To extract features related to range, velocity, and angle, we feed each heatmap into a dedicated heatmap encoder. As illustrated in Fig. 6, each heatmap resembles an image, making convolutional neural networks (CNNs) an effective choice for pattern extraction, as CNNs excel at capturing patterns in mmWave heatmaps [127]. Each heatmap encoder consists of two single-channel CNN layers followed by a max-pooling layer. The outputs from the three encoders are then concatenated and passed through a two-layer linear block to integrate features across all heatmaps.

For the RFID branch, we obtain a matrix of shape $(N_{frame}^R, N_{tag}, N_{ant})$, comprising the $\Delta\phi$ phase differences for each tag-antenna pair at each frame. We observe that the phase differences from one tag to multiple antennas form a coordinate-like vector related to the tag's position. Inspired by prior work with 3D coordinates [128], our phase encoder uses linear layers to project the input into a high-dimensional space, providing better capacity to distinguish motion patterns. To ensure stable training and effective feature normalization, we use Layer Normalization [3] instead of the commonly used Batch Normalization [46] within each block. It is because Layer Normalization normalizes along the feature dimension only, while Batch Normalization operates across batches and can disrupt temporal dependencies, potentially reducing model performance.

3.3.3 Temporal features extraction. To address the varying frame rate of RFID data, we extract temporal features independently within each modality and then fuse them via concatenation.

For the mmWave branch, after spatial feature extraction, we obtain a matrix of shape (N_{frame}^M, N_{dim}^M) , where N_{frame}^M is the number of frames and N_{dim}^M is the feature dimension for each frame. We add trigonometric positional encodings [111] and process the sequence through three Transformer Encoder layers. To summarize temporal information, mean pooling is applied across the time dimension, resulting in a vector of size N_{dim}^M for each sample.

For the RFID branch, we apply a similar temporal transformer to obtain a vector of dimension N_{dim}^R . The word embeddings of object names are projected using a linear layer into the same feature space, N_{dim}^R . We concatenate the temporal features and projected embeddings, and further align the vector dimension to N_{dim}^M using a linear block, preparing the features for fusion with the mmWave branch.

3.3.4 Fusion and decoupling output. After processing within each modality, we concatenate the features from mmWave and RFID modalities and feed them into two output branches: an action classifier and a target identifier. In this way, both branches utilize information from both modalities for decision-making, while disentangling the two outputs for a better generalizability and interpretability. Each output branch is composed of linear layers and produces either the action category or the target object, respectively. Let $loss_{action}$ and $loss_{target}$ denote the

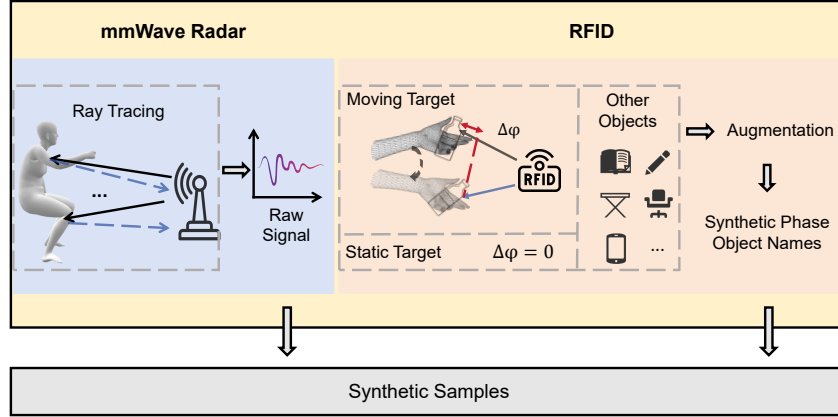


Fig. 8. Overview of the RF signal simulator. Our simulator can synthesize both mmWave and RFID data given a mesh sample.

cross-entropy losses for the action classifier and target identifier. The total loss is defined as their weighted sum:

$$loss_{total} = loss_{action} + \alpha \cdot loss_{target}, \quad (1)$$

where α is a hyperparameter that balances the two losses. Finally, the model outputs the predicted action category and the tag ID of the target object among the input RFID matrix.

3.4 Generating Synthetic Multimodal RF Data

To boost model performance with limited real-world training data and enhance generalizability to unseen factors, we propose generating large amounts of synthetic data using a custom-built simulator. As illustrated in Fig. 8, the proposed simulator supports multimodal RF data synthesis, *i.e.*, both mmWave and RFID data. In this section, we detail the design of our simulator.

3.4.1 Determine principles of simulation. Existing RF signal simulators for wireless sensing are typically based on either data-driven methods [2, 16, 22] or physics-based methods [9, 14, 23, 54, 61]. Although data-driven simulators are versatile, their models are usually trained on non-HOI datasets and lack interpretability. As a result, synthesizing HOI data using data-driven approaches may lead to inconsistent representations between modalities, such as mismatched object movements and user actions. To ensure both explainability and reproducibility, we adopt a physics-based approach to build our RF signal simulator.

3.4.2 Prepare input to the simulator. We choose ray tracing to simulate mmWave signals, as it is the most widely used technique among physics-based methods. However, ray tracing requires well-annotated 3D mesh data as input, which is particularly challenging for HOI tasks. Unlike HAR tasks—where generative models can produce human mesh data—generating meshes for HOI is more difficult due to the added complexity from diverse objects and their geometries [24, 94, 125]. This complexity can cause mismatches between human motion and real object movements, making datasets from motion capture systems preferable for realism.

Fortunately, several HOI datasets with human mesh have been released in the computer vision community [49–51]. Since our simulator can accept input from any HOI dataset that provides 3D mesh data and HOI category annotations, it is compatible with datasets either from motion capture or generative models in the future. Among the options, we select the TRUMANS dataset [51] as the most suitable, because it is fully public and richly

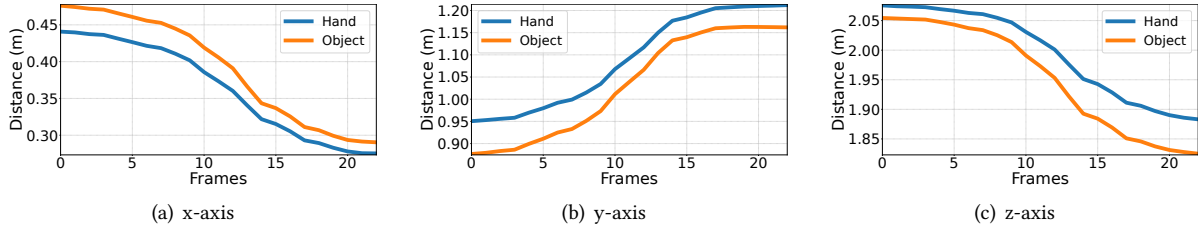


Fig. 9. The coordinates in three axes of the object and the hand in a ‘pick up’ action.

annotated with action names, target object names, and details about each interaction—whereas other datasets generally lack object information.

3.4.3 Build the simulator. After preparing the mesh dataset, we develop a physics-based simulator for *RF-HOI* that jointly synthesizes mmWave and RFID data. Specifically, our mmWave signal synthesis module builds upon RFGGen [14], an open-source ray tracing framework originally designed for HAR. Instead of using generative models for human mesh, we adapt this framework to accept 3D mesh data from public HOI mesh datasets. Although the heatmap representations used in our system contain reflection intensity information, we do not aim to precisely model absolute radar cross section (RCS) [100] in simulation. Instead, all heatmaps are normalized to the $[0,1]$ range using min-max normalization, which substantially reduces the influence of absolute intensity values that depend on distance, angle, and material properties. This design choice encourages the model to focus on relative spatiotemporal patterns and motion-induced structures that are more consistent between simulated and real-world data, rather than relying on absolute reflection strength. We note that spectrum-based representations have been widely adopted in prior radar sensing work [9, 53, 129], and these studies have also demonstrated that simulated radar data can effectively enhance downstream performance when combined with appropriate domain adaptation strategies.

However, *RF-HOI* requires not only mmWave data but also RFID data, which is not supported by any existing tools. Moreover, ray tracing cannot be directly applied for RFID simulation due to two reasons: (1) Ray tracing introduces large errors with RFID frequencies due to geometric approximations [109], as RFID operates at much lower frequencies than mmWave; and (2) Accurate object coordinates are needed to correctly position meshes for ray tracing, yet only a small portion of the TRUMANS dataset includes such annotations. This constraint would significantly reduce the number of usable samples for ray tracing if applied to RFID.

To address the challenges above, we design the RFID data synthesis module to directly simulate RFID signal phase differences based on physical principles, focusing on object movement. We use phase differences ($\Delta\phi$) to capture the kinematic features relevant to our model. In this way, we do not require explicit object position annotations, since $\Delta\phi(t)$ depends solely on the distance change between adjacent frames. For static objects, we set $\Delta\phi = 0$. For moving objects, the TRUMANS dataset provides information on which hand the user uses during interaction, and the SMPL-X human body mesh [92] supplies precise hand locations. As shown in Fig. 9, the hand and the target object’s trajectories are highly similar during interaction. Thus, we use the corresponding hand location as a proxy for the object’s location to calculate $\Delta\phi$. Additionally, for non-target candidate objects, we assign them random object names from TRUMANS and set $\Delta\phi = 0$ for their phases.

To bridge the gap between synthetic and real-world data, we apply three data augmentation steps: (1) Simulate packet loss by randomly dropping samples, emulating EPC Gen2 RFID standard collisions [124]; (2) Add Gaussian white noise to the synthesized phase signals; and (3) Normalize the data to reduce the impact of outliers. Fig. 10

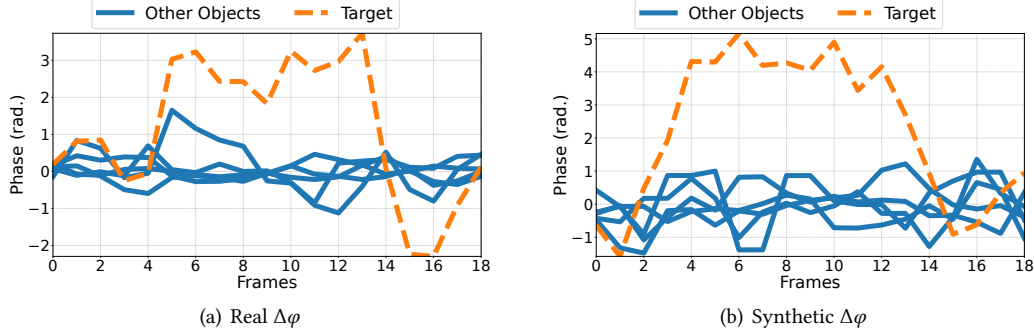


Fig. 10. Comparing normalized tag phase difference in a ‘pick up’ action.

Algorithm 1: Synthetic object category expansion

Input: Original object categories O_{old} ; new object categories O_{new} ; object-agnostic action list $\mathcal{A}$; original synthetic dataset $\mathcal{D}$

Output: Expanded synthetic dataset $\mathcal{D}^*$

```

1  $O_{all} \leftarrow O_{old} \cup O_{new}$ ;
2  $\mathcal{D}^* \leftarrow \mathcal{D}$ ;
3 foreach sample  $(a, o) \in \mathcal{D}^*$  do
4   if  $a \in \mathcal{A}$  then
5     Randomly sample a new object category  $o' \sim O_{all}$ ;
6     Replace the original object category  $o$  with  $o'$ ;
7     Update the corresponding object name embedding using  $o'$ ;
8   end
9 end
10 return  $\mathcal{D}^*$ 

```

presents a comparison of our synthesized $\Delta\phi$ during a pick-up action with real-world measurements (using one antenna and six objects). In both cases, the target’s $\Delta\phi$ is clearly distinguishable from those of other candidates.

3.5 Generalization to New Object Categories

In HOI recognition, the space of object categories is typically much more diverse than that of action categories, as similar interaction dynamics (e.g., *pick up* or *put down*) can involve a wide range of objects. This makes generalization to unseen object categories particularly challenging.

3.5.1 Challenges. The difficulty mainly arises from two aspects. First, in conventional formulations that treat each HOI as an independent class, the model is required to recognize previously unseen action–object pairs, which correspond to unseen categories at test time. Second, wireless signals are highly affected by environmental factors and object properties, making it non-trivial to disentangle motion-related patterns from object-specific variations without explicit modeling. To address these challenges, we design *RF-HOI* to support generalization to new object categories from both the model and simulation perspectives.

3.5.2 Model perspective. An intuitive approach to HOI recognition is to treat each HOI category as an independent category and train a single-branch classifier over all combinations. However, such a formulation cannot generalize to unseen object categories, as each new action-object pair corresponds to an unseen category. In contrast, our model adopts a decoupled output design that separates action recognition from object identification. This design is motivated by the observation that many HOI categories share common interaction dynamics across different categories of objects. The object branch predicts the index of the target object within the candidate set for each sample, rather than assigning a fixed global category label.

In addition, the RFID branch uses an input representation consisting of phase-based temporal features together with object identity embeddings. For interaction primitives such as *pick up* and *put down*, the motion-induced phase variations are primarily determined by interaction dynamics rather than specific object instances. As a result, objects with similar interaction patterns can share transferable representations.

This enables action representations to generalize across objects, while object identity is inferred separately from the RFID branch.

3.5.3 Simulation perspective. To further support generalization beyond the limited real-world dataset, we extend object categories at the simulation level. Based on the observation that many interaction primitives are largely object-agnostic in terms of motion dynamics, the simulation pipeline enables systematic reassignment or expansion of object categories associated with the same interaction primitives. In practice, each action can be paired with either existing or newly introduced object categories during RF data synthesis, without requiring additional motion capture or mesh reconstruction.

3.5.4 Putting everything together. To extend *RF-HOI* to unseen object categories, we first expand the synthetic dataset, as summarized in Algorithm 1. Specifically, we generate synthetic samples using our simulator and update the object categories with the expanded object set. The resulting synthetic dataset is used to pre-train the model, followed by fine-tuning on the same real-world dataset containing only the original object categories. The model is then evaluated on real-world samples with unseen object categories.

4 System Implementation

4.1 Data Collection Platform

4.1.1 Hardware. We build the *RF-HOI* hardware platform using commercial components to collect real-world data. As shown in Fig. 11(a) and Fig. 11(b), our setup includes a TI AWR1843AOP mmWave radar [45], an Impinj R420 RFID reader [43] with four antennas [82], passive UHF RFID tags [90], and a Femto Bolt RGB camera [88]. In Fig. 11, we mark the attachment of RFID tags with yellow rectangles. The camera and the mmWave radar are placed closely at about one meter above the ground, surrounded by the RFID antennas.

The mmWave radar transmits Frequency Modulated Continuous Wave (FMCW) signals with a frame length of 0.1 seconds, operating as a 3TX 4RX MIMO system (emulating a 3×4 virtual array with $1/2 \lambda$ antenna spacing). Each frame consists of 128 chirps (65 μ s duration, 256 sampling points per chirp), using a 3.9 GHz bandwidth swept from 77 to 80.9 GHz. This configuration supports a maximum sensing range of 11 meters, range resolution of 4.3 cm, maximum detectable velocity of 4.5 m/s, and velocity resolution of 7.1 cm/s. As is shown in Fig. 11(a), the radar and RGB camera are mounted 1 meter above the ground, while the RFID antennas are placed around the radar with 45 cm spacing.

The RFID reader is FCC-compliant and performs frequency hopping between 902.75 and 927.25 MHz across 50 channels. This induces phase discontinuities, which we calibrate by profiling the initial phase readings as described in [121].

4.1.2 Software. We developed data collection software in Python to control all devices and collect multimodal data. The RFID data acquisition module uses public Python APIs [89, 104]. Since the mmWave radar lacks a

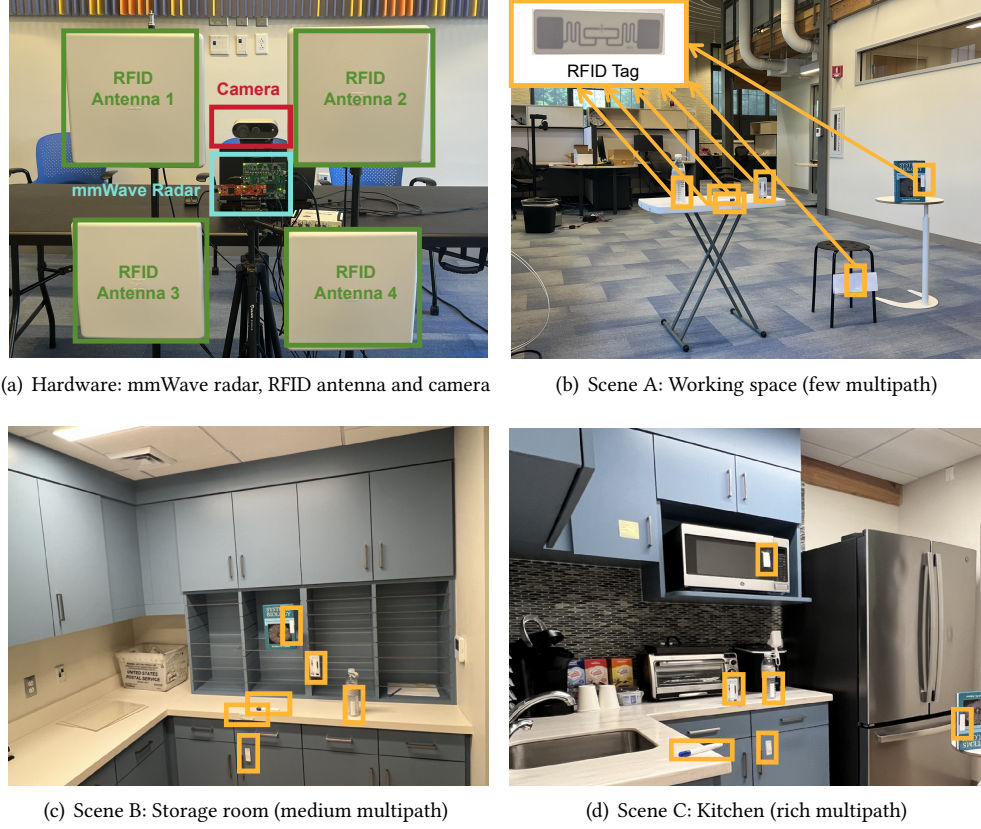


Fig. 11. Hardware platform and experiment setups. *RF-HOI* is evaluated in three different scenes with various multipath.

Python API but streams data via Ethernet, we capture data in real-time by sniffing the Ethernet interface. We employ multi-processing to accelerate data handling and compress raw images to JPEG format for memory efficiency. With these solutions, we enable real-time data collection from all sensors using a Dell G15 laptop [21] with 16 GB of memory.

4.2 Real-world Data Collection

As shown in Fig. 11, to evaluate the robustness of *RF-HOI*, we collect real-world data across three distinct scenes: a working space, a storage room, and a kitchen. Depending on the density of non-target objects, the three environments exhibit different degrees of static multipath interference. The working space has few multipaths, the storage room has medium multipaths, and the kitchen has rich multipaths.

In total, the dataset contains 17 HOI categories, including seven actions and eight objects, selected based on the interaction taxonomy of the TRUMANS dataset [51]. While TRUMANS includes additional HOI categories, we focus on those with sufficient samples to ensure reliable learning of interaction dynamics. Rather than exhaustively enumerating all interactions, we focus on fundamental action primitives (e.g., pick, put, open, and close) that capture core motion patterns and serve as building blocks for more complex behaviors. This design facilitates learning transferable representations and supports generalization to unseen HOI compositions and

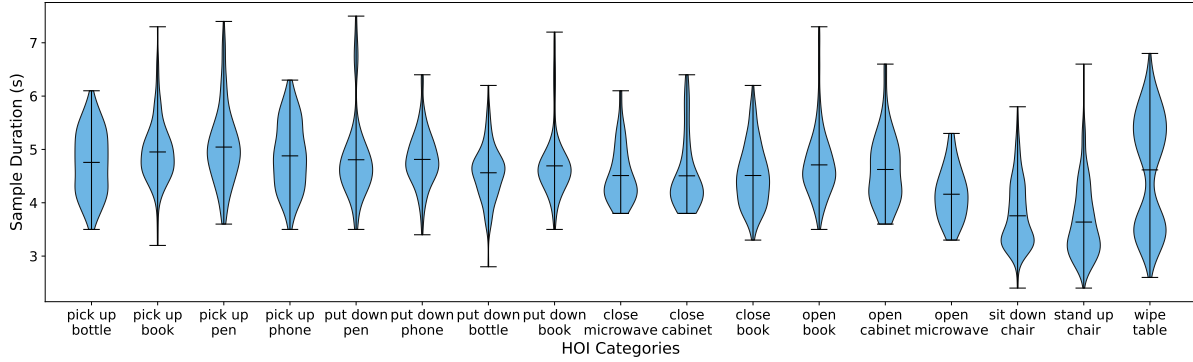


Fig. 12. Real-world sample durations of 17 HOI Categories. The proposed modality fusion network is designed to support samples with various durations.

object categories. During data collection, five other candidate objects are included in addition to the target object. One instance of each object type is present at a time, each equipped with an RFID tag. Fig. 12 shows the distribution of the duration of all samples in each category. We can see that in the real world, the duration distribution varies both within and across categories, which highlights the importance of supporting variable-length samples when we design the neural network.

We recruited five users (both male and female, heights ranging from 163 cm to 191 cm). For each sample, one user stands in front of the sensors and performs a single $\langle \text{action}, \text{object} \rangle$ tuple. The subject-to-sensor distance varies between 1 m and 3 m, within a $\pm 45^\circ$ -degree range in front of the sensors. Each action is collected in seven user orientations, ranging from -90° to 90° (see Fig. 4). Data from mmWave radar, RFID, and RGB camera are acquired for each sample.

In addition to HOI categories, each sample is annotated with its setup: a combination of environment, user identity, user-sensor distance, and user orientation. Data are collected across 78 unique setups, with each user repeating interactions 5 to 10 times in each setup. In total, the data collection process spans approximately 1.5 months and yields 3,615 real-world samples (617 GB).¹ The data are randomly partitioned, with 30% used for training and 70% for testing. This main dataset is used for overall evaluation of *RF-HOI*, including ablation studies and assessing robustness across diverse setups. In addition, we collect an extra set of 625 real-world samples dedicated exclusively to generalization evaluation, which are strictly excluded from any training or model selection process. This additional dataset is used to evaluate the model under more challenging conditions, including unseen object categories (five new objects leading to 10 unseen HOI categories), human-induced occlusions (from hands or body), and dynamic environments with multi-user interference.

4.3 Data Synthesis Detail

We utilize the TRUMANS dataset [51] to provide mesh data for our simulator. TRUMANS contains 15 hours of motion-captured data with fine-grained HOI annotations, featuring four female and three male participants and a variety of target objects—including both static items (e.g., chairs) and dynamic items (e.g., bottles).

Prior to data synthesis, we filter out samples that do not match the categories shown in Fig. 12, and further remove around 20% of samples that are either too short or too long. To enrich and expand the dataset, we

¹The data collection was approved by the IRB of the authors' institution.

systematically vary the distance between the human body mesh and the sensors from 1 to 3 m, and place sensors at different orientations relative to the human mesh for each sample.

Using the *RF-HOI* simulator, we generate 7,518 synthetic samples comprising mmWave and RFID data with well-annotated labels. The simulator also supports flexible object category assignment for generating diverse HOI samples, which is later used to study generalization to unseen objects. By default, the synthetic dataset is randomly split in half for training and validation.

4.4 Model Implementation and Training

We implement our proposed model (Sec. 3.3) using PyTorch [91], setting $N_{map} = 256$, $N_{embed} = 300$, $N_{dim}^M = 1024$, and $N_{dim}^R = 128$. Each Transformer Encoder Layer in the temporal transformer block has two heads, and we use LeakyReLU [79] as the activation function. The hyperparameter α is set to 0.5 to balance the two losses.

Different from some prior works [9, 113, 142], our model handles variable-length sequence inputs by leveraging transformer blocks for temporal feature extraction. Since samples are processed in batches during training, we pad all samples within a batch to the same length and provide both the padded sequences and corresponding padding masks to the model.

For pre-training, the learning rate is set to 1×10^{-5} . We save the model weights after each epoch and select the checkpoint with the best performance on the validation set for subsequent fine-tuning. During fine-tuning, we use a learning rate of 1×10^{-6} . In both stages, AdamW [74] serves as the optimizer. Unless otherwise specified, all model training and evaluation are conducted on a single NVIDIA A100 GPU.

5 Experiments

5.1 Metrics

In our experiment, for each sample, the model must classify the action among seven categories and identify the target from six candidate objects. A sample's HOI recognition is considered correct only if both the predicted action and target object match the ground truth. We define the following metrics to evaluate performance:

- (1) **HOI Accuracy** is the ratio of test samples for which both the action and target object are recognized correctly.
- (2) **Action Accuracy** is the ratio of test samples for which the action classified correctly, independent of the target object identification.
- (3) **Object Accuracy** is the ratio of test samples for which the target object is identified correctly, independent of the action classification.

5.2 Advantages of Modality Fusion

We first evaluate the effectiveness of our modality fusion network by comparing the proposed *RF-HOI* with the following methods:

- (1) ***RF-HOI* w/o Fusion**: The concatenation of the two modality features is removed; the action classifier receives only mmWave features, while the target identifier receives only RFID features.
- (2) **RFID-only**: Only RFID features are used for both the action classifier and target identifier.
- (3) **mmWave-only**: Only mmWave features are used for both the action classifier and target identifier.

For each baseline, we first pre-train the model with the same amount of synthetic data as *RF-HOI*, then fine-tune the model on the same amount of real-world data. We vary the percentage of real-world training samples used in fine-tuning and repeat each experiment with five random seeds, reporting the average result to reduce variance due to randomness.

As is shown in Fig. 13(a), increasing the amount of real-world training data leads to performance improvements for all baselines, with *RF-HOI* consistently outperforming others. On average, the HOI accuracy of *RF-HOI* exceeds that of *RF-HOI* w/o Fusion, RFID-only, and mmWave-only by 21.29%, 32.90%, and 73.20%, respectively. Specifically,

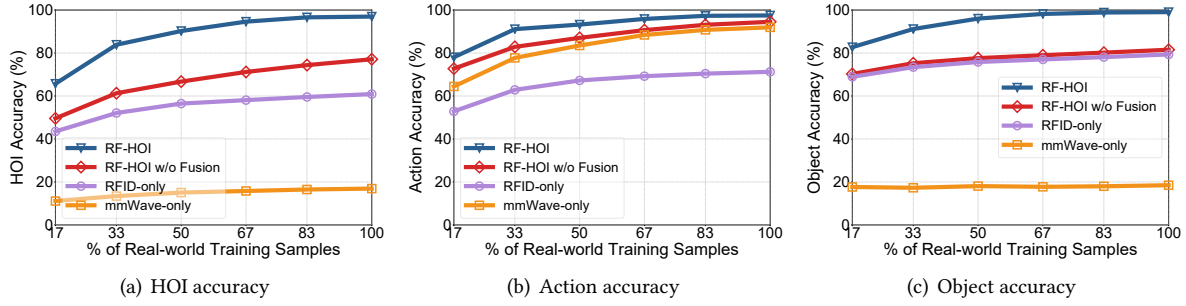


Fig. 13. Effectiveness of modality fusion.

when using all real-world training samples, *RF-HOI* reaches 96.97% HOI accuracy, which is 19.87%, 36.07%, and 80.06% higher than *RF-HOI* w/o Fusion, RFID-only, and mmWave-only, respectively.

To further elucidate the advantages of *RF-HOI*, we present the action accuracy in Fig. 13(b) and the object accuracy in Fig. 13(c). To be more specific, in terms of the action accuracy, *RF-HOI* outperforms that of *RF-HOI* w/o Fusion, RFID-only, and mmWave-only by 5.35%, 26.54%, and 9.38%, correspondingly. Similarly, with regard to object accuracy, *RF-HOI* outperforms that of *RF-HOI* w/o Fusion, RFID-only, and mmWave-only by 17.03%, 18.90%, and 76.47%, correspondingly. Across both metrics, *RF-HOI* consistently outperforms all other baselines, confirming the effectiveness of modality fusion for both action classification and target identification. Notably, using only mmWave yields high action accuracy (exceeding 90%), but its object accuracy is close to random guessing among six candidate objects (approximately 17%; see Fig. 13(c)). This limitation results in poor overall HOI accuracy. Conversely, using only RFID achieves higher object accuracy but significantly lower action accuracy than all other methods. Moreover, since features of interactive objects may reveal some action patterns and action patterns can help narrow the candidate targets, *RF-HOI* consistently achieves higher action accuracy than mmWave-only and higher Object accuracy than RFID-only.

Collectively, these results underscore the necessity of modality fusion and validate the effectiveness of the modal fusion design in *RF-HOI*.

5.3 Close Performance Between *RF-HOI* and Vision

Next, we evaluate the performance gap between *RF-HOI* and vision-based HOI recognition. Our vision baseline utilizes RGB videos as input, extracting image features with a pre-trained ResNet [35] and modeling temporal features using transformer blocks. Following InteractNet [31], the baseline includes two output branches for action and target object identification, respectively. Each experiment is repeated with five different random seeds to mitigate random variance.

Fig. 14 compares the results of *RF-HOI* and the vision-based method across varying percentages of real-world training samples. While the video baseline performs slightly better than *RF-HOI*, the average gap is only 2.50%. Furthermore, as the training set size increases, the gap narrows—reaching just 1.96% when using 100% of the training data. Nevertheless, *RF-HOI* maintains key advantages over the vision baseline by avoiding privacy concerns and remaining robust under poor ambient lighting conditions.

5.4 Effectiveness of Synthetic Data

As RF data collection for HOI recognition is expensive, synthetic data is important to enhance *RF-HOI*'s generalizability. In this section, we evaluate the effectiveness of the proposed simulator for multimodal RF data synthesis.

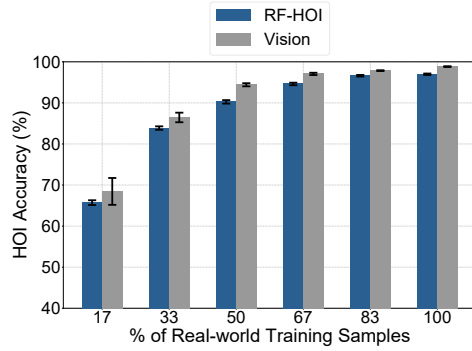


Fig. 14. Compare *RF-HOI* with the vision model in terms of HOI accuracy.

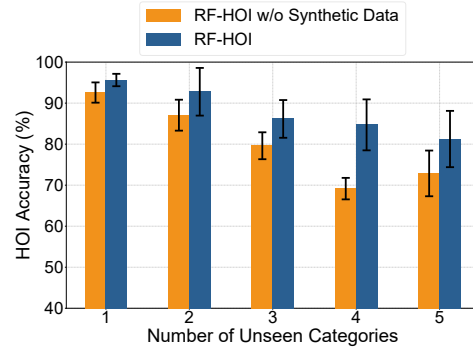


Fig. 15. Generalizability to unseen HOI categories.



Fig. 16. Five unseen objects: jar, lunchbox, bowl, cup, and eyeglass case.

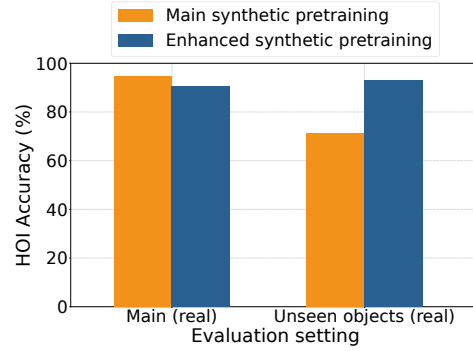


Fig. 17. Generalizability to unseen object categories.

For a fair comparison, we fine-tune *RF-HOI* on the same real-world training set used by the baseline, while the baseline trains the proposed *RF-HOI* model from scratch, without any synthetic data or pre-trained weights.

5.4.1 Generalizability to unseen HOI categories. First, benefiting from the decoupled output branches of actions and objects in our model design, we evaluate whether *RF-HOI* can generalize to unseen HOI categories, i.e., unseen action-object pairs, when the constituent actions and objects have been observed during training. This setting focuses on compositional generalization, which is important for scaling HOI recognition beyond a predefined set of categories.

To this end, we retain all HOI categories in the synthetic dataset, which provides broad coverage over action-object pairs. In contrast, for the real-world dataset, we randomly select varying numbers of HOI categories as unseen and reserve them exclusively for testing, while the remaining categories are used for training. Importantly, we ensure that all action categories and object categories remain present in the training data, so that only the action-object pairs are unseen. For a fixed number of unseen HOI categories, we repeat the random selection five

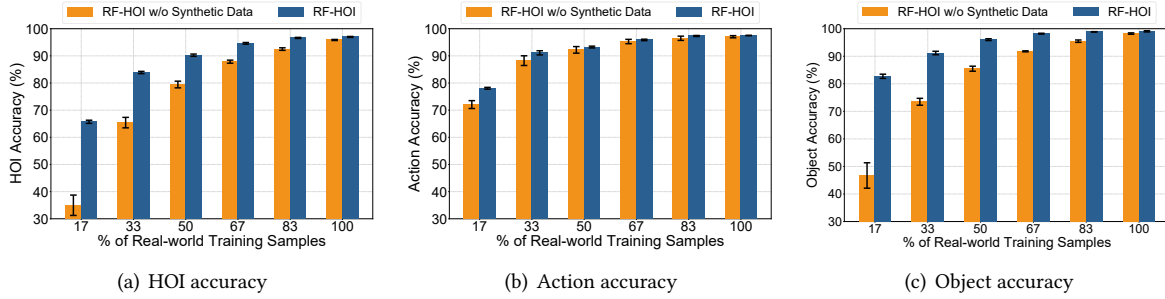


Fig. 18. Impact of training set size.

times to reduce sampling bias. Fig. 15 compares HOI accuracy between *RF-HOI* and the baseline. On average, *RF-HOI* outperforms the baseline by 7.84% in HOI accuracy and maintains an accuracy of 81.25% even when five HOI categories are completely unseen during training.

The results indicate that our design on decoupling action and object predictions enables effective compositional generalization, and that synthetic data further enhances this capability by expanding the coverage of the action-object pairs space.

5.4.2 Generalizability to unseen object categories. Second, we evaluate generalization to unseen object categories. In this experiment, we focus on ‘pick up’ and ‘put down’ actions, where the target objects never appear in the real-world training data, and where the underlying motion patterns are largely consistent across different objects.

As shown in Fig. 16, we introduce five additional objects: jar, lunchbox, bowl, cup, and eyeglass case. This corresponds to 10 unseen HOI categories in total, formed by five unseen objects and two actions (‘pick up’ and ‘put down’). We collect 145 real-world samples for evaluation only, covering three orientations, and do not use these samples in any training or model selection process. This evaluation set is constructed to specifically assess generalization to unseen object categories. In addition, we generate an enhanced synthetic dataset of 7,518 samples with expanded HOI categories following our design in Sec. 3.5, which provides sufficient coverage for pre-training with extended object diversity. Models are pre-trained either on the original synthetic dataset or on the enhanced synthetic dataset, and are fine-tuned using the same original real-world training data.

The results are shown in Fig. 17. Here, “Main (real)” and “Unseen objects (real)” denote evaluation on the original real-world dataset and the newly collected unseen-object dataset, respectively, while the two bars correspond to models pre-trained on the original and enhanced synthetic datasets. When pre-trained on the original synthetic dataset, the model achieves high HOI accuracy on seen objects (94.5%) but suffers a substantial drop to 71.3% on unseen objects. In contrast, pre-training on the enhanced synthetic dataset significantly improves accuracy on unseen objects to 93.1%, while maintaining comparable performance on seen objects (90.5%). The slight decrease on seen objects reflects a natural trade-off when expanding object coverage during pre-training, rather than a degradation of the model’s recognition capability.

These results demonstrate that *RF-HOI* can effectively generalize to new object categories without additional real-world training data, provided that the underlying action patterns are similar and that object-level expansion is appropriately incorporated at the simulation stage.

5.4.3 Performance boost under the same amount of fine-tuning data. Third, we demonstrate that pre-training with synthetic data significantly boosts *RF-HOI*’s performance, particularly when the available real-world training

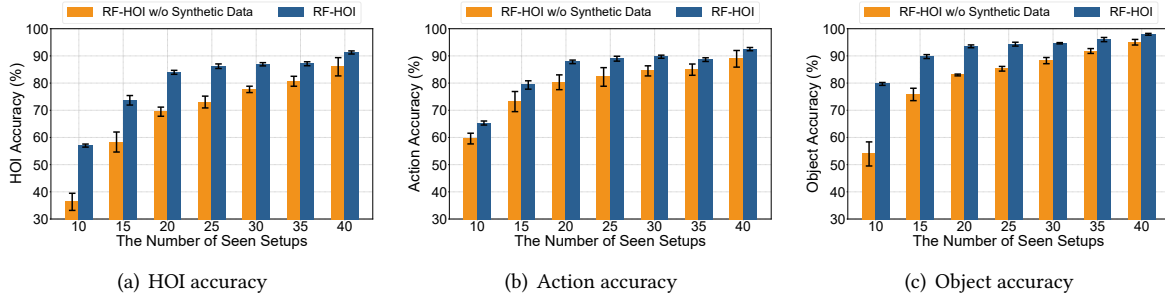


Fig. 19. Comparison between different numbers of unseen setups.

data is limited. Each experiment is repeated with five different random seeds to ensure reliability. By gradually increasing the percentage of real-world training samples, we report HOI accuracy, action accuracy, and object accuracy in Fig. 18. While both *RF-HOI* and the baseline improve as more real-world data are provided, *RF-HOI* consistently outperforms the baseline given the same number of samples, with average gains of 12.00% in HOI accuracy, 1.97% in action accuracy, and 12.48% in object accuracy. The benefit is especially pronounced with fewer training samples: when only 17% of real-world data is used, *RF-HOI* surpasses the baseline by 30.74% in HOI accuracy, 6.00% in action accuracy, and 36.01% in object accuracy.

5.4.4 Generalizability to unseen setups. Finally, we show that incorporating synthetic data improves *RF-HOI*'s generalizability to unseen setups when the number of seen real-world training setups is limited. To evaluate this, we randomly select varying numbers of setups as seen for training, with the remaining setups reserved for testing. For each setup in the training set, five samples are included. We ensure that every possible value for most factors (including environment, user, and distance) appears in the training data, but not necessarily in all combinations. For the orientation factor, we specifically include -90° , 0° , and 90° instead of all orientations. Each experiment is repeated with five different random seeds to ensure reliability.

As presented in Fig. 19, when the number of seen setups increases, both *RF-HOI* and the baseline improve, but *RF-HOI* consistently achieves higher accuracy on the test (unseen) setups. On average, *RF-HOI* outperforms the baseline by 12.09% in HOI accuracy, 5.50% in action accuracy, and 10.40% in object accuracy. The advantage is especially marked with fewer seen setups; for example, with only 20 seen setups, *RF-HOI* surpasses the baseline by 14.48% in HOI accuracy, 7.51% in action accuracy, and 10.54% in object accuracy.

5.5 Micro-benchmark Performance

Here, we would evaluate the performance of *RF-HOI* under various system settings.

5.5.1 Impact of fusion strategies. To validate the design choice of the modality fusion module in *RF-HOI*, we compare it with two widely adopted state-of-the-art fusion strategies for multimodal learning:

- (1) **FiLM** [26, 95]: Feature-wise Linear Modulation (FiLM), conditions the feature representations of one modality using linear layers as feature-wise affine transformations generated from the other modality.
- (2) **Attention** [44, 76]: Attention-based fusion models cross-modal interactions by allowing one modality to attend to the other. We implement a cross-attention-based fusion module at the representation level.

For all baselines and *RF-HOI*, we adopt a consistent training pipeline. Specifically, all models are first pre-trained on the same synthetic dataset, and then fine-tuned and evaluated on the same real-world dataset. During

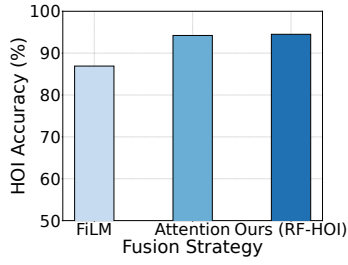


Fig. 20. Impact of fusion strategies.

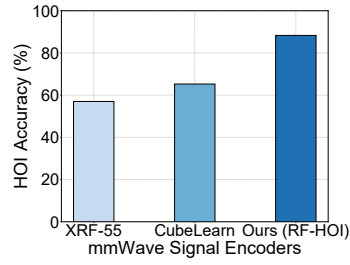


Fig. 21. Comparison for different mmWave signal encoders.

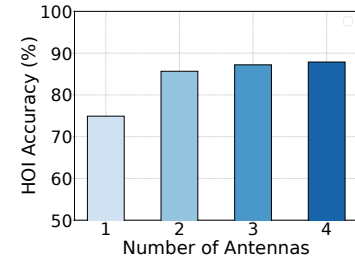


Fig. 22. Comparison for different numbers of antenna.

fine-tuning, we use 67% of the real-world training samples. To ensure a fair comparison, all models share identical modality-specific encoders and training protocols, with the fusion module being the only component that differs.

As illustrated in Fig. 20, *RF-HOI* achieves the highest accuracy among the three fusion strategies, while the performance gap between *RF-HOI* and attention-based fusion remains below 1%. This observation indicates that, under the current experimental setting, more complex fusion mechanisms provide only marginal improvements over our lightweight design. We note that the proposed framework is modular and can readily incorporate more expressive fusion designs in the future.

5.5.2 Impact of mmWave encoders. To validate our design for the mmWave branch, we compare *RF-HOI* with two state-of-the-art encoders proposed in previous HAR solutions:

- (1) **XRF-55** [113]: a multimodal network for HAR, where we substitute its mmWave branch into our model. Since XRF-55 uses CNNs for temporal feature extraction, input samples are cropped to a fixed length for compatibility.
- (2) **CubeLearn** [142]: a model that processes raw complex-valued mmWave data via a learnable preprocessing module, followed by LSTM [37] blocks to track temporal features.

All models are trained from scratch using 67% of the real-world training samples. As illustrated in Fig. 21, *RF-HOI* achieves 31.30% higher accuracy than XRF-55, and surpasses CubeLearn by 23.04%. The superior performance over XRF-55 can be attributed to *RF-HOI*'s transformer blocks, which handle variable-length samples more effectively. In comparison to CubeLearn, our design's efficient preprocessing module enables the model to learn decision boundaries more accurately when the dataset is limited.

5.5.3 Impact of the number of RFID antennas. In this section, we evaluate the impact of using different numbers of RFID antennas. The model is trained from scratch using 67% of the real-world training samples, with the number of antennas varied from one to four. As shown in Fig. 22, the HOI accuracies corresponding to 1, 2, 3, and 4 antennas are 74.91%, 85.66%, 87.20%, and 87.87%, respectively.

Notably, increasing the number of antennas from one to two yields the largest performance improvement (10.75%), while further increasing the array size from two to four results in only marginal gains (less than 3%). This suggests that, although we use four antennas in our default configuration to achieve the best performance, the proposed system does not fundamentally rely on a large antenna array. In practice, a compact configuration with two antennas already achieves competitive performance, substantially reducing system complexity and hardware cost while retaining most of the recognition accuracy.

5.5.4 Impact of balancing action classification and target identification. In this section, we evaluate the effect of α , the hyperparameter controlling the balance between $loss_{action}$ and $loss_{target}$ in Equation 1. The model is trained from scratch using 67% of the real-world training samples. As shown in Fig. 23, increasing α initially improves HOI accuracy, which then drops sharply beyond a certain point. More specifically, when the value of

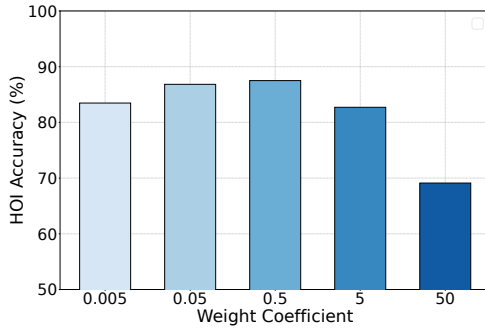


Fig. 23. Impact of weight coefficient α of loss function.

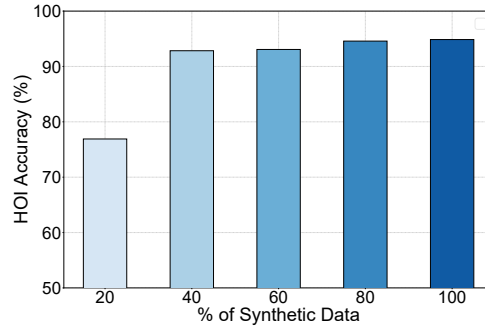


Fig. 24. Impact of pre-training dataset size.

α increases from 0.005 to 50, starting at 83.47%, the corresponding HOI accuracy first rises to 87.50% and then drops to 69.11%. Notably, a larger (or smaller) α causes the model to prioritize action (or target) classification, respectively. We therefore select a moderate value $\alpha = 0.5$, as both action and target identification are essential for robust HOI recognition.

5.5.5 Impact of the pre-training dataset size. In this section, we evaluate how the scale of the pre-training dataset affects performance. In this experiment, we vary the percentage of synthetic samples used for pre-training. After pre-training, all models are fine-tuned with 67% of the real-world training samples and evaluated on the real-world test set. As shown in Fig. 24, increasing the size of the pre-training dataset leads to a sharp improvement in HOI accuracy, especially when the number of synthetic training samples is limited. For example, the accuracy rises from 76.90% to 92.84% as the proportion of synthetic data increases from 20% to 40%. And the performance gains gradually saturate and plateau around 94%. This result indicates that our synthesized dataset is sufficiently large for current experimental settings; however, expanding the synthetic dataset may be necessary if more HOI categories or setups are introduced in the future.

5.6 Robustness to occlusion and dynamic environment

5.6.1 Impact of people occlusion. We evaluate the robustness of *RF-HOI* under three people-occlusion conditions:

- (1) **No occlusion:** All RFID tags remain unobstructed.
- (2) **Hand occlusion:** The user partially covers the RFID tag on the target object with fingers during interaction.
- (3) **Body occlusion:** Tagged non-target objects are placed behind the user, resulting in body-induced occlusion along the antenna–user–object path. In the body-occlusion setting, the distance between the user and the occluded objects ranges from approximately 20 to 50 cm.

For all three conditions, we collect a total of 270 real-world samples, covering three user orientations and four actions that involve direct hand–object contact: pick, put, open, and close.

We first analyze the impact of people occlusion at the signal level. Fig. 25 compares the RSSI distributions of the target object under no occlusion and hand occlusion. When the tag is partially covered by the user’s hand, the RSSI distribution exhibits a clear left shift, with an average attenuation of 8.30 dB. This result confirms that hand occlusion introduces substantial signal attenuation on the target tag, while still allowing the RFID signal to remain readable within the considered range. This observation is consistent with prior works [59, 63, 119], which show that RSSI is more susceptible to environmental interference than to motion patterns, and therefore provides limited discriminative power for interaction recognition.

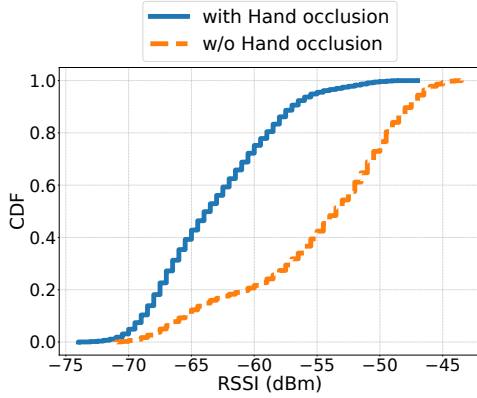


Fig. 25. Impact of hand occlusion on RFID RSSI.

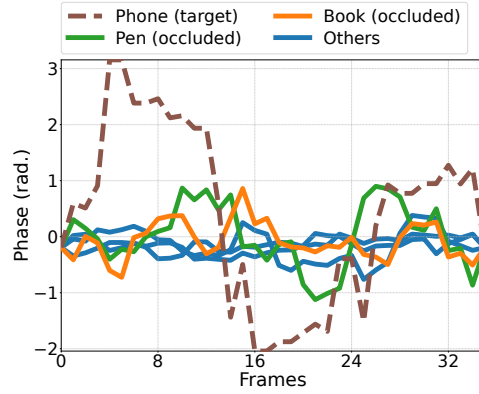


Fig. 26. Normalized tag phase differences under body occlusion.

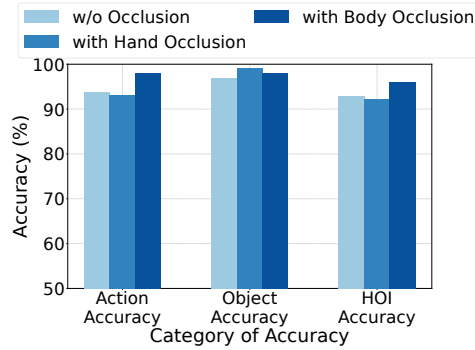
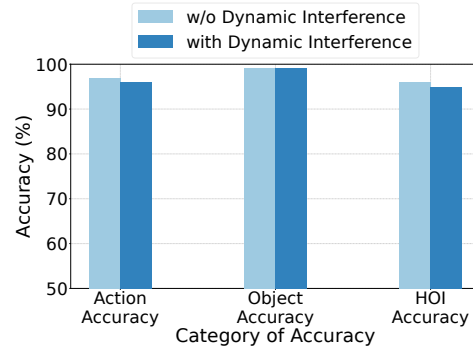


Fig. 27. Impact of human occlusion.

Fig. 28. Test *RF-HOI* with dynamic inference.

We note that the design choice of *RF-HOI* to use phase features is primarily motivated by their higher sensitivity to object motion, which provides more discriminative cues for interaction recognition, rather than by occlusion considerations alone. Fig. 26 visualizes one representative example of normalized RFID phase differences from one antenna to multiple objects under body occlusion. Although the non-target objects placed behind the user exhibit noticeable phase fluctuations due to body-induced dynamic occlusion, the target object (phone) consistently shows larger phase variations than the occluded non-target objects, such as pen and book. This preserves a clear contrast between the target and non-target objects.

We then present action, object, and HOI accuracy under the three occlusion conditions in Fig. 27. Across all settings, *RF-HOI* achieves over 90% HOI accuracy, indicating strong robustness to people-induced occlusion. We observe that hand occlusion only slightly degrades HOI accuracy, likely because the phase variations of the target object's tag remain observable, even though hand occlusion attenuates the RSSI. In addition, body occlusion does not degrade performance and even leads to a slight improvement. A possible explanation is that placing non-target objects behind the user reduces their interference, while the mmWave radar's line of sight to the human body remains unobstructed, allowing reliable capture of motion features.

Overall, these results demonstrate that *RF-HOI* can maintain stable HOI recognition performance under realistic people-occlusion scenarios, including both direct hand occlusion on the target tag and body-induced occlusion affecting non-target objects.

5.6.2 Impact of dynamic multipath interference. We evaluate the robustness of *RF-HOI* under dynamic and multi-user environments, addressing the concern that non-target objects or surrounding people may introduce time-varying multipath effects in real-world settings. Although *RF-HOI* focuses on single-user HOI recognition, it does not assume a static environment and is designed to tolerate moderate dynamic interference.

In this experiment, we ask one volunteer to perform HOI interactions in 105 samples, covering three user orientations and seven actions. To introduce dynamic multipath interference, a second volunteer is asked to perform random daily activities involving non-target objects (e.g., drinking from a bottle or typing on a keyboard) around the interacting user at a distance of approximately 1–2 meters. These objects are not equipped with RFID tags and only serve as dynamic reflectors in the environment. As a baseline, we also collect 105 samples under identical setups without dynamic multipath interference. Fig. 28 compares the performance of *RF-HOI* under the two conditions. The results show that the difference in HOI accuracy between static and dynamic multipath settings is within 1%. This indicates that *RF-HOI* can maintain stable recognition performance even in the presence of additional users and mild environment dynamics.

Overall, this experiment demonstrates that *RF-HOI* is robust to slight dynamic multipath interference and can generalize beyond strictly static, single-user environments, supporting its applicability in more realistic real-world scenarios.

5.7 System Computation Overhead

5.7.1 Evaluation on fine-tuning time cost. We evaluate the time required to fine-tune *RF-HOI* for real-world deployment while maintaining strong performance, highlighting its adaptability. Using 67% of the real-world training samples, we fine-tune the pre-trained model for varying durations. As shown in Fig. 29, the accuracy of HOI initially increases with fine-tuning time and steadily levels off after approximately 20 minutes. Specifically, as the fine-tuning duration increases from 5 minutes to 25 minutes, the corresponding HOI accuracy improves from 84.49% to 95.09%, with the most swift gain occurring in the first 15 minutes, and then converges. In practice, this enables rapid adaptation to new setups by collecting a small amount of real-world data for fine-tuning. Therefore, the low time cost for fine-tuning facilitates quick and flexible deployment of *RF-HOI* across various setups.

5.7.2 Evaluation on system latency. We measure the system latency of *RF-HOI*, including both data preprocessing and model inference times. Although the MUSIC algorithm is computationally intensive, we implement the mmWave preprocessing pipeline in PyTorch to leverage GPU acceleration, reducing the average preprocessing time for one sample from 13.39 seconds using CPU to 1.75 seconds using GPU on a Dell G15 laptop [21].

We then analyze the latency on systems equipped with either an NVIDIA L40S, NVIDIA A100, or (on the laptop) an NVIDIA RTX 4050 GPU. All experiments on the server are conducted using a single GPU, with 8 CPU cores and 128 GB RAM allocated per job. As illustrated in Fig. 30, both model inference and preprocessing are fastest on the L40S, followed by the A100, with the RTX 4050 being the slowest. On the laptop, the end-to-end latency averages 1.79 seconds per sample. Since the shortest sample duration in our dataset is 2.40 seconds, these results demonstrate the feasibility of *RF-HOI* for supporting real-time HOI recognition. In addition, Fig. 30 shows that the system latency is dominated by the preprocessing stage, while the model inference itself is relatively lightweight, taking only 36.86 ms per sample on the RTX 4050 GPU. We note that GPU performance is workload-dependent: in our preprocessing-dominated pipeline, the L40S shows better efficiency, likely due to architectural differences and optimizations for inference-oriented workloads.

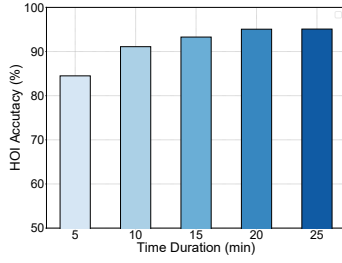


Fig. 29. Impact of fine-tuning duration.

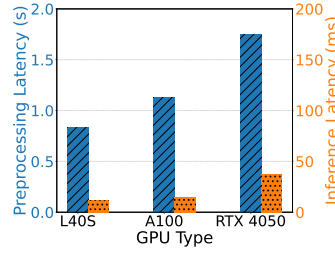


Fig. 30. Sample level latency on preprocessing and model inference.

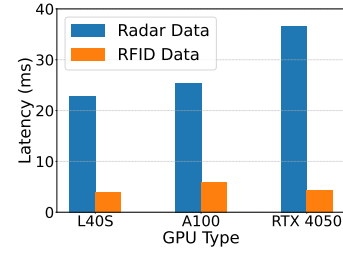


Fig. 31. Frame level preprocessing latency on mmWave radar data and RFID data.

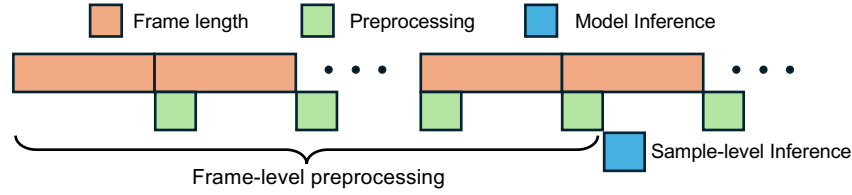


Fig. 32. Streaming pipeline of *RF-HOI* for real-time processing.

To further assess the system's responsiveness in an online setting, we analyze the frame-level preprocessing latency of each sensing modality in Fig. 31. As shown in the figure, the per-frame preprocessing cost of mmWave radar data ranges from 22.83 ms to 36.52 ms across different GPUs, while the corresponding cost for RFID data remains below 6 ms. The effective acquisition interval of an RFID frame—corresponding to one complete read of all tags—is not strictly fixed, but is typically close to 100 ms in our system. With a 100 ms frame interval for the mmWave radar and a comparable effective acquisition interval for RFID data, the frame-level preprocessing latency is well within the real-time budget. Therefore, by performing preprocessing incrementally as each frame arrives in parallel to data acquisition, the system can further improve its responsiveness in an online setting, as illustrated in Fig. 32.

5.8 Case Study: Goal Inference

To demonstrate the real-world applications of *RF-HOI*, we apply *RF-HOI* to infer humans' goals in household scenarios. Such goal inference capacity is crucial for engineering assistive home systems to help humans without the need for privacy-intrusive sensors, e.g., cameras. As is illustrated in Fig 1, given a sequence of human-object interactions observed so far, we ask a goal inference model to infer the goal of the human. Unlike prior work, we leverage *RF-HOI* to recognize the HOI sequence from privacy-preserving sensors instead of relying on video inputs. Following the baselines proposed in [135], we take LLMs as the goal inference models. We provide a predefined goal space for LLMs to choose from [97] and provide some pairs between HOI sequences and goals for few-shot prompting. The goal inference accuracy is defined as the proportion of correctly inferred sequences to the total number of sequences. We present our prompt as follows:

Human has been performing a task that involves executing a sequence of interactions. The goal of task can only be one of the following: {Candidate goals}.

You are a helpful assistant. In order to help human, please infer the most corresponding goals selected from the goal space for the following interaction sequences, in order: {Test HOI sequences}.

Hints: There is a set of example pairs showing the relationships between interaction sequences and their corresponding goals: {Few-shot prompting pairs between HOI sequences and goals}.

The order of actions in each sequence is very important. Each sequence has a one-to-one correspondence with its goal.

Output requirement: For each provided sequence, you have to choose the most relevant goal from the goal space, even if you are uncertain.

Table 2. Five examples of HOI sequences and corresponding goals.

HOI sequences	Goals
[<stand up, chair>, <pick up, book>, <open, book>, <pick up, pen>]	standing and taking notes in book with pen
[<sit down, chair>, <stand up, chair>, <pick up, phone>, <put down, phone>]	standing and answering the phone
[<open, microwave>, <close, microwave>, <open, microwave>, <close, microwave>]	using microwave
[<pick up, bottle>, <open, cabinet>, <put down, bottle>, <close, cabinet>]	trying to put bottle in cabinet
[<wipe, table>, <pick up, bottle>, <put down, bottle>, <wipe, table>]	cleaning table

In total, we define 30 different pairs between HOI sequences and goals, where each HOI sequence may have four or five HOIs. Table 2 shows five pairs as an illustration. Among all 30 pairs, eight pairs are for few-shot prompting, which is designed to help the LLM understand the goal inference task. The remaining 22 pairs are for testing. During the test, we ask the LLM to infer the corresponding goals for the HOI sequences in the test pairs from all 30 candidate goals.

We ask one volunteer to perform according to the 22 test HOI sequences in both the working space and the kitchen. We collect the corresponding RF data and get the recognized HOIs by *RF-HOI*. We compare three types of input for the goal inference:

- (1) **Action-only** only inputs sequences of actions recognized by *RF-HOI* to the LLM.
- (2) **Object-only** only inputs sequences of objects recognized by *RF-HOI* to the LLM.
- (3) ***RF-HOI*** inputs sequences of full HOI categories recognized by *RF-HOI* to the LLM.

We evaluate the performance of the three methods on three LLMs: Gemini 2.5 Flash [32], GPT-5 [87], and DeepSeek-V3.2 [20], and present the results in Table 3. We can see that for all LLMs, *RF-HOI* brings higher goal inference accuracy than Action-only and Object-only. Specifically, on average, *RF-HOI* outperforms action-only by 68.18% and object-only by 22.72% in goal inference accuracy.

We observe that, in the action-only setting, most failures stem from misidentifying the target object, whereas in the object-only setting, errors primarily arise from misinterpreting the intended action. For example, a

sequence like [<open>, <close>, <open>, <close>] can lead the LLM to confuse using the microwave with opening and closing a cabinet. Conversely, a sequence like [<book>, <book>, <book>, <book>] can cause the LLM to infer merely checking books rather than reading them, due to the absence of the <open> action. These results demonstrate the effectiveness of *RF-HOI* when adapted to downstream applications and highlight the importance of jointly recognizing actions and objects for a better user understanding.

Table 3. Goal Inference Accuracy

LLM \ Input	Action-only	Object-only	<i>RF-HOI</i>
Gemini 2.5 Flash	18.18%	59.09%	95.45%
GPT-5	27.27%	77.27%	86.36%
DeepSeek-V3.2	18.18%	63.64%	86.36%

6 Related Work

6.1 Vision-based Human-Object Interaction Recognition

Over the past decade, significant effort has been devoted to camera-based HOI recognition [12, 33, 78, 133], given its importance for scene and human understanding. Many subsequent works [31, 40, 98] extend object detection frameworks, typically structuring HOI recognition as a multi-task problem using a two-branch network: one branch detects humans and objects, while the other classifies actions and infers target objects through human-object relationships. Recent advancements include the use of sophisticated model architectures such as Transformers [66] and Mamba [126]. Other approaches introduce multimodal techniques, incorporating modalities such as text prompts [131]. Specifically, RF-camera [72] fuses vision and RFID modalities to recognize human gestures and achieve human-object matching. In addition, there is growing interest in collecting 3D HOI scene mesh data [24, 42, 49, 51, 94, 125], advancing HOI understanding from 2D to 3D environments.

However, vision-based approaches raise privacy concerns for in-home user sensing, and often suffer under poor lighting conditions. In contrast, *RF-HOI* leverages RF modalities, which are less invasive to privacy and robust to ambient lighting variations.

6.2 Wireless Sensing

The advancement of deep learning has facilitated rapid progress in wireless sensing systems for human activity recognition [9, 70, 103, 113, 116, 117, 132, 136, 140, 142]. To achieve vision-level perception while preserving user privacy, prior work has explored various wireless modalities, including mmWave [9, 70, 103, 117, 136, 142], RFID [28, 41, 108, 113, 119], and WiFi [116, 132, 140], for sensing human activities. To improve user convenience, most systems operate in device-free settings, although some RFID-based approaches require tags attached to the user's body [52, 114, 115, 120, 122, 130]. However, identifying specific objects in such settings remains highly challenging [71, 137], let alone jointly recognizing both human actions and their corresponding target objects.

To address the challenge of object identification, a line of prior work adopts device-based approaches by attaching RFID tags or similar markers to objects and inferring user actions from changes in the communication channel between tags and readers [30, 60, 96, 106, 119, 139]. These approaches avoid instrumenting users and have been widely explored in RFID-based sensing. However, they often struggle to distinguish actions that do not induce significant channel variations, such as interactions without object movement or direct tag coverage, thereby limiting the diversity of recognizable actions. RF-Diary [27] combines radar sensing with a pre-defined

floor map of object locations to recognize HOI. While effective, maintaining accurate object maps in everyday environments remains challenging.

Other works [56–58] employ acoustic signals collected by wearable devices to sense HOI, yet these approaches require wearables and fail to capture whole body movements, such as stand and sit. Recent works [55, 107] explore HOI recognition with modality fusion, but their solutions require access to personal devices, such as audio and IMU data from smartphone, while *RF-HOI* does not assume any devices on users.

In contrast, *RF-HOI* combines the strengths of both strategies by fusing mmWave and RFID data, enabling effective recognition of both human actions and target objects without compromising user convenience.

6.3 Wireless Signal Synthesis for Sensing

Recent studies have explored the use of synthetic wireless signals to augment training data for human activity recognition. Technically, these works employ either ray tracing-based methods [9, 14, 23, 54, 61] or data-driven approaches [2, 16, 22]. As a physics-based method, ray tracing models signal propagation by computing paths affected by reflection, diffraction, and scattering using physics and geometry formulas. In contrast, the simulation of data-driven approaches usually relies on empirical models, which lack physical interpretability, and may produce unrealistic outputs without physical priors.

However, prior research typically models only the human body as the primary signal reflector, treating other objects in the environment as interference. This makes it difficult to extend such techniques to HOI recognition tasks. In contrast, *RF-HOI* explicitly simulates separate signals from both the human body and target objects, enabling the generation of multimodal synthetic data tailored for HOI recognition.

7 Discussion and Future Work

7.1 Increase the Size and Diversity of the Data

As an initial effort to explore RF-based sensing for HOI recognition, *RF-HOI* collects data at a moderate scale and focuses on a set of common HOI categories encountered in daily life under diverse setups and sensing conditions. Compared with large-scale vision-based HOI datasets, our real-world dataset remains limited in scale, particularly in the number of participants, object categories, and interaction types. This limitation comes from the cost of collecting mmWave–RFID data under various real-world setups, and motivates future efforts toward larger community benchmarks for RF-based HOI recognition. We acknowledge that expanding the number of HOI categories and participants is an important direction toward broader real-world deployment. However, extending to entirely new action categories in the current simulation pipeline is constrained by the availability of realistic motion priors in the underlying TRUMANS dataset, which limits the ability to synthesize physically plausible RF data for additional action categories. The selected action categories (e.g., pick, put, open, and close) are therefore intentionally chosen as fundamental interaction primitives. These primitives are sufficient to support a range of downstream tasks under state-of-the-art frameworks, such as goal inference [97, 135, 141]. Rather than solely increasing the category count in the training set, we evaluate *RF-HOI* along more challenging generalization axes, including unseen object categories and unseen action-object combinations, without using any additional real-world training data for new categories.

In future works, the collected data can be further expanded in both size and diversity and can include more HOI categories to support broader applications and more comprehensive evaluations. We observe that there is a growing number of motion capture-based mesh datasets [49, 77, 138]. In parallel, recent advances in HOI mesh synthesis [24, 42, 65, 84, 94, 125, 134] enable generative models to more faithfully represent the dynamics of human-object interactions, paving the way for large-scale RF data generation to support HOI recognition tasks. Accordingly, *RF-HOI* has the potential to enlarge its synthetic data by utilizing more mesh datasets from either motion capture or generative models. Furthermore, in terms of the real-world data, we envision that *RF-HOI* will

inspire future research on RF-based HOI recognition using custom-built real-world test sets. Specifically, the decoupled representation of actions and objects enables systematic expansion of HOI categories by combining existing primitives with new object categories, reducing the need for exhaustive data collection. The proposed recognition model can serve as a baseline to quantify the difficulty of the dataset, while the insights from our data collection provide practical guidance for designing future RF-based HOI datasets. Although the data that a single work can contribute may be limited, we may integrate these efforts into a universe benchmark, including more HOI categories and setups.

7.2 Downstream Application

HOI recognition serves as a crucial component in various downstream applications, including VR, AR, and assistive robots. In VR [8], recognizing an HOI enables the system to generate appropriate sensory feedback—such as the sound of a closet opening—enhancing user immersion. In AR [48], the recognized HOI can trigger updates to the virtual environment, such as overlaying the text to write when the user picks up a pen. In general, *RF-HOI* can enhance the sense of realism in both VR and AR. Besides, as shown in our case study, HOI sequences can help robots to infer the user’s goal. Then assistive robots can provide help correspondingly to reach the goal. For instance, observing the sequence [*<sit down, chair>*, *<stand up, chair>*, *<pick up, book>*, *<put down, book>*] may lead the robot to infer that the user intends to check books, prompting it to deliver another book to the user for checking. These scenarios demonstrate the potential of integrating *RF-HOI* as a core module in interactive systems that require semantic understanding of HOI. The above applications are more readily deployable in environments where RFID infrastructure is already available or can be naturally integrated, such as smart homes, healthcare facilities, assisted living spaces, and retail environments with tagged objects. However, deploying *RF-HOI* in fully uninstrumented environments would require additional tagging effort and reader placement, which may limit its immediate applicability. Future work can reduce this deployment burden by exploring opportunistic RFID infrastructure, lower-cost tagging strategies, or complementary object-identification cues.

8 Conclusion

This paper proposes *RF-HOI*, a novel HOI recognition system that leverages only RF signals. By fusing mmWave and RFID data, *RF-HOI* harnesses the strengths of different modalities, making it particularly well-suited for HOI recognition applications. To enhance generalizability, *RF-HOI* uses a physics-based simulator to generate synthetic mmWave and RFID data, pre-trains the model on these synthetic data, and subsequently fine-tunes it with a small amount of real-world data. Extensive experiments demonstrate that *RF-HOI* consistently outperforms single-modality baselines, with synthetic data yielding significant performance improvements when real-world training samples or setups are limited, thus reducing data collection costs. We anticipate that *RF-HOI* will inspire further research in wireless sensing for HOI recognition.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. We also thank Zhehao Zhang and Jienan Chen for their help with data collection.

References

- [1] Aakriti Adhikari, Hem Regmi, Sanjib Sur, and Srihari Nelakuditi. 2022. Mishape: Accurate human silhouettes and body joints from commodity millimeter-wave devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3 (2022), 1–31.
- [2] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 2021. Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–10.
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* 1, 1 (2016), 1.

- [4] Kang Min Bae, Namjo Ahn, Yoon Chae, Parth Pathak, Sung-Min Sohn, and Song Min Kim. 2022. OmniScatter: extreme sensitivity mmWave backscattering using commodity FMCW radar. In *The 20th Annual International Conference on Mobile Systems, Applications and Services*. Association for Computing Machinery, New York, NY, USA, 316–329.
- [5] Kang Min Bae, Hankyeol Moon, and Song Min Kim. 2024. SuperSight: Sub-cm NLOS Localization for mmWave Backscatter. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services*. Association for Computing Machinery, New York, NY, USA, 278–291.
- [6] Bluesight. 2025. RFID Medication Management and Tracking System. <https://bluesight.com/rfid-medication-management/>.
- [7] Yanling Bu, Lei Xie, Yinyin Gong, Chuyu Wang, Lei Yang, Jia Liu, and Sanglu Lu. 2020. RF-dial: Rigid motion tracking and touch gesture detection for interaction via RFID tags. *IEEE Transactions on Mobile Computing* 21, 3 (2020), 1061–1080.
- [8] Ryan Canales and Sophie Jörg. 2020. Performance is not everything: Audio feedback preferred over visual feedback for grasping task in virtual reality. In *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games*. Association for Computing Machinery, New York, NY, USA, 1–6.
- [9] Qiming Cao, Hongfei Xue, Tianci Liu, Xingchen Wang, Haoyu Wang, Xincheng Zhang, and Lu Su. 2024. mmCLIP: Boosting mmWave-based Zero-shot HAR via Signal-Text Alignment. In *The 22nd ACM Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 184–197.
- [10] Chainway. 2026. MC51 5G Built-in RFID Reader. <https://www.chainway.net/Products/Info/172>.
- [11] Chainway. 2026. MR20 Wearable RFID Reader. <https://www.chainway.net/mobile/Products/Info/98>.
- [12] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *IEEE International Conference on Computer Vision*. IEEE, Piscataway, NJ, USA, 1017–1025.
- [13] Jian Chen, Lan Du, Guanbo Guo, Linwei Yin, and Di Wei. 2022. Target-attentional CNN for radar automatic target recognition with HRRP. *Signal Processing* 196 (2022), 108497.
- [14] Xingyu Chen and Xinyu Zhang. 2023. RF genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models. In *The 21st ACM Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 28–42.
- [15] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. 2019. Destruction and construction learning for fine-grained image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 5157–5166.
- [16] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. 2024. RF-diffusion: Radio signal generation via time-frequency diffusion. In *The 30th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 77–92.
- [17] Han Cui, Shu Zhong, Jiacheng Wu, Zichao Shen, Naim Dahnoun, and Yiren Zhao. 2023. Milipoint: A point cloud dataset for mmwave radar. *Advances in Neural Information Processing Systems* 36 (2023), 62713–62726.
- [18] Shenghong Dai, Shiqi Jiang, Yifan Yang, Ting Cao, Mo Li, Suman Banerjee, and Lili Qiu. 2025. Babel: A scalable pre-trained model for multi-modal sensing via expandable modality alignment. In *The 23rd ACM Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 240–253.
- [19] Decathlon. 2025. Product traceability and RFID technology at DECATHLON. <https://sustainability.decathlon.com/product-traceability-and-rfid-technology-at-decathlon>.
- [20] DeepSeek AI. 2025. DeepSeek 3.2 model page. <https://chat.deepseek.com/>.
- [21] Dell. 2025. Dell G15 5535 laptop. <https://www.dell.com/en-us/shop/dell-laptops/g15-gaming-laptop/spd/g-series-15-5535-laptop>.
- [22] Kaikai Deng, Dong Zhao, Qiaoyue Han, Zihan Zhang, Shuyue Wang, Anfu Zhou, and Huadong Ma. 2023. Midas: Generating mmWave radar data from videos for training pervasive and privacy-preserving human sensing tasks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1 (2023), 1–26.
- [23] Kaikai Deng, Dong Zhao, Zihan Zhang, Shuyue Wang, Wenxin Zheng, and Huadong Ma. 2023. Midas++: generating training data of mmWave radars from videos for privacy-preserving human sensing with mobility. *IEEE Transactions on Mobile Computing* 23, 6 (2023), 6650–6666.
- [24] Christian Diller and Angela Dai. 2024. Cg-hoi: Contact-guided 3d human-object interaction generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 19888–19901.
- [25] e tag. 2025. RFID In Smart Kitchens: Enhancing Efficiency and Reducing Waste. <https://e-tagrfid.com/rfid-in-smart-kitchens-enhancing-efficiency-reducing-waste>.
- [26] Junqiao Fan, Haocong Rao, Jiarui Zhang, Jianfei Yang, and Lihua Xie. 2026. MMPred: Radar-Based Human Motion Prediction in the Dark. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 1. AAAI Press, Palo Alto, California, USA, 1.
- [27] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. 2020. In-home daily-life captioning using radio signals. In *European Conference on Computer Vision*. Springer, Springer, Berlin, Germany, 105–123.
- [28] Chao Feng, Jiashen Chen, Shuo Liang, Xiaopeng Peng, Baizhou Yang, Xuan Wang, Zexuan Huang, Xianjia Meng, and Xiaojiang Chen. 2026. RFusion: Dynamic Multimodal RF Fusion for Few-Shot Human Activity Recognition. *IEEE Transactions on Mobile Computing* 1-1, 1-1 (2026), 1–1.

- [29] Yixu Feng, Shuo Hou, Haotian Lin, Yu Zhu, Peng Wu, Wei Dong, Jinxiu Sun, Qingsen Yan, and Yanning Zhang. 2024. Difflight: integrating content and detail for low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 6143–6152.
- [30] Chuhan Gao, Yilong Li, and Xinyu Zhang. 2019. Livetag: Sensing human-object interaction through passive chipless wi-fi tags. *GetMobile: Mobile Computing and Communications* 22, 3 (2019), 32–35.
- [31] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and recognizing human-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 8359–8367.
- [32] Google DeepMind. 2025. Gemini 2.5 Flash model page. <https://gemini.google.com/app>.
- [33] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 10 (2009), 1775–1789.
- [34] Guorong He, Shaojie Chen, Dan Xu, Xiaojiang Chen, Yaxiong Xie, Xinhui Wang, and Dingyi Fang. 2023. Fusang: Graph-inspired robust and accurate object recognition on commodity mmWave devices. In *The 21st Annual International Conference on Mobile Systems, Applications and Services*. Association for Computing Machinery, New York, NY, USA, 489–502.
- [35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 770–778.
- [36] H&M. 2025. H&M Group: Taking steps to grow the customer experience through tech. <https://hmgroup.com/our-stories/hm-group-taking-steps-to-grow-the-customer-experience-through-tech/>.
- [37] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [38] Markus Höll, Markus Oberweger, Clemens Arth, and Vincent Lepetit. 2018. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces*. IEEE, Piscataway, NJ, USA, 175–182.
- [39] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spaCy: Industrial-strength Natural Language Processing in Python. <https://spacy.io/>.
- [40] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation networks for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 3588–3597.
- [41] Anna Huang, Dong Wang, Run Zhao, and Qian Zhang. 2019. Au-id: Automatic user identification and authentication through the motions captured from sequential human activities using rfid. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2 (2019), 1–26.
- [42] Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, et al. 2025. HOIGPT: Learning Long-Sequence Hand-Object Interaction with Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 7136–7146.
- [43] Impinj. 2025. Impinj R420 reader. <https://www.impinj.com/products/readers/impinj-speedway>.
- [44] Naveed Imran, Jian Zhang, Jehad Ali, Sana Hameed, Houbing Herbert Song, and Byeong-hee Roh. 2025. SwinLSTM-EmoRec: A Robust Dual-Modal Emotion Recognition Framework Combining mmWave Radar and Camera for IoT-Enabled Multimedia Applications. *IEEE Internet of Things Journal* 1 (2025), 1.
- [45] Texas Instruments. 2025. AWR1843AOPEVM. <https://www.ti.com/tool/AWR1843AOPEVM>.
- [46] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*. pmlr, PMLR, San Diego, CA, USA, 448–456.
- [47] Cesar Iovescu and Sandeep Rao. 2020. The fundamentals of millimeter wave radar sensors. 7 pages.
- [48] Rahul Jain, Jingyu Shi, Runlin Duan, Zhengzhe Zhu, Xun Qian, and Karthik Ramani. 2023. Ubi-TOUCH: Ubiquitous Tangible Object Utilization through Consistent Hand-object interaction in Augmented Reality. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 1–18.
- [49] Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. 2024. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*. Association for Computing Machinery, New York, NY, USA, 1–11.
- [50] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. 2023. Full-body articulated human-object interaction. In *IEEE/CVF International Conference on Computer Vision*. IEEE, Piscataway, NJ, USA, 9365–9376.
- [51] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. 2024. Scaling up dynamic human-scene interaction modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 1737–1747.
- [52] Haojian Jin, Zhijian Yang, Swarun Kumar, and Jason I Hong. 2018. Towards wearable everyday body-frame tracking using passive RFIDs. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4 (2018), 1–23.
- [53] Vinay Joshi, Shengkai Xu, Qiming Cao, Yi Zhu, Pu Wang, and Hongfei Xue. 2024. Towards Robust mmWave-based Human Activity Recognition using Large Simulated Dataset for Model Pretraining. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, Piscataway, NJ, USA, 7332–7337.

- [54] Belal Korany, Chitra R Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for through-wall person identification from candidate video footage. In *The 25th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 1–15.
- [55] Bo Lan, Pei Li, Jiaxi Yin, Yunpeng Song, Ge Wang, Han Ding, Jinsong Han, and Fei Wang. 2025. XRF V2: A Dataset for Action Summarization with Wi-Fi Signals, and IMUs in Phones, Watches, Earbuds, and Glasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 98 (Sept. 2025), 41 pages. doi:10.1145/3749521
- [56] Gierad Laput and Chris Harrison. 2019. Sensing fine-grained hand activity with smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.
- [57] Chi-Jung Lee, Jiaxin Li, Tianhong Catherine Yu, Ruidong Zhang, Vipin Gunda, François Guimbretière, and Cheng Zhang. 2025. Grab-n-Go: On-the-Go Microgesture Recognition with Objects in Hand. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3 (2025), 1–27.
- [58] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, et al. 2024. Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–21.
- [59] Chenglong Li, Emmeric Tanghe, David Plets, Pieter Suanet, Jeroen Hoebeke, Eli De Poorter, and Wout Joseph. 2020. ReLoc: Hybrid RSSI- and phase-based relative UHF-RFID tag localization with COTS devices. *IEEE Transactions on Instrumentation and Measurement* 69, 10 (2020), 8613–8627.
- [60] Hanchuan Li, Can Ye, and Alanson P Sample. 2015. IDSense: A human object interaction detection system based on passive UHF RFID. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2555–2564.
- [61] Jiamu Li, Dongheng Zhang, Zhi Wu, Cong Yu, Yadong Li, Qi Chen, Yang Hu, Qibin Sun, and Yan Chen. 2024. SBRF: A fine-grained radar signal generator for human sensing. *IEEE Transactions on Mobile Computing* 1-1, 1-1 (2024), 1–1.
- [62] Rong Li, Tao Deng, Siwei Feng, Mingjie Sun, and Juncheng Jia. 2025. ConSense: Continually Sensing Human Activity with WiFi via Growing and Picking. In *The AAAI Conference on Artificial intelligence*, Vol. 39. AAAI Press, Palo Alto, California, USA, 14292–14300.
- [63] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall S Burd. 2016. Deep learning for rfid-based activity recognition. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 164–175.
- [64] Zhengxiong Li, Baicheng Chen, Zhuolin Yang, Huining Li, Chenhan Xu, Xingyu Chen, Kun Wang, and Wen Yao Xu. 2019. Ferrotag: A paper-based mmwave-scannable tagging infrastructure. In *The 17th Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 324–337.
- [65] Zhe Li, Jiakun Li, Mingqi Gao, Jinyu Yang, Wei Wang, and Feng Zheng. 2026. ContA-HOI: Towards Physically Plausible Human-Object Interaction Generation via Contact-Aware Modeling. In *Thirteenth International Conference on 3D Vision*. IEEE, Piscataway, NJ, USA, 1–1.
- [66] Zhuolong Li, Xingao Li, Changxing Ding, and Xiangmin Xu. 2024. Disentangled pre-training for human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 28191–28201.
- [67] Bo Liang, Purui Wang, Renjie Zhao, Heyu Guo, Pengyu Zhang, Junchen Guo, Shunmin Zhu, Hongqiang Harry Liu, Xinyu Zhang, and Chenren Xu. 2023. {RF-Chord}: Towards deployable {RFID} localization system for logistic networks. In *20th USENIX Symposium on Networked Systems Design and Implementation*. IEEE, Piscataway, NJ, USA, 1783–1799.
- [68] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. 2020. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 482–490.
- [69] Mingzhi Lin, Teng Huang, Han Ding, Cui Zhao, Fei Wang, Ge Wang, and Wei Xi. 2026. Active Domain Adaptation for Mmwave-Based HAR Via Rényi Entropy-Based Uncertainty Estimation. *IEEE Transactions on Mobile Computing* 1-1, 1-1 (2026), 1–1.
- [70] Haipeng Liu, Kening Cui, Kaiyuan Hu, Yuheng Wang, Anfu Zhou, Liang Liu, and Huadong Ma. 2022. mTransSee: Enabling environment-independent mmWave sensing based gesture recognition via transfer learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1 (2022), 1–28.
- [71] Ting Liu, Yao Zhao, Yunchao Wei, Yufeng Zhao, and Shikui Wei. 2019. Concealed object detection for activate millimeter wave image. *IEEE Transactions on Industrial Electronics* 66, 12 (2019), 9909–9917.
- [72] Xiulong Liu, Dongdong Liu, Jiuwu Zhang, Tao Gu, and Keqiu Li. 2021. RFID and camera fusion for recognition of human-object interactions. In *The 27th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 296–308.
- [73] Logitech. 2025. Logitech Spot: Presence and Environmental Sensor. <https://www.logitech.com/en-us/products/video-conferencing/accessories/spot-sensor.950-000107.html>.
- [74] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* 1-1, 1-1 (2017), 1–1.

- [75] Haofan Lu, Mohammad Mazaheri, Reza Rezvani, and Omid Abari. 2023. A millimeter wave backscatter network for two-way communication and localization. In *Proceedings of the ACM SIGCOMM 2023 Conference*. Association for Computing Machinery, New York, NY, USA, 49–61.
- [76] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019), 1.
- [77] Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. 2025. HUMOTO: A 4D Dataset of Mocap Human Object Interactions. *arXiv preprint arXiv:2504.10414* 1-1, 1-1 (2025), 1–1.
- [78] Tianlun Luo, Steven Guan, Rui Yang, and Jeremy Smith. 2023. From detection to understanding: A survey on representation learning for human-object interaction. *Neurocomputing* 543 (2023), 126243.
- [79] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, Vol. 30. Atlanta, GA, PMLR, San Diego, CA, USA, 3.
- [80] Valerio Magnago, Luigi Palopoli, Alice Buffi, Bernardo Tellini, Andrea Motroni, Paolo Nepa, David Macii, and Daniele Fontanelli. 2019. Ranging-free UHF-RFID robot positioning through phase measurements of passive tags. *IEEE Transactions on Instrumentation and Measurement* 69, 5 (2019), 2408–2418.
- [81] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. 2023. HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Assistive roBOTS. In *7th Annual Conference on Robot Learning*, Vol. 1-1. PMLR, San Diego, CA, USA, 1–1.
- [82] Fairview Microwave. 2025. FM51FP1006 high performance antenna. <https://www.fairviewmicrowave.com/product/antennas/directional-antennas/panel-antennas/wide-antenna-9-dbi-gain-n-fm51fp1006.html>.
- [83] Mahathir Monjur and Shahriar Nirjon. 2026. mmWEAVER: Environment-Specific mmWave Signal Synthesis from a Photo and Activity Description. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, Piscataway, NJ, USA, 1875–1884.
- [84] Lingzhou Mu, Wang Qiang, Fan Jiang, Mengchao Wang, Mu Xu, and Kai Zhang. 2026. FantasyHSI: Video-Generation-Centric 4D Human Synthesis in Any Scene Through a Graph-Based Multi-Agent Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. AAAI Press, Palo Alto, California, USA, 8116–8124.
- [85] RMS Omega. 2026. Track Critical Healthcare Supplies with RFID Technology. <https://rmsomega.com/healthcare/track/rfid-in-healthcare/>.
- [86] Omnicell. 2025. Medical RFID Cabinet. <https://www.omnicell.me/medical-supply-automation/rfid-cabinet/>.
- [87] OpenAI. 2025. ChatGPT (GPT-5) model page. <https://chat.openai.com/>.
- [88] ORBBEC. 2025. Femto Bolt. <https://www.orbbec.com/products/tof-camera/femto-bolt/>.
- [89] ORBBEC. 2025. Python Bindings for Orbbec SDK. <https://github.com/orbbec/pyorbbecsdk>.
- [90] OSSIFYGAR. 2025. UHF RFID tag 9662 Long Distance Passive Alien H3 Adhesive Inlay Markable Lable. <https://www.amazon.com/dp/B0B1H7Z6KZ>.
- [91] A Paszke. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* 1, 1 (2019), 1.
- [92] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 10975–10985.
- [93] Mingtao Pei, Yunde Jia, and Song-Chun Zhu. 2011. Parsing video events with goal inference and intent prediction. In *International Conference on Computer Vision*. IEEE, IEEE, Piscataway, NJ, USA, 487–494.
- [94] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2025. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 2878–2888.
- [95] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32. AAAI Press, Palo Alto, California, USA, 1.
- [96] Swadhin Pradhan, Eugene Chai, Karthikeyan Sundaresan, Lili Qiu, Mohammad A Khojastepour, and Sampath Rangarajan. 2017. RIO: A pervasive RFID-based touch gesture interface. In *The 23rd Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 261–274.
- [97] Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. 2023. Nopa: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. In *IEEE International Conference on Robotics and Automation*. IEEE, IEEE, Piscataway, NJ, USA, 7628–7634.
- [98] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. 2018. Learning human-object interactions by graph parsing neural networks. In *European conference on computer vision*. Springer, Berlin, Germany, 401–417.
- [99] RFIDlabel. 2025. Upgrade Your Kitchen with Smart Microwaves and RFID Tagging in The Future! <https://www.rfidlabel.com/upgrade-your-kitchen-source-smart-microwaves-with-rfid-tagging-today/>.
- [100] Mark A Richards et al. 2005. *Fundamentals of radar signal processing*. Vol. 1. McGraw-hill New York, New York, NY, USA.
- [101] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *IEEE/CVF International Conference on Computer Vision*.

- IEEE, Piscataway, NJ, USA, 9339–9347.
- [102] Ralph Schmidt. 1986. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation* 34, 3 (1986), 276–280.
 - [103] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*. Association for Computing Machinery, New York, NY, USA, 51–56.
 - [104] sllurp. 2025. A Python library to interface with RFID readers. <https://github.com/sllurp/sllurp>.
 - [105] Elahe Soltanaghahi, Akarsh Prabhakara, Artur Balanuta, Matthew Anderson, Jan M Rabaey, Swarun Kumar, and Anthony Rowe. 2021. Millimetro: mmWave retro-reflective tags for accurate, long range localization. In *The 27th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 69–82.
 - [106] Andrew Spielberg, Alanson Sample, Scott E Hudson, Jennifer Mankoff, and James McCann. 2016. RapID: A framework for fabricating low-latency interactive objects with RFID tags. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 5897–5908.
 - [107] 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. doi:10.1145/3643553
 - [108] Xiaoqi Sun, Yanwen Wang, Chenwei Zhang, Zheng Wang, Xiaokang Shi, and Yuanqing Zheng. 2024. LoDiHAR: A Low-Cost Distributed Human Activity Recognition System Based on RFID. In *The 21st Annual IEEE International Conference on Sensing, Communication, and Networking*. IEEE, IEEE, Piscataway, NJ, USA, 1–9.
 - [109] L Talbi. 2001. Simulation of indoor UHF propagation using numerical technique. In *Canadian Conference on Electrical and Computer Engineering 2001. Conference Proceedings (Cat. No. 01TH8555)*, Vol. 2. IEEE, IEEE, Piscataway, NJ, USA, 1357–1362.
 - [110] UROVO. 2026. Enhancing Retail with Handheld RFID Readers. <https://en.urovo.com/blog/RFID-handheld-readers/enhancing-retail-with-handheld-rfid-readers.html>.
 - [111] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 1–1.
 - [112] Anandghan Waghmare, Youssef Ben Taleb, Ishan Chatterjee, Arjun Narendra, and Shwetak Patel. 2023. Z-ring: Single-point bio-impedance sensing for gesture, touch, object and user recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.
 - [113] Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. 2024. Xrf55: A radio frequency dataset for human indoor action analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1 (2024), 1–34.
 - [114] Hengbo Wang, Lvqing Yang, Yishu Qiu, Zhipeng Liu, Kai Li, Bo Yu, and Shihui Guo. 2026. RFGAT: Generative Adversarial Teacher for Cross-Domain RFID Activity Recognition. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, Piscataway, NJ, USA, 5811–5815.
 - [115] Liang Wang, Tao Gu, Xianping Tao, and Jian Lu. 2016. Toward a wearable RFID system for real-time activity recognition using radio patterns. *IEEE Transactions on Mobile Computing* 16, 1 (2016), 228–242.
 - [116] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *The 21st Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 65–76.
 - [117] Yuheng Wang, Haipeng Liu, Kening Cui, Anfu Zhou, Wensheng Li, and Huadong Ma. 2021. m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Association for Computing Machinery, New York, NY, USA, 8298–8302.
 - [118] Yida Wang, Yu Lu, Yuxuan Zhou, Yifei Shen, Lili Qiu, Zeyuan Lai, Yi-Chao Chen, Hao Pan, Juntao Zhou, Dian Ding, et al. 2025. High-resolution mmWave Imaging using Metasurface and Diffusion. In *The 23rd Annual International Conference on Mobile Systems, Applications, and Services*. Association for Computing Machinery, New York, NY, USA.
 - [119] Yanwen Wang and Yuanqing Zheng. 2018. Modeling RFID signal reflection for contact-free activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4 (2018), 1–22.
 - [120] Ziqi Wang and Shiwen Mao. 2025. Generative AI-Empowered RFID Sensing for 3D Human Pose Augmentation and Completion. *IEEE Open Journal of the Communications Society* 1-1, 1-1 (2025), 1–1.
 - [121] Teng Wei and Xinyu Zhang. 2016. Gyro in the air: tracking 3d orientation of batteryless internet-of-things. In *The 22nd Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 55–68.
 - [122] Jiahao Wu, Yishu Qiu, Lvqing Yang, Shaoqin Shen, Kai Li, Bo Yu, and Shihui Guo. 2026. RF-OnlineHAR: A Real-Time RFID Activity Recognition Framework with Memory-Augmented Transformer. *IEEE Internet of Things Journal* 1-1, 1-1 (2026), 1–1.
 - [123] Xiaomi. 2025. Xiaomi Mijia Human Presence Sensor Detector for Smart Wireless Home. <https://www.aliexpress.us/item/3256807256719195.html>.

- [124] Xin Xie, Xiulong Liu, Xibin Zhao, Weilian Xue, Bin Xiao, Heng Qi, Keqiu Li, and Jie Wu. 2019. Implementation of differential tag sampling for COTS RFID systems. *IEEE Transactions on Mobile Computing* 19, 8 (2019), 1848–1861.
- [125] Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. 2025. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 7048–7060.
- [126] Yongchao Xu, Jiawei Liu, Sen Tao, Qiang Zhang, and Zheng-Jun Zha. 2025. HOIMamba: Efficient Mamba-based Disentangled Progressive Learning for HOI Detection. In *The AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI Press, Palo Alto, California, USA, 8987–8995.
- [127] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. 2023. Towards generalized mmwave-based human pose estimation through signal augmentation. In *The 29th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 1–15.
- [128] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *The 19th Annual International Conference on Mobile Systems, Applications, and Services*. Association for Computing Machinery, New York, NY, USA, 269–282.
- [129] Yifan Yan, Shuai Yang, Xiuzhen Guo, Xiangguang Wang, Wei Chow, Yuanchao Shu, and Shibo He. 2025. mmExpert: Integrating Large Language Models for Comprehensive mmWave Data Synthesis and Understanding. In *Proceedings of the Twenty-sixth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*. Association for Computing Machinery, New York, NY, USA, 1–10.
- [130] Chao Yang, Lingxiao Wang, Xuyu Wang, and Shiwen Mao. 2022. Environment adaptive RFID-based 3D human pose tracking with a meta-learning approach. *IEEE Journal of Radio Frequency Identification* 6 (2022), 413–425.
- [131] Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. 2024. Open-world human-object interaction detection via multi-modal prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 16954–16964.
- [132] Zheng Yang, Yi Zhang, Kun Qian, and Chenshu Wu. 2023. {SLNet}: A Spectrogram Learning Neural Network for Deep Wireless Sensing. In *20th USENIX Symposium on Networked Systems Design and Implementation*. USENIX Association, Berkeley, CA, USA, 1221–1236.
- [133] Bangpeng Yao and Li Fei-Fei. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 17–24.
- [134] Wei Yao, Yunlian Sun, Hongwen Zhang, Yebin Liu, and Jinhui Tang. 2026. Hosig: Full-body human-object-scene interaction generation with hierarchical scene perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. AAAI Press, Palo Alto, California, USA, 11901–11909.
- [135] Lance Ying, Xinyi Li, Shivam Aarya, Yizirui Fang, Yifan Yin, Jason Xinyu Liu, Stefanie Tellex, Joshua B. Tenenbaum, and Tianmin Shu. 2025. Pragmatic Embodied Spoken Instruction Following in Human-Robot Collaboration with Theory of Mind. *arXiv preprint arXiv:2409.10849* 1-1, 1-1 (2025), 1–1.
- [136] Bo Zhang, Zimeng Zhou, Boyu Jiang, and Rong Zheng. 2024. SUPER: Seated Upper Body Pose Estimation using mmWave Radars. In *IEEE/ACM 9th International Conference on Internet-of-Things Design and Implementation*. IEEE, Association for Computing Machinery, New York, NY, USA, 181–191.
- [137] Feng Zhang, Chenshu Wu, Beibei Wang, and KJ Ray Liu. 2020. mmEye: Super-resolution millimeter wave imaging. *IEEE Internet of Things Journal* 8, 8 (2020), 6995–7008.
- [138] Juzhe Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. 2024. HOI-M³: Capture Multiple Humans and Objects Interaction within Contextual Environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 516–526.
- [139] Qian Zhang, Dong Wang, Run Zhao, Yufeng Deng, and Yinggang Yu. 2019. Shopeye: fusing rfid and smartwatch for multi-relation excavation in physical stores. In *The 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 86–95.
- [140] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2021. Widar3. 0: Zero-effort cross-domain gesture recognition with Wi-Fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8671–8688.
- [141] Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. 2025. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. *arXiv preprint arXiv:2502.15676* 1-1, 1-1 (2025), 1–1.
- [142] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, Niki Trigoni, and Andrew Markham. 2023. Cubelearn: End-to-end learning for human motion recognition from raw mmwave radar signals. *IEEE Internet of Things Journal* 10, 12 (2023), 10236–10249.
- [143] Ruoyu Zhao, Yushu Zhang, Tao Wang, Wenying Wen, Yong Xiang, and Xiaochun Cao. 2025. Visual content privacy protection: A survey. *Comput. Surveys* 57, 5 (2025), 1–36.
- [144] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. 2022. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Association for Computing Machinery, New York, NY, USA, 19568–19577.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009