HiSync: Spatio-Temporally Aligning Hand Motion from Wearable IMU and On-Robot Camera for Command Source Identification in Long-Range HRI

Chengwen Zhang
Department of Computer Science and Technology, BNRist
Tsinghua University
Beijing, China
zcw25@mails.tsinghua.edu.cn

Chun Yu*
Department of Computer Science and Technology, BNRist, College of AI
Tsinghua University
Beijing, China
chunyu@tsinghua.edu.cn

Borong Zhuang
Department of Computer Science and Technology
Tsinghua University
Beijing, China
brucezhuang@bupt.edu.cn

Haopeng Jin
Beijing University of Posts and Telecommunications
Beijing, China
meiqin1shengtui@126.com

Qingyang Wan
Academy of Arts & Design
Tsinghua University
Beijing, China
wanqy23@mails.tsinghua.edu.cn

Zhuojun Li
Department of Computer Science and Technology
Tsinghua University
Beijing, China
lizj23@mails.tsinghua.edu.cn

Zhe He
Department of Computer Science and Technology
Tsinghua University
Beijing, Beijing, China
hz23@mails.tsinghua.edu.cn

Zhoutong Ye
Department of Computer Science and Technology
Tsinghua University
Beijing, China
yezt24@mails.tsinghua.edu.cn

Yu Mei
Department of Computer Science and Technology
Tsinghua University
Beijing, China
meiy24@mails.tsinghua.edu.cn

Chang Liu
Department of Computer Science and Technology
Tsinghua University
Beijing, China
c-liu21@tsinghua.org.cn

Weinan Shi*†
Department of Computer Science and Technology
Tsinghua University
Beijing, China
swn@tsinghua.edu.cn

Yuanchun Shi*
Department of Computer Science and Technology, BNRist
Tsinghua University
Beijing, China
Qinghai University
Xining, China
shiyc@tsinghua.edu.cn

Abstract

Long-range Human-Robot Interaction (HRI) remains underexplored. Within it, Command Source Identification (CSI) — determining who issued a command — is especially challenging due to multi-user and distance-induced sensor ambiguity. We introduce HiSync, an optical-inertial fusion framework that treats hand motion as binding cues by aligning robot-mounted camera optical flow with hand-worn IMU signals. We first elicit a user-defined (N=12) gesture set and collect a multimodal command gesture dataset (N=38) in long-range multi-user HRI scenarios. Next, HiSync extracts frequency-domain hand motion features from both camera and IMU data, and a learned CSINet denoises IMU readings, temporally aligns modalities, and performs distance-aware multi-window fusion to compute cross-modal similarity of subtle, natural gestures, enabling robust CSI. In three-person scenes up to 34 m, HiSync achieves 92.32% CSI accuracy, outperforming the prior SOTA by 48.44%. HiSync is also validated on real-robot deployment. By making CSI reliable and natural, HiSync provides a practical primitive and design guidance for public-space HRI.

CCS Concepts

• **Human-centered computing** → **Gestural input**.

Keywords

Human Robot Interaction, Optical-Inertial Fusion, Gesture

*Also with Key Laboratory of Pervasive Computing, Ministry of Education.
†Corresponding author.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790345

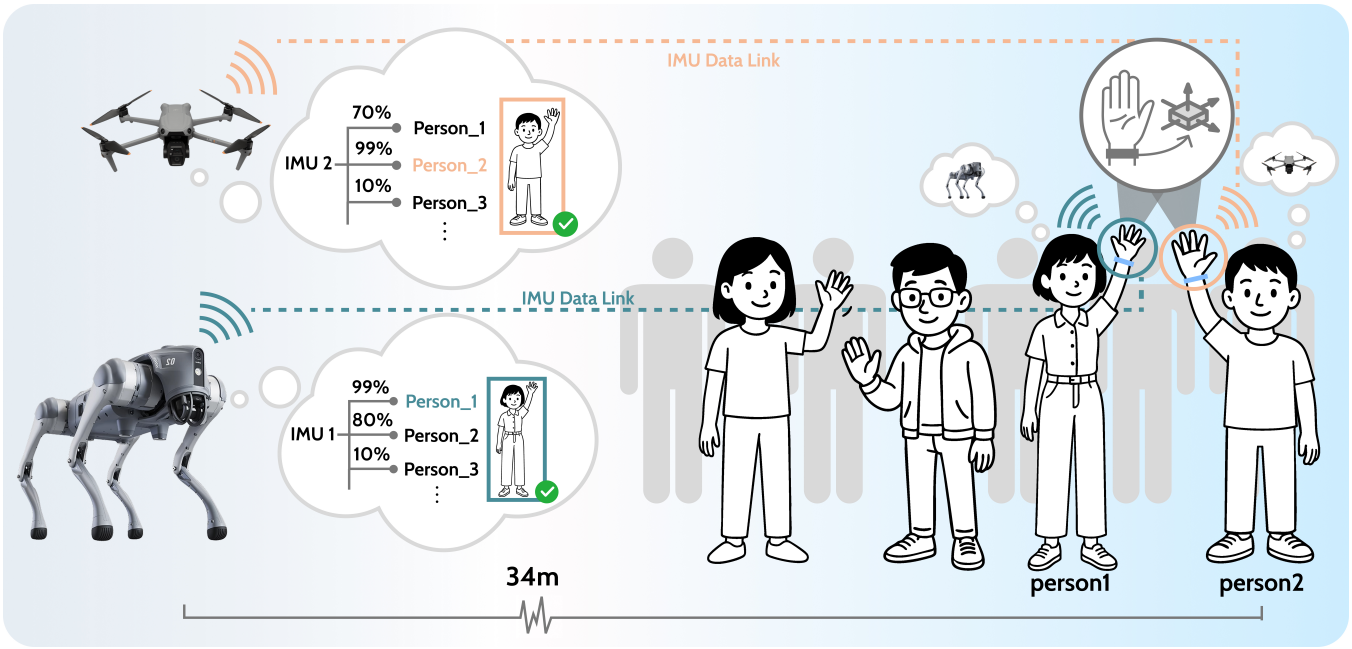


Figure 1: Demonstration of HiSync. Multiple people perform similar gestures at a distance of 34 m. The figure illustrates the application scenario: Person 1 (blue link) controls the quadruped, while Person 2 (orange link) simultaneously controls the drone. Other bystanders act as visual distractors. Each robot receives an inertial stream from its paired device. HiSync enables the robot to identify its bound command issuer, effectively rejecting distractors regardless of their kinematic similarity.

ACM Reference Format:
Chengwen Zhang, Chun Yu, Borong Zhuang, Haopeng Jin, Qingyang Wan, Zhuojun Li, Zhe He, Zhoutong Ye, Yu Mei, Chang Liu, Weinan Shi, and Yuanchun Shi. 2026. HiSync: Spatio-Temporally Aligning Hand Motion from Wearable IMU and On-Robot Camera for Command Source Identification in Long-Range HRI. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3772318.3790345

1 INTRODUCTION

Human-robot interaction (HRI) is a vibrant and fast-growing area within HCI [30, 72]. However, the vast majority of HRI research [31, 36, 40, 48, 91] focuses on near-range or single-person settings — within 10 m where the operator is visually salient and easy to disambiguate. At the same time, advances in robotics and real-world deployments are driving interactions into larger public spaces and more diverse embodiments [51, 76]. Many emerging use cases require interaction at a distance. For example, summoning an autonomous car across a parking lot, directing a drone outdoors, or calling a service robot in a large venue [5, 13, 89]. These settings introduce long distances ($\geq$ 10 m), dynamic multi-user crowds, and cluttered, uncertain environments, demanding interfaces that remain reliable from afar.

A foundational perceptual capability in such long-range scenarios is **Command Source Identification (CSI)** — determining who in the scene issued the command. In near-range HRI, CSI is often trivial or implicitly handled because the operator is close, prominent, and rarely confused with bystanders [6, 58, 86]. In long-range contexts, however, distance-induced visual ambiguity and simultaneous gestures from multiple users make CSI difficult. This gap motivates our work: we extend HRI beyond near-range by focusing on robust CSI as a prerequisite for reliable long-range interaction. To scope this work, we consider scenarios in which a single user has already established a one-to-one connection with the robot, while other users only act as visual distractors.

We identify three key challenges introduced by long-range CSI, which were not covered by existing literature. **(C1) Unnatural Interaction**. Existing methods [55, 66, 75] require users to perform substantial movements or set up additional devices, which can be unnatural and burdensome. **(C2) Visual ambiguities**. At long ranges, hands and subtle wrist motions collapse to only a few pixels on typical robot-mounted cameras; appearance cues such as texture, skin tone, and fingers are degraded, and hand estimators become unreliable (Fig. 2). **(C3) Data noise**. At long ranges, both optical flow and IMU signals are plagued by noise. Visual flow is distorted by background clutter and motion blur [3], while IMU readings are affected by inherent noise and grip jitter [67]. Moreover, sensor lags and sample-rate mismatch further complicate reliable CSI [21].

We introduce HiSync, which tackles the challenges of command source identification in long-range human-robot interaction. Firstly, to address **C1**, we use common low-cost on-robot camera and IMU as CSI sensors. Moreover, we focus on hand gestures as the interaction modality, as they are intuitive and commonly used in daily life. To this end, we conduct a formative study to elicit a user-defined gesture set from 12 participants.

Secondly, to solve **C2**, HiSync is designed as an optical-inertial fusion framework that treats hand motion as a binding cue: it aligns robot-mounted camera optical flow with hand-worn IMU traces to disambiguate the command issuer. By treating RGB as optical flow and matching candidates on motion cues, HiSync remains reliable when visual details are blurred.

Thirdly, to handle **C3**, HiSync maps optical flow and IMU to the frequency domain, leveraging a novel CSINet that incorporates the following mechanisms. Quality-Aware Feature Modulation reweights IMU spectral features by quality cues to suppress frames with low signal-to-noise ratio (SNR). IMU-Anchored Cross-Modal Attention aligns flow-spectral sequences to the IMU time reference, mitigating time jitter, missed detections, and phase drift. Scale-Aware Multi-Window Fusion aggregates predictions across multiple window lengths conditioned on distance, stabilizing decisions across ranges. Together, these modules yield a more robust cross-modal similarity in noisy long-range settings.

Furthermore, to address the lack of public datasets for CSI in long-range multi-user settings, we curate, to our knowledge, the first large-scale multimodal dataset of long-range gestural interactions. In this dataset, each frame has at least 2 people gesturing at the on-robot camera. Only one person was equipped with an IMU, while bystanders acted as visual-only distractors. We acknowledge that this setup simplifies the CSI problem compared to scenarios where multiple users simultaneously transmit IMU data to the same robot. However, evaluations on this setting still show that HiSync demonstrates the feasibility of optical-inertial binding for CSI and exhibits robustness to visual modality interference.

On this dataset, HiSync achieved 97.82% overall accuracy from 3–34 m, exceeding the previous SOTA by 27.3% points. Even at 34 m, it maintained 94.31% accuracy, while the previous SOTA achieved only 43.88%. These results demonstrate HiSync's applicability to open, multi-user environments. The dataset also establishes a reproducible benchmark for this underexplored problem.

In summary, our main contributions are as follows:

- We present a systematic user-defined natural gesture set for multiple robot types at long range, showing how gesture properties scale with distance.
- We introduce HiSync, an optical-inertial fusion architecture that leverages spectral motion cues extracted from low-cost and widely deployed sensors (on-robot camera and wearable IMU) to achieve robust CSI at long range.
- We evaluate HiSync and release a large, annotated multi-modal dataset[1] spanning 3-34 m and multiple gesture categories, collected from 38 participants, to enable reproducible research evaluation on long-range HRI.
- We deploy HiSync on a quadruped robot to validate its real-world feasibility. Field experiments confirm robust performance. User studies demonstrate superior usability and low burden compared to standard interfaces.

Together, these contributions establish gesture-based command source identification as a distinct research challenge in long-range HRI, offering both empirical insights and technical foundations that open new avenues for designing robust, scalable human-robot interaction systems.

[1]https://github.com/OctopusWen/HiSync

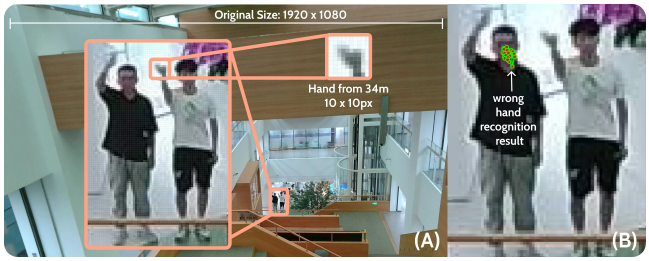


Figure 2: Visual Ambiguity at a Distance of 34 m. Figure (A) shows a real sample from our dataset (1920 × 1080 resolution). The inset highlights that the hand region occupies fewer than 10 × 10 pixels. Even with a manually zoomed-in view of the detection result like (B), YOLOv11x [34] fails to identify the hand. This demonstrates the inherent visual ambiguity in long-range interactions.

2 RELATED WORK

2.1 Long-range Human-Robot Interaction

Long-range human-robot interaction (HRI) has explored multiple input channels, including microphones for speech or prosodic cues [11, 12, 71, 85, 88], screen-based or on-robot displays for directing and confirming actions [14], hand-held or wearable controllers for precise teleoperation at a distance [17, 45, 68, 69, 77], and free-hand gestures detected by onboard or remote cameras [5, 9, 19, 50]. Among these, hand gesture interaction is often favored for its low barrier and social legibility across users and settings [2, 12, 30, 42].

Yet most gesture-based HRI systems are designed for close-range use, typically within arm's reach to a few meters, where tracking and disambiguation are simpler and environments are more controlled [17, 20, 42]. Only a small subset explicitly addresses far-field perception, and these works largely prioritize recognition robustness over interaction breadth [5, 9, 50]. Despite these technical explorations, studies situated in public spaces or crowd settings remain sparse, and few examine how people would naturally gesture to summon and direct robots from tens of meters away without prior training [12, 30, 42]. This gap motivates our focus on free-hand commands for long-range, multi-user scenarios: we first conduct a formative study to elicit a user-defined gesture set and characterize distance-mediated communication, which then informs our system design and evaluation.

2.2 Gesture-based Command Source Identification

The core goal of HiSync is *command source identification (CSI)* for long-range, multi-user, gesture-based HRI — given several people and a detected command, determine who issued it [12, 30, 42, 78]. To the best of our knowledge, far-field hand-gesture CSI has not been explicitly defined and evaluated as a standalone task in prior HRI literature, which motivates our formulation of CSI as distinct from person localization, person search, and ReID [15, 43, 44, 62, 79, 90]. Closely related pipelines in vision emphasize detection-association-metric learning for persistent identity, whereas CSI binds an ephemeral command to its current actor without long-term

Table 1: Representative works related to long-range gesture CSI.

Work	Sensors	Data Modalities	Supported Distance
PTGAN[79]	Camera	RGB images	5–20 m
OSNet[90]	Camera	RGB images	5–20 m
CASE-Net[43]	Camera	Video	5–20 m
ABD-Net[15]	Camera	RGB images	5–20 m
AprilTag[55]	Camera + Fiducial Tag	Visual Markers	3–10 m
DeepArUco++[8]	Camera + Fiducial Tag	Learned Fiducial Detection	3–10 m
UWB[82]	UWB Tags	RF Ranging	30 m *Need Base Stations
GPS[52]	GPS Receiver	Global Location	*Error Range:7–13 m
Person-in-WiFi[75]	MIMO WiFi Tx&Rx	RF	3–6 m
HuPR[41]	mmWave Radar	mmWave	3–6 m
HPERL[22]	Expensive LiDAR	Laser	5–20 m
Simultaneous[28]	Camera + IMU	Video + Inertial Orientation	5-10 m
VIPL[66]	Camera + IMU	Optical Flow + IMU	<10 m *Need Walking
HiSync	Camera + IMU	Optical Flow + IMU on Spectral	3–34 m

identity enrollment [15, 43, 79]. For completeness, we summarize input modalities that inform multi-user gesture settings in Table 1.

Existing routes for CSI fall into three families — vision-only deep learning [15], device-assisted localization [39, 82], and visual-inertial (VI) fusion [28, 66]. **Vision-only** ReID/person-search match learned appearance features but typically require a gallery and mid-range conditions; at $\geq$ 30 m, sparse resolution, occlusion, and confounds make identity brittle without priors [15, 25, 32, 38, 43, 44, 79, 90]. **Device-assisted localization** leverages "privileged data" from user to bind identity, yet struggles in long-range ad-hoc scenarios. High-precision solutions require expensive hardware (e.g., LiDAR [22]) and pre-deploying infrastructure in the scene (e.g., AprilTag, ArUco and UWB [8, 35, 39, 55, 82, 83]). Ubiquitous signals like WiFi [75], GPS [52] and mmWave [41] lack the fine-grained spatial resolution required to disambiguate multiple users in common public HRI scenarios, where individuals often stand closely at a distance. **VI** association links on-body IMU with visual motion (e.g., heading matching and contrastive phone-IMU alignment), but still degrades when people occupy few pixels or gestures are brief [28, 66]. Building on this line, HiSync binds the command via hand-motion spectra, aligning robot-mounted cameras with low-cost and widely deployed IMU within the command window to avoid identity enrollment while preserving uninstrumented, free-hand interaction [12, 30, 42, 57].

2.3 Vision-Inertial Fusion Approaches

Recent vision-inertial (VI) work, like VIPL [66], learns shared embeddings that align optical flow with IMU signals via contrastive training. Nevertheless, its reliance on walking scenarios with 5-second full-body motion, assumption of tight video-IMU synchronization, and short-range limitation (<10 m) collectively restrict its deployment in real-world HRI settings. And inertial signals also suffer long-term noise and bias [33, 74]; IMU-only estimates are especially sensitive [28]. Without quality-aware gating, such errors propagate through fusion and reduce robustness.

Beyond synchronization, alignment is intrinsically hard due to distance: optical flow and IMU differ in units, sampling rates, and noise profiles, so simple shared embeddings underperform without additional constraints or priors [16, 46, 53, 61]. Temporal window length strongly shapes VI performance [54, 66], yet most systems fix a single window and under-discuss its impact across users and motions. These gaps motivate fusion that is synchronization-tolerant, quality-aware, and adaptive in the temporal context.

3 FORMATIVE STUDY

Existing user-defined gesture sets for HRI rarely account for long-range, open, multi-user settings, limiting their usefulness for Command Source Identification (CSI) [30]. To inform algorithm and system design under these conditions, we conducted a formative study to **(1)** elicit CSI-capable, user-preferred gestures for long-range HRI and **(2)** examine how these gestures change with distance to guide the design and optimization of long-range CSI systems.

To elicit natural gestures, we employed three real robotic platforms as interaction targets: a quadruped robot, a human-height wheeled robot, and a drone. Participants were instructed to issue gesture-based commands to these robots across varying distances and in multi-person settings.

3.1 Participants

We recruited 12 participants (8 male, 3 female, 1 preferred not to disclose), aged 20-31 years (M_{age} = 24.4, SD = 3.9). They represented diverse national, cultural, and educational backgrounds. Five participants had non-STEM backgrounds and reported little to no prior knowledge of robotics or robotic perception, while the other 7 participants' STEM backgrounds ranged from unrelated fields to expert in robotics. Two participants were left-handed. All participants received compensation of $14 per hour.

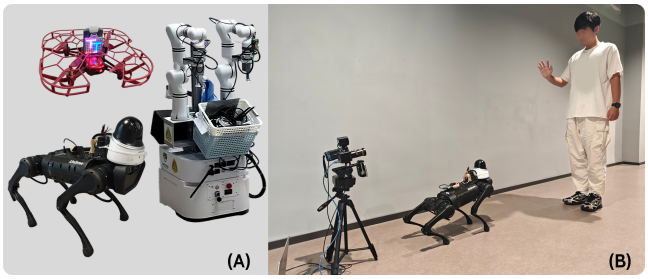


Figure 3: Apparatus of Formative Study. (A) Three robot forms used in the study. (B) Example of a participant performing a gesture toward a quadruped robot.

3.2 Apparatus

A robot's morphology and height can shape both its perceived social role and its perceptual capabilities [29, 60], and following the design rationale in [48], we employed three robot embodiments that map to distinct perceptual perspectives: downward-looking (drone), eye-level (human-height wheeled platform), and upward-looking (quadruped).

Specifically, as shown in Fig. 3(a), we used a Unitree A1 quadruped robot, an AgileX Cobot wheeled platform (approximately human height), and a DJI Tello Talent drone (equipped with propeller guards) as representative embodied AI systems. The quadruped robot carried a front-mounted white camera, the wheeled robot and the drone were outfitted with LED lights to facilitate visual localization by participants — thereby improving users' ability to identify the intended interaction target [23].

3.3 Settings

We manipulated two independent variables: robot embodiment (three forms) and interaction distance (six levels ranging from 3 m to 30 m), resulting in $3 \times 6 = 18$ experimental conditions. To minimize fatigue-related drift in participants' gestures, we randomized the trial order for each participant. Drone trials were conducted outdoors, while the quadruped and wheeled robots were tested in a corridor. Each condition involved two participants simultaneously, and in both environments, occasional passersby provided natural background activity and potential sources of distraction. Additional cameras were positioned near the participants to record their movements for subsequent analysis. All robots were remotely operated by a research team member, who remained out of the participants' immediate view to avoid influencing their gestures.

3.4 Procedure

The experiment comprised two stages, designed to capture both the commands participants wished to issue to different robots and the gestures they employed to convey these commands across distances. At the outset, we measured participants' upper-arm, forearm, palm, and index-finger lengths to provide quantitative references for subsequent analysis. Participants were then introduced to a background scenario (see supplementary materials) crafted to elicit natural and spontaneous gesturing.

In Stage 1 (Command Elicitation), each robot was positioned nearby and teleoperated to showcase its range of motion and interaction capabilities. Participants were asked to propose the commands they would want the robot to execute. They continued until they judged the command set to be comprehensive. All proposed commands were recorded and later categorized by experts.

In Stage 2 (Distance-Conditioned Gesturing), the same robots were placed at distances of 3, 5, 10, 15, 20, and 30 meters (the drone was evaluated only at ≥ 10 m for safety). Participants were instructed to perform gestural realizations of the previously elicited commands while verbally articulating their intended meaning. To preserve ecological validity, we occasionally introduced deliberate pauses before the robot responded, simulating recognition failures and allowing observation of participants' recovery behaviors.

3.5 Gestures Defined By Users

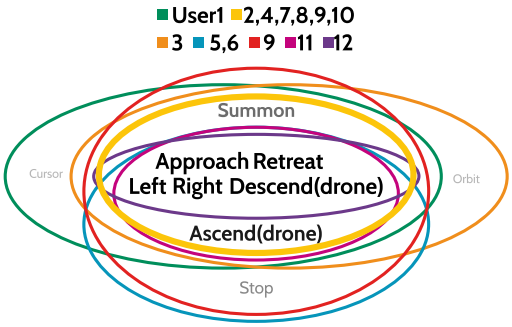


Figure 4: Command Vocabulary Proposed by Participants.

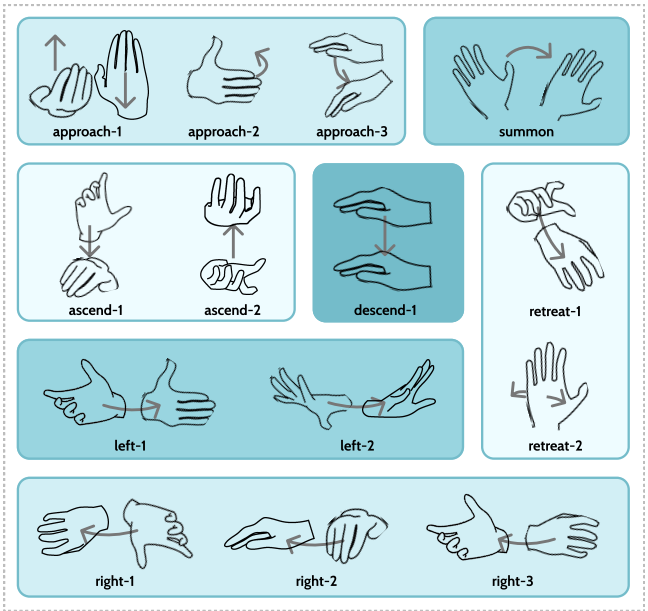


Figure 5: Illustration of the User-defined Gesture Set.

Fig. 4 shows the command vocabulary proposed by participants in Stage 1. From these proposals, we distilled a core set comprising seven commands — "Approach", "Retreat", "Left", "Right", "Ascend" (drone), "Descend" (drone) and "Summon". The first six motion commands unambiguously specify movement direction, whereas "Summon" exhibited two interpretations: (i) attracting the robot's attention and expecting an acknowledgment, and (ii) instructing the robot to continuously approach and stop near the operator (3m on average). We thus expose it as a user-configurable behavior to align system response with user intent. Several low-frequency suggestions were excluded, including "Orbit" for drones (N=1) and "Cursor"-like steering (N=2). We excluded "Stop": a few participants (N=3) used it to compensate for limited obstacle avoidance, but most (N=7) expected autonomous handling, and it is ill-suited to CSI as a static pose.

In Stage 2, the user-defined gesture set is illustrated in Fig. 5. Participants showed high within-class consistency for gestures labeled "Left" and "Retreat", whereas "Right" often involved an additional wrist rotation. Several participants preferred to use both hands or to switch hands during operation. We further observed cross-class visual similarity among "Left", "Approach" and "Descend", indicating potential ambiguity for our system. Results demonstrate high consistency in gesture execution across different robot embodiments when following the same control protocol. This finding is further corroborated by user feedback (P1, P2, P5, P8, P11, P12).

3.6 Quantitative Results

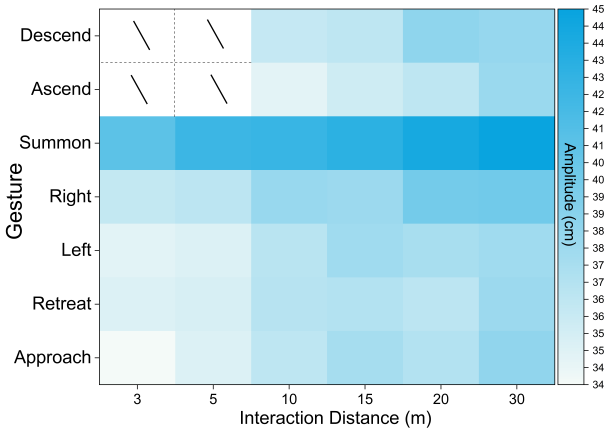
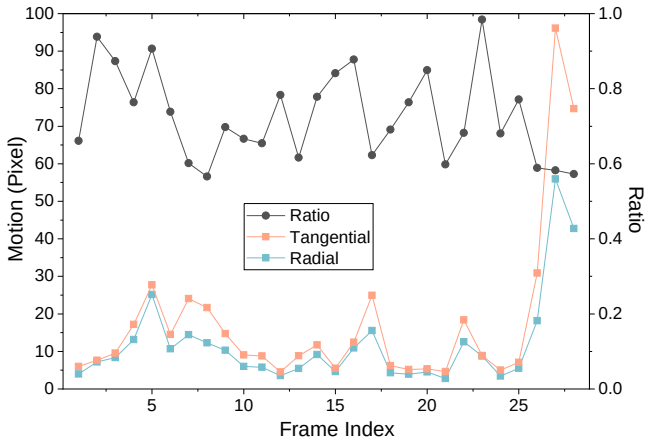
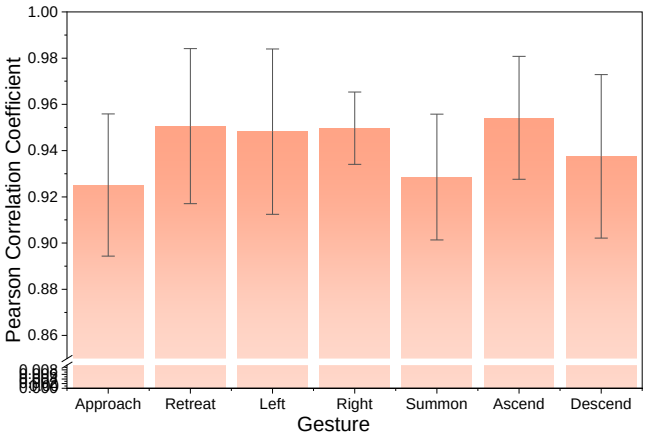


Figure 6: Gesture Amplitude vs. Interaction Distance. The color intensity represents the average amplitude. A general trend of motion amplification is observed as distance increases. Note: Drone-specific gestures (Ascend, Descend) were not evaluated at close ranges (3-5 m) due to safety constraints.

Using each participant's limb measurements (Sec. 10) and an estimated skeleton model [47], we compute gesture amplitude as the displacement of the middle finger-tip. We then analyze amplitude over different robot embodiments and participants (Sec 3.6.1), interaction distance (Sec 3.6.2), and time under a non-response



(a) Illusition of 3D Radial and 2D Tangential Motion. A sampled sequence showing 2D tangential motion (orange) tightly aligns with 3D radial magnitude (blue). Although with fluctuating ratio, this high correlation validates monocular 2D flow as a reliable proxy for 3D dynamics.



(b) Pearson Correlation Coefficient. Average correlations between 3D radial and 2D tangential motion exceed 0.925 for all classes. This consistently high alignment confirms 2D flow as a reliable proxy for 3D dynamics, independent of specific gesture trajectories. Error bars denote standard deviation.

Figure 7: Correlation between 3D Radial and 2D Tangential Motion.

condition (Sec 10.3). Finally, we empirically validate that 2D pixel motion serves as a reliable proxy for 3D gesture (Sec 3.6.3).

3.6.1 Amplitude over Embodiments and Participants. Our gesture set exhibited high within-participant consistency across robot embodiments, suggesting cross-platform generalizability (detailed in App. A). By contrast, we observed huge between-participant differences in amplitude profiles (minimum: User 12 at 16.53 cm; maximum: User 5 at 70.64 cm), indicating strong idiosyncratic kinematics that can support CSI.

3.6.2 Amplitude over Distance. As shown in Fig. 6, gesture amplitude generally increased with interaction distance. However, this trend was not universal: for example, User 1 (detailed in App. A) exhibited nearly distance-invariant amplitude. We also observed that participants without STEM backgrounds rarely consider the robot's sensing limits; they tended to use more conservative motions even with a long interaction distance, thereby potentially degrading recognition accuracy. These behaviors impose stricter requirements on HiSync's ability to capture subtle motion cues.

3.6.3 Correlation between 3D Radial and 2D Tangential Motion. To investigate whether 2D camera data can effectively represent gesture motion information in 3D space, we computed the radial motion magnitude from optical flow data. We then compared it with the tangential optical flow. As shown in Fig. 7, the radial motion exhibited a very strong correlation with the magnitude ($\rho > 0.925$). This finding indicates that, under natural conditions, participants rarely performed movements strictly perpendicular to the camera's focal plane. Consequently, a 2D camera is sufficient to capture the essential motion characteristics of users' gestures.

4 TASK DEFINITION

We address the challenge of Command Source Identification (CSI) in long-range, multi-user environments. In our setup, we identify a specific **target operator** (O_{target}) who holds interaction authority and is equipped with a paired IMU.

As illustrated in Fig. 1, the system input consists of a single continuous inertial stream $\mathcal{I}_{target}$ transmitted from the target's device, and a visual stream $\mathcal{V}$ capturing N individuals $\mathcal{P} = \{P_1, ..., P_N\}$. Crucially, this scene may include bystanders ($P_j, j \neq k$) who are actively performing gestures **(e.g., controlling their own separate robots or interacting with human peers)**.

In this formulation, these bystanders **constitute visual-only negative samples**. Since their inertial data is not transmitted to *this* robot, the system must distinguish the true commander solely by cross-referencing the received $\mathcal{I}_{target}$ with the visual motion fields of all candidates. The goal is to identify the target O_{target} by validating which candidate's visual motion spectrally aligns with the reference inertial signal, thereby rejecting visual distractors regardless of their gestural similarity to O_{target}.

5 DATASET

To address the absence of public datasets for CSI under long-range, multi-user conditions, we construct a multimodal dataset grounded in user-defined natural hand gestures, with synchronized monocular RGB flow and finger-worn IMU streams to enable and evaluate long-range CSI approaches.

5.1 Dataset Characteristics

Gesture Taxonomy. We distilled seven gesture classes (Sec. 3.5) plus a "no-gesture" class composed of unconstrained free-motion sequences. The class distribution is shown in Fig. 9.

Participants. We recruited 38 participants (23 male, 14 female, 1 preferred not to disclose; mean age = 24.66 years, variance = 11.1), including two left-handed individuals. Participants self-reported diverse professional backgrounds and cultural origins. Additionally, we collected measurements of their upper arm, forearm, palm, and index finger lengths.

Figure 8: Visualization of data perspectives. The three camera angles—Upward, Eye-level, and Downward—are designed to emulate the perceptual perspectives of a quadruped, service robot, and drone, respectively.

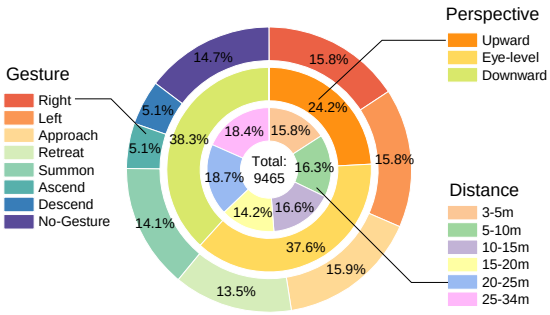


Figure 9: Distribution of Dataset. The chart visualizes the composition of 9,465 sequences, demonstrating a relatively uniform coverage across all dimensions to minimize bias. The rings, from outer to inner, represent gesture categories, camera perspectives, and interaction distances.

Visual. Fig. 8 shows that data were recorded from three viewpoints: eye-level, overhead, and low-angle. Video data were captured using two industrial cameras that are commonly used by robots (Model HBS4800 and Model J5), at a resolution of 1920×1080 and 30 fps. The field of view (FOV) of the main Camera HBS4800 was 75°, while another J5 was a telephoto camera with a focal length range of 10-50 mm, enabling the capture of high-definition close-up

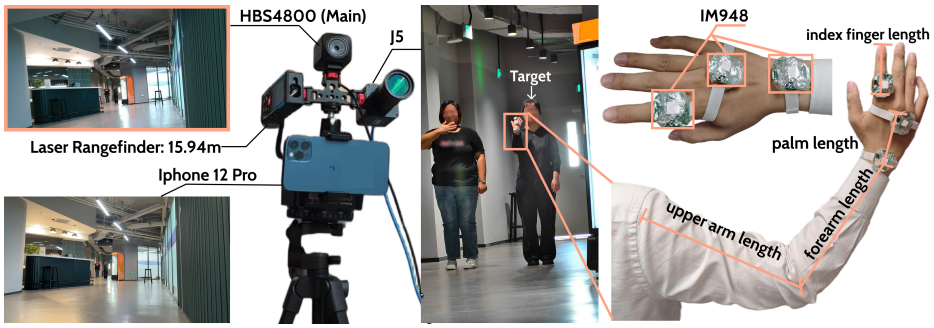


Figure 10: Dataset Characteristics. (Left) Recording apparatus with cameras and laser ranging, paired with visualizations of their data. The orange-framed inset displays the main industrial camera view used for experiments. (Right) Detailed sensor placement and biometrics. IMUs are positioned to simulate smart ring, handheld phone, and smartwatch form factors, alongside anthropometric measurements collected.

views of participants for extracting ground-truth motion trajectories. In addition, we used a smartphone (iPhone 12 Pro) camera with a larger FOV and higher image quality, also recording at 1920×1080 and 30 fps.

Inertial. For inertial sensing, we employed three IM948 9-axis IMUs that logged tri-axial accelerometer, gyroscope, and magnetometer signals; orientation quaternions were then derived. The sensors were affixed at three hand locations — wrist, palm, and the base of the index finger (Fig. 10) — to emulate three common consumer form factors: a smartwatch, a phone held in hand, and a smart ring [27], respectively.

Synchronization and Labels. Each recording consists of three time-synchronized video streams and three IMU signals. Most sequences intentionally include more than two participants to introduce multi-person interference, while all IMUs are worn by a single (target) participant. Ground-truth annotations specify the target participant's distance to the camera, camera viewpoint, gesture class, and the 2D bounding box of the target participant. Labels were produced by an automated detection pipeline and subsequently refined via manual verification to ensure high annotation fidelity.

Dataset Size. The dataset comprises 9465 gesture sequences, totaling 452055 frames of RGB data and corresponding data across all viewpoints and sensors. Data from 8 participants were reserved as the test set, while the remaining 30 participants formed the training set. In total, the training set contains 7530 sequences and 1935 sequences for test set.

5.2 Dataset Collection

Participants were instructed to treat the distant camera as the target device and to perform their naturally preferred gesture for each command. The first 12 dataset participants are the same as those in the formative study. The additional 26 participants did not introduce any new gesture types with non-negligible frequency (i.e., appearing in $\geq$ 1% sequences) beyond the gesture set in the formative study (Fig. 5), thus ensuring the consistency between the formative elicitation and the dataset gesture. Participants repeated each gesture five times to ensure sufficient coverage and consistency.

Data were collected in outdoor environments. To simulate aerial viewpoints for drone interactions, we leveraged building height differences to capture gestures from low-, medium-, and high-angle perspectives. In each recording session, at least two participants were present: one designated participant wore an IMU sensor for gesture tracking, while the other(s) moved freely or actively issued the same commands (accounting for 70.4% of the data). This setup naturally introduced negative samples and simulated multi-user interference. Passersby sometimes entered the field of view, adding uncontrolled motion and occlusion that increased the ecological validity of the dataset.

Recordings captured both natural-scale gestures and adaptive variations. Participants initially performed gestures at a comfortable amplitude and speed, but if the target device did not respond, they may naturally amplify the movement amplitude or frequency. Commands could also be alternated within a recording, allowing us to capture a broad spectrum of gesture expressions, including subtle, exaggerated, and transitional forms like their occurrence in real world.

5.3 Data Processing

Target Human Annotation. For ground-truth identification of the target participant, we used the skeleton keypoints "Right Shoulder", "Left Shoulder", "Right Hip", "Left Hip" to form a body region. This region was then compared with predefined clothing color profiles of participants, and the person with the smallest color difference was selected as the ground-truth subject. Two researchers checked the results independently and resolved all disagreements collaboratively through discussion afterwards.

Temporal Synchronization. We utilized the on-chip CPU perf of the IMU for time alignment. After connecting the IMU to the computer used for video recording, we performed a linear regression to map IMU timestamps to the corresponding video frame times. The final time error is less than one frame of the camera (about 33 ms).

5.4 Privacy Protection

All participants signed an informed consent form. They all consented to the release of their recorded gesture data and associated metadata, and were informed of their right to withdraw their data from the dataset at any time. To ensure participant anonymity, all faces in RGB videos were blurred [56].

6 HiSync ALGORITHM

CSI using IMUs and cameras faces three main challenges. **Firstly**, both modalities are highly noisy: IMU signals suffer from sensor drift and motion artifacts, while cameras at long range produce ambiguities as people occupy only a few pixels with blurred outlines. **Secondly**, IMUs capture acceleration whereas cameras capture positional trajectories. In theory, trajectories can be integrated from IMU data or accelerations derived by differentiating positions, but in practice such transformations are highly inaccurate, leading to inconsistent feature spaces [63, 80]. **Thirdly**, precise temporal synchronization between IMU and camera streams is difficult, as wireless transmission (e.g., Bluetooth) introduces delays and clock drift that accumulate over distances. Unsynced data makes trivial feature matching infeasible for CSI.

6.1 Overview

Fig. 11 provides an overview of HiSync, whose goal is gesture-based command-source identification in multi-user, long-range human-robot interaction. HiSync fuses robot-mounted RGB video and hand-worn IMU streams and reasons in the frequency domain to obtain robust CSI results. The pipeline has two components: (i) a Motion Feature Extractor (Sec. 6.2) that estimates per-person 2D image-plane velocity from RGB video and 3D linear accelerations from the IMU, then transforms both into spectral representations, and (ii) CSINet (Sec. 6.3), a spectral neural network tailored to cross-modal matching under loose synchronization and low SNR. Within CSINet, a Quality-Aware Feature Modulation recalibrates unreliable spectral channels; an IMU-Anchored Cross-Modal Attention module performs soft temporal alignment between the IMU spectra and each person's visual spectra; a Similarity Head produces dimensionless match scores by combining normalized $\ell 2$ and cosine distances; and a Scale-Aware Multi-Window Fusion aggregates evidence over window sizes that differ with camera-estimated range. The network outputs a score for every visible person; the arg-max is taken as the command source and forwarded to the robot controller for downstream tasks.

6.2 Motion Feature Extractor

6.2.1 2D Velocity Estimation. Our objective is to estimate *image-plane hand velocities* for every visible person from the RGB stream. We decompose the task into two stages: (a) **hand-ROI localization** and (b) **hand velocity estimation**.

Hand-ROI Detection. Long-range imagery suffers from a small person scale and frequent keypoint dropouts. We therefore adopt a two-stage pose pipeline. First, a high-recall person detector (YOLOv11x [34], conf > 0.2) first crops regions and then passed to YOLOv11x-pose to obtain a rough skeleton. However, monocular 3D Human Pose Estimation (HPE) remains unreliable at such distances, even state-of-the-art methods [84, 87] fail to capture usable motion cues for CSI. We thus define a *hand region of interest (ROI)* with side length $s_t^{(i)} = \alpha\, h_{t,\mathrm{bbox}}^{(i)}$, where $h_{t,\mathrm{bbox}}^{(i)}$ is the height of person i's bounding box at time t and $\alpha = 0.1$ is selected via experiment. When hand keypoints are missing or low-confidence, we propagate the last valid ROI with short-term velocity prediction and, if needed, infer a proxy ROI from the elbow-to-wrist direction.

VideoFlow. To capture fine 2D motion $\tilde{\mathbf{u}}_t$ under low pixel counts, we compute dense optical flow with *VideoFlow-MOP* [65], which pairs a *Tri-frame Optical Flow (TROF)* module — jointly predicting flows from the middle frame to its preceding and succeeding frames — with a *Motion Propagation (MOP)* module that passes motion cues across adjacent triads to enlarge the temporal receptive field and improve accuracy. For each time t, let $\mathbf{u}_t^+$ denote the flow from frame t to $t{+}1$, and $\mathbf{u}_t^-$ the flow from t to $t{-}1$. We form a per-pixel velocity field $\tilde{\mathbf{u}}_t = \frac{1}{2}\left(\mathbf{u}_t^+ - \mathbf{u}_t^-\right)$.

Flow Pooling. We estimate the hand velocity $\mathbf{v}_t^{\mathrm{hand},(i)}$ for person i by mean pooling $\tilde{\mathbf{u}}_t$ within the hand ROI. Moreover, if the camera exhibits residual motion (e.g., slight pan/tilt), we estimate a global background flow $\mathbf{g}_t$ via robust homography fitting on rigid regions and compensate for the $\mathbf{v}_t^{\mathrm{hand},(i)}$.

6.2.2 3D Linear Acceleration Estimation. We estimate linear acceleration $\mathbf{a}_t^{\mathrm{lin}}$ from a 150 Hz 9-DoF IMU. We compute attitude q_t via *Madgwick AHRS* [49], fusing gyroscope ($\boldsymbol{\omega}_t$), magnetometer ($\mathbf{m}_t$), and accelerometer ($\mathbf{a}_t^{\mathrm{meas}}$). We reject dynamic disturbances by scaling the accelerometer gradient with Madgwick parameters: base gain $\beta = 0.035$ and weight $w_a = \mathrm{clip}(1 - |\|\mathbf{a}_t^{\mathrm{meas}}\| - g|/3, 0, 1)$. World-frame linear acceleration is then obtained by removing pre-calibrated sensor bias $\mathbf{b}_a$ and gravity $\mathbf{g}_w$: $\mathbf{a}_t^{\mathrm{lin}} = \mathbf{R}(q_t)(\mathbf{a}_t^{\mathrm{meas}} - \mathbf{b}_a) - \mathbf{g}_w$, followed by a 15 Hz Butterworth low-pass filter.

6.2.3 Spectral Analysis. Directly matching raw 2D image-plane velocities from RGB with 3D linear accelerations from the IMU is brittle in long-range, multi-person settings due to sensor and estimation noise and loose cross-modal synchronization. We therefore transform both streams into the frequency domain. Working in frequency space (i) provides a common, modality-comparable basis for both velocity and acceleration, (ii) reduces sensitivity to temporal offsets, and (iii) suppresses uncorrelated noise through windowing and band-wise aggregation. This spectral projection serves as the frontend for CSINet.

Validation of Comparing Acceleration and Velocity on Spectral. Let $v(t)$ be a velocity component and $a(t) = \dot{v}(t)$ the corresponding acceleration. By the *differentiation property* of the Fourier transform $\mathcal{F}$, their spectra are related by: $A(\omega) = \mathcal{F}\{\dot{v}(t)\} = j\omega\, V(\omega)$, where $A(\omega)$ and $V(\omega)$ are the Fourier transforms of $a(t)$ and $v(t)$. Consequently, for any angular frequency $\omega \neq 0$:

$$|A(\omega)| = |\omega|\,|V(\omega)|, \quad \angle A(\omega) = \angle V(\omega) + \frac{\pi}{2}\mathrm{sgn}(\omega). \tag{1}$$

Here, $|\cdot|$ is the complex magnitude, $\angle(\cdot)$ is the principal phase and $\mathrm{sgn}(\cdot)$ denotes the sign function. Therefore, defining a normalized acceleration spectrum $\tilde{A}(\omega) \triangleq A(\omega)/(j\omega)$ yields $\tilde{A}(\omega) \equiv V(\omega)$. This confirms that acceleration features inherently encode velocity

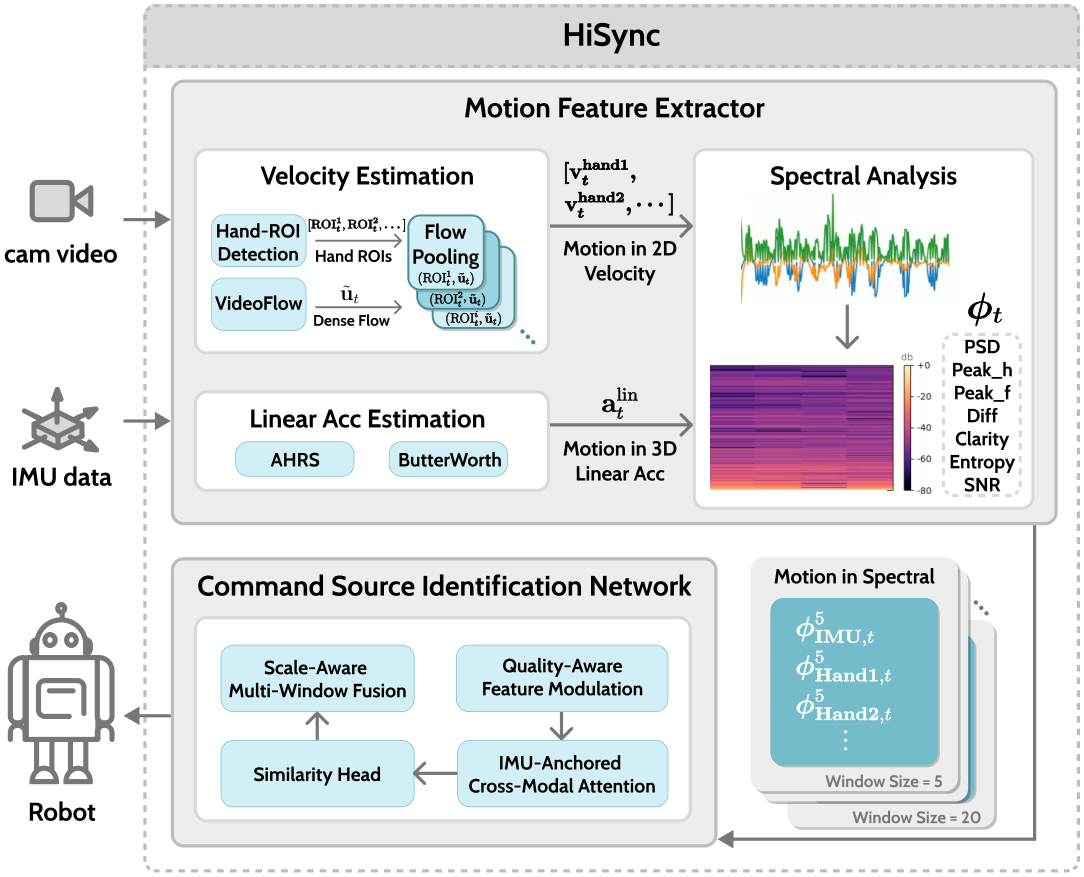


Figure 11: Overview of HiSync. The framework fuses robot-mounted RGB video and wearable IMU signals in the frequency domain for robust Command Source Identification (CSI). It consists of two main stages: (1) a Motion Feature Extractor that estimates and transforms visual velocity and inertial acceleration into spectral representations; and (2) CSINet, which aligns and matches these cross-modal features using Quality-Aware Feature Modulation, IMU-Anchored Attention, and Scale-Aware Multi-Window Fusion to identify the target operator.

patterns, differing only by a known frequency-dependent scaling and phase shift.

Validation of Utilizing Spectral to Avoid Time Shift. *Spectral features are shift-tolerant within a window.* If a global time shift τ exists between v and a (e.g., from sensor/video misalignment), then $A(\omega) = e^{-j\omega\tau}\, j\omega V(\omega)$; subtracting the linear phase $e^{-j\omega\tau}$ restores the result.

Validation of Spectral-domain Denoising. Unlike time-domain matching, which is vulnerable to spikes and drift, our spectral approach naturally attenuates these artifacts. Specifically, transient disturbances disperse energy across wide bands, minimizing local impact; narrowband sensor noise [1] is implicitly masked via band selection by the downstream neural network; and low-frequency drift or gravity leakage [24] is eliminated by excluding the near Direct Current neighborhood. This ensures robustness to modality-specific noise while preserving discriminative motion cues.

Spectral Feature $\boldsymbol{\phi}$. Given a length-N sequence x, we apply a window size w and compute the discrete Fourier transform (DFT) $X[k]$, where k denotes the frequency bin index. We then form the power spectral density (PSD) $S[k] = |X[k]|^2$ which serves as our primary feature denoted as **PSD**. From the PSD, we derive a compact set of interpretable descriptors: peak height (p), peak frequency (f), spectral clarity (κ), spectral entropy (H), frequency spacing (Δf), in-band SNR (**SNR**), and average power (P_{avg}). Detailed definitions and ablations are provided in the App. B. These descriptors, concatenated with the full PSD, form the final feature vector:

$$\boldsymbol{\phi}(x) = [\mathbf{PSD}; \text{Feat}] = \left[\mathbf{PSD},\, p,\, f,\, \kappa,\, H,\, \Delta f,\, \text{SNR},\, P_{\text{avg}}\right]. \quad (2)$$

Different DFT window lengths capture complementary spectral cues, much like different receptive fields in a CNN. We compute spectral features for window lengths $w \in \{5, \ldots, 20\}$ (frames) and denote the feature vector from window w as ϕ^w. These multi-scale spectral descriptors offer complementary time-frequency resolutions.

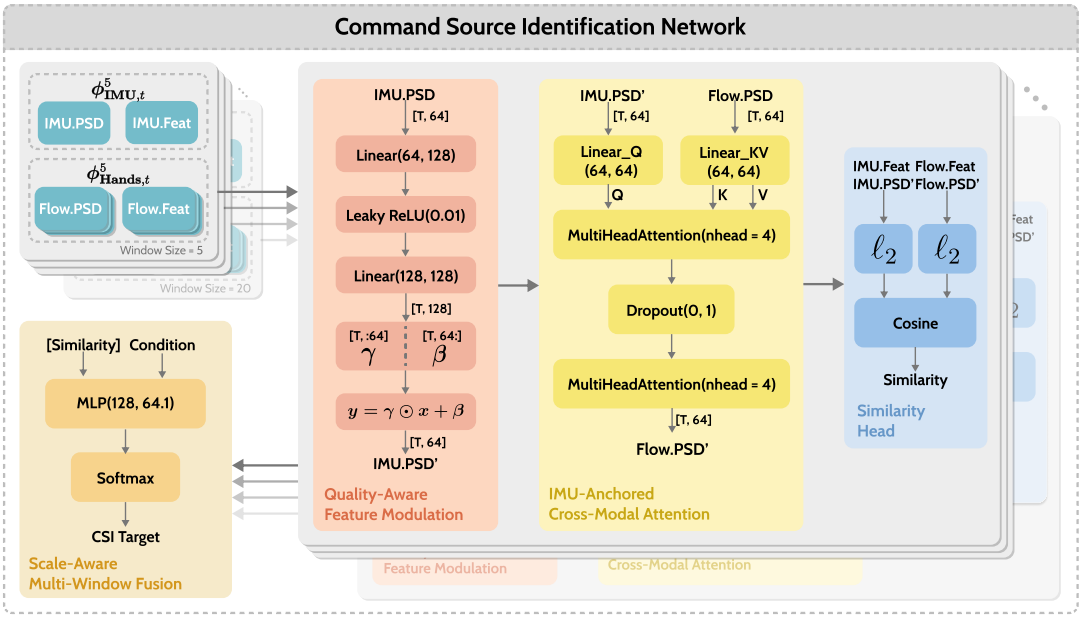


Figure 12: Architecture of Command Source Identification Network. Taking spectral features as input, for each window size, the network first applies Quality-Aware Feature Modulation to suppress noise in IMU signals. It then utilizes IMU-Anchored Cross-Modal Attention to align visual flow sequences with the inertial reference. Finally, the Similarity Head computes cosine scores, which are dynamically aggregated by Scale-Aware Multi-Window Fusion across multiple window sizes to output the final similarity.

6.3 Command Source Identification Network

6.3.1 Model Architecture. As illustrated in Fig. 12, the Command Source Identification Network (CSINet) operates on precomputed spectral feature vector ϕ and comprises four modules: (1) *Quality-Aware Feature Modulation*, which recalibrates IMU spectral features using per-frame quality cues; (2) *IMU-Anchored Cross-Modal Attention*, which aligns each candidate's flow-spectral sequence to the IMU time reference, mitigating jitter, missed detections, and phase drift; (3) *Similarity Head* that ℓ_2-normalizes temporally aggregated embeddings and outputs a cosine-similarity score for each IMU-candidate pair; and (4) *Scale-Aware Multi-Window Fusion*, which evaluates results from multiple window lengths and fuses their scores via learned weights conditioned on a distance proxy. This design yields robust command-source identification in crowded, long-range settings.

Quality-Aware Feature Modulation. We observed that some segments contain weak or non-purposeful motion, degrading similarity estimates. To emphasize informative frames while suppressing unreliable channels, we adopt a FiLM operator [59] on the IMU spectral features. Given a per-frame PSD embedding $x_t \in \mathbb{R}^D$ (composed of IMU.PSD and Flow.PSD), a lightweight MLP predicts channel-wise affine parameters $(\gamma_t, \beta_t) = \mathrm{MLP}(x_t)$. The modulated representation $y_t = \gamma_t \odot x_t + \beta_t$ recalibrates frequency channels (e.g., down-weights those affected by occlusion or low detection confidence) prior to cross-modal alignment. Compared with simple concatenation, Quality-Aware Feature Modulation improves robustness with low-quality. In HiSync, Quality-Aware Feature Modulation is applied to IMU features, producing IMU.PSD′.

IMU-Anchored Cross-Modal Attention. Treating the IMU sequence as the temporal anchor, we perform soft alignment and evidence selection over each candidate's flow-spectrum embeddings. We compute

$$\begin{aligned} Q &= W_q \cdot \text{IMU.PSD}', \quad K = W_k \cdot \text{Flow.PSD}, \\ V &= W_k \cdot \text{Flow.PSD}, \quad \text{Flow.PSD}' = \mathrm{MHA}(Q, K, V), \end{aligned} \tag{3}$$

where MHA is multi-head attention with 4 heads, stacked for two layers with dropout 0.1. The resulting Flow.PSD′ represents the candidate's motion "as seen from the IMU's perspective," mitigating frame-level mismatch due to jitter, intermittent detection, or phase shifts.

Similarity Head. Because IMU and optical flow live in different units (m/s^2 vs. pixels/frame) yet share discriminative spectral amplitude, we remove scale effects by ℓ_2-normalizing temporally aggregated embeddings and use cosine similarity as the cross-modal score:

$$\mathrm{sim}(\phi_{\text{IMU}}, \phi_{\text{Flow}}) = \frac{\phi_{\text{IMU}}^{\top} \phi_{\text{Flow}}}{\|\phi_{\text{IMU}}\|_2 \, \|\phi_{\text{Flow}}\|_2}, \tag{4}$$

ϕ is the spectral feature defined in Spectral Feature Selection

This score quantifies the motion-pattern consistency between the IMU segment and each candidate and is used for ranking and final identification.

Scale-Aware Multi-Window Fusion. User motion varies with gesture and distance; a single temporal extent may be sub-optimal. We therefore evaluate similarities over a set of DFT window lengths

$\{w\}_5^{20}$. These DFT windows are different from the sliding window used to segment the streaming input sequence. And fuse them conditioned on an estimated scale s (approximated by the ratio between the person's bounding-box width and the image width). An MLP takes s (optionally concatenated with quality statistics) and outputs fusion weights $\alpha_m = \mathrm{softmax}(g(s))$. The final similarity is the weighted average of sim under various window sizes m.

$$\mathrm{Sim}_{\mathrm{final}} = \sum_{m=1}^{M} \alpha_m \, \mathrm{sim}_m, \tag{5}$$

adapting the effective temporal context to distance while preserving robustness across heterogeneous motion regimes.

6.3.2 Model Training.

Training. We train CSINet end-to-end in PyTorch using AdamW with cosine annealing. We run for 50 epochs with early stopping after 3 epochs without validation improvement, batch size 4, initial learning rate 1e − 4, and weight decay $1{\times}10^{-4}$. The scheduler uses $T_{\max} = 25$ (half of total epochs) and $\eta_{\min} = 0.01\times$ the initial learning rate.

Batching and masking. To accommodate multi-person scenes and variable-length sequences, we perform batchwise *align-and-pad*: within each mini-batch, we pad time steps, the number of temporal scales, and the number of candidates to their within-batch maxima and construct frame and person masks to ignore padded entries. Temporal features are aggregated via *masked mean pooling* to obtain segment-level embeddings. We then apply ℓ_2 normalization to the IMU representation and to the attention-aligned flow representation to improve numerical stability and the effectiveness of contrastive learning.

Loss. The principal component is the InfoNCE contrastive loss, computed over per-segment, per-candidate scores. To prevent overfitting and ensure generalization, we incorporate an L_2 regularization term on the network parameters θ. For the i-th sample with similarity logits $\mathbf{s}_i = [s_{i,1}, \ldots, s_{i,P_i}]$ and positive index y_i, the contrastive term is defined as:

$$\mathcal{L}_{\mathrm{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s_{i,y_i}/\tau)}{\sum_{j=1}^{P_i}\exp(s_{i,j}/\tau)} + \lambda\|\theta\|_2^2, \tag{6}$$

In practice, we clip logits to a safe range and sanitize NaN/Inf values to prevent gradient pathologies.

Augmentation. To strengthen discrimination against distractors, we adopt *cross-sequence negative augmentation*: for each training instance we append time series from randomly sampled people in other sequences along the candidate dimension, so each batch contains 5 external negatives. The original positive index y_i is preserved throughout.

7 EXPERIMENTS on DATASETS

7.1 Evaluation Protocol

We evaluate HiSync on our long-range gestural interaction dataset (Sec. 5). Unless otherwise noted, all results are reported as the mean over 5 runs with different random seeds.

7.1.1 Implementation and Training Strategy. Since the official implementations of the baselines [5, 9, 66] are not publicly available and were originally designed for shorter ranges, we re-implemented their architectures with necessary adjustments to adapt to far-range scenarios. For image-based baselines, input sequences are sampled at 5 fps. To ensure a rigorous subject-independent evaluation, we use data from 30 participants for training, while the remaining 8 participants are reserved as the **held-out** test set. Within the training set, we split at 80/20 for training and validation.

7.1.2 Robustness Test Set Construction. All experiments are evaluated on the 8 held-out participants. To rigorously assess the system's robustness against real-world synchronization failures [21], we introduce artificial temporal noise into the synchronized test set. Specifically, we apply three types of temporal perturbations between the input streams:

- **Global Offset.** To simulate transmission latency, we apply a global time shift Δt for each sequence. The offset is sampled from a uniform distribution: $\Delta t \sim \mathcal{U}(-500, 500)$ ms.
- **Clock Drift:** To mimic the sampling rate mismatch between independent device clocks (e.g., camera vs. IMU crystal oscillators), we apply a linear scaling factor $\alpha \sim \mathcal{U}(0.98, 1.02)$ to the timestamps.
- **Jitter:** To simulate packet arrival jitter and processing fluctuations, we add independent Gaussian noise to the timestamp: $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where $\sigma = 60$ ms (approx. 2 frame duration).

7.2 Baselines

To assess the perceptual limits of vision-based models at long range, we compare against two representative long-distance gesture recognition research and a cross-modal CSI method (VIPL), and we additionally construct a skeleton-based variant of our pipeline to probe the benefit of spectral representations for CSI.

7.2.1 DNN [9]. This vision-only baseline takes an RGB picture as input. A shallow spatiotemporal stem extracts low-dimensional features; a patch-selection subnetwork highlights gesture-relevant regions and suppresses background; and a patch-level classifier predicts the gesture class. The original work reports "long-distance" results at roughly 4 m. We introduced smaller patch sizes beyond 10 m and a higher loss weight on patch selection to encourage tighter focus.

7.2.2 URGR [5]. URGR targets ultra-long-range (up to 25 m) gesture recognition with a single RGB camera by combining an image enhancement module (HQ-Net) with a GVIT backbone that fuses GCN and a modified ViT. To test distances beyond 25 m, we replace HQ-Net with a state-of-the-art image restoration model [81].

7.2.3 VIPL [66]. VIPL tackles command source identification without prior appearance enrollment. While sharing similar input modalities with HiSync, it fundamentally differs by modeling temporal motion dynamics in the time domain, heavily relying on large-scale full-body walking cues and assuming tight synchronization. Specifically, it pairs visual features (optical flow and keypoints) with raw inertial signals (acceleration and angular velocity). A dual-branch LSTM-based network maps both streams into a shared embedding space optimized by a triplet loss, selecting the person with the minimum feature distance at inference.

7.3 Metrics

Unless otherwise noted, a segment is counted as correct under *strict accuracy* if the predicted candidate exactly matches the annotated command source. But for the vision-only baseline DNN (Sec. 7.2.1) and URGR (Sec. 7.2.2), which is a single-frame gesture recognizer rather than a multi-person CSI model, we adapt its outputs as follows: a frame is deemed correct if its predicted gesture matches the ground-truth command issued by the true source; segment-level decisions are then obtained by majority vote over frames within the segment. Because this baseline has no capability to resolve which person performed the gesture, the above protocol effectively grants it oracle source attribution; therefore, the resulting CSI accuracy should be interpreted as an *upper bound* on what a purely vision-based, single-frame recognizer could achieve in our multi-person setting.

7.4 Results and Analysis

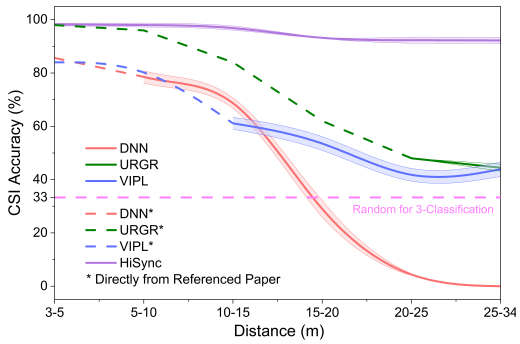


Figure 13: Results of Experiments. This figure shows the CSI Accuracy vs. Distance. The dashed line represents the data obtained directly from the referenced paper. The results show that HiSync outperform baselines, especially at long distances. Error bar denotes standard deviation.

Figure 14: Failure Case. This is a failure sequence where the two persons move in near-perfect sync. However, because the HiSync can capture subtle differences in movements, the failure rate is very low even in such a rare case.

7.4.1 Overall Performance. HiSync sustains reliable command-source identification (CSI) across long ranges, outperforming both purely visual baselines and state-of-the-art optical-inertial methods. With about 3 people visible per frame on average and a sequence length of 3 seconds, it achieves 99.77% CSI accuracy at $\leq$ 10m, 15.77% over VIPL's reported 84% and comparable with a vision-only upper bound (100% at 3-5 m; 98% at 5-10 m). Between 10-30 m, HiSync maintains 96.64% accuracy, beating the VIPL baseline (70.34%) by 26.30% as that method degrades sharply with distance. Even at the extreme range (up to 34 m), HiSync still attains 94.31% accuracy, whereas baselines approach random (VIPL, three-category classification: 43.88%; vision-only gesture recognizer, eight-category: 44.48%).

7.4.2 Fail Case Studies. As shown in Fig. 14, in rare instances two participants exhibit gestures in near-perfect sync — sharing dominant frequencies, harmonic structure, and amplitude envelopes, thus yielding nearly equal cross-modal similarity to the query IMU, a condition we term a "spectral collision". In such cases, leveraging a longer temporal context is warranted; see Sec. 7.5.3 for details.

7.5 Ablation

We did ablation studies on **Optical Flow** (Skeleton), **Quality-Aware Feature Modulation** (-FiLM), **Scale-Aware Multi-Window Fusion** (-Multi-window) for HiSync. See Fig. 15 for quantitative comparisons.

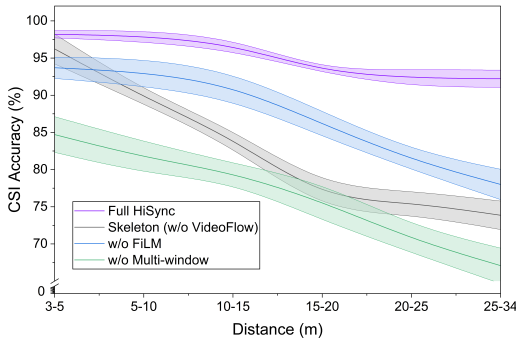


Figure 15: Ablation Results. We compare the full system against ablated variants across distances. The significant performance drop in incomplete models highlights the necessity of each component for maintaining robust CSI. Error bar denotes standard deviation.

7.5.1 Ablation on Optical Flow. To isolate the value of Optical Flow feature for CSI at long range, we replace the **Velocity Estimation** (Sec. 6.2.1) in our method with per-person 2D skeleton trajectories and encode them using a lightweight MLP in lieu of Spectral Analysis (Sec. 6.2.3 for 2D velocity). Specifically, for each candidate, we form a concatenated keypoint trajectory vector over the window, reduce it to 64 dimensions with a linear layer, and keep all downstream components unchanged. This variant performs markedly worse at longer distances, where pose detection becomes unreliable as keypoints are often missing, jittery, or mislocalized, confirming

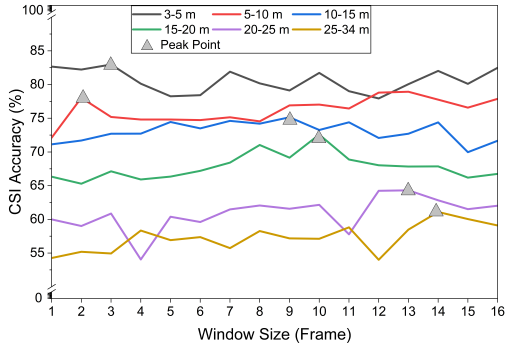


Figure 16: Impact of Window Size on Accuracy. This figure shows a clear trend where optimal window size increases with distance. This validates our design choice to fuse the different temporal window sizes.

that Optical Flow motion features provide greater robustness under low resolution and high noise.

7.5.2 Ablation on Quality-Aware Feature Modulation. We disable the **Quality-Aware Feature Modulation** (Sec. 6.3.1) and keep the rest of the architecture and training schedule fixed. Without Quality-Aware Feature Modulation, the model cannot down-weight low-quality or noise-dominated frequency channels, leading to consistently lower accuracy across distances and harder training convergence. Errors concentrate in segments with micro-motions that reduce SNR of IMU, indicating that Quality-Aware Feature Modulation's channel-wise modulation is critical for stabilizing learning and improving robustness under persistent IMU noise.

7.5.3 Ablation on Scale-Aware Multi-Window Fusion. We replace **Scale-Aware Multi-Window Fusion** (Sec. 6.3.1) with a single-scale baseline and sweep several window lengths for fairness. We report per-distance accuracies for each window size in Fig. 16. Removing scale-aware fusion yields a substantial overall drop, and the window that maximizes accuracy varies with distance. Generally, larger windows perform better at longer ranges while shorter windows suffice at close range. Intuitively, long windows capture global motion characteristics (for example, overall swing frequency), whereas short windows better capture fine-grained cues (for example, arm length proxies and personalized way of exerting force). These results support our design choice to fuse temporal scales rather than rely on a single window.

7.6 Evaluation on Temporal Misalignment

Beyond the theoretical justification (Sec. 6.2.3), we empirically evaluate the method's robustness against temporal misalignment under progressively severe desynchronization scenarios.

7.6.1 Ablation on Spectral Transformation. To validate the necessity of the spectral transformation, we bypass the Spectral Analysis (Sec. 6.2.3) step. Instead, the raw IMU linear acceleration and visual velocity sequences are fed directly into a network with an architecture identical to CSINet. This "w/o Spectral" setting serves to verify the impact of spectral features on alignment robustness. Furthermore, to benchmark against traditional temporal alignment techniques, we employ Linear Regression [21] (Linear), Cross-Correlation [37] (XCorr) and Dynamic Time Warping [7] (DTW) to pre-align the visual and inertial streams in the time domain before feeding them into the "w/o Spectral" network.

7.6.2 Progressive Noise. To simulate real-world conditions ranging from ideal to severe, we construct four test sets with varying degrees of temporal noise. Beyond the synchronized set T_{clean} and the severe noise set T_3 (detailed in Sec. 7.1), we introduce two intermediate sets. T_1 features only a mild global offset $\Delta t \sim \mathcal{U}(-200, 200)$ ms. T_2 adopts the global offset and clock drift from T_3, but excludes transmission jitter to isolate the impact of irregular fluctuations.

7.6.3 Results and Analysis. The comparative results in Tab. 2 reveal significant performance divergences under temporal noise. The "w/o Spectral" baseline deteriorates sharply on noisy sets, confirming the fragility of direct time-domain matching against misalignment. Although traditional pre-alignment methods (Linear, XCorr, DTW) mitigate simple global offsets (T_1), they fail to address the compounding drift and jitter in T_3. In contrast, HiSync maintains high robustness across all noise levels with only marginal drops, empirically validating that spectral magnitude features provide inherent invariance to temporal noise, effectively decoupling recognition performance from precise synchronization.

8 ONLINE EXPERIMENTS on REAL ROBOT

To thoroughly evaluate the system's performance in real-world deployment and gather subjective user feedback, we deployed HiSync on a quadruped robot and conducted both correctness tests and interactive usability experiments in dynamic indoor and outdoor environments.

8.1 Participants and Apparatus

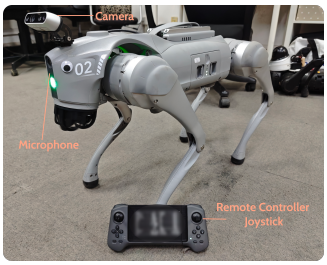


Figure 17: Real Robot Apparatus. The real-robot experiment is conducted on a *Unitree Go2* quadruped robot, equipped with a *Realsense D435* camera. The onboard microphone and official joystick are used in the *Voice Control* and *Remote Controller* baseline respectively.

We recruited 12 participants (N = 12, 6 male and 6 female; mean age = 23.9, SD = 3.7) who had not participated in the formative study or dataset collection. Their expertise with robotics varied from novice to expert.

As shown in Fig. 17, the system was deployed on a *Unitree Go2* quadruped robot equipped with a *Realsense D435* camera running

Table 2: Evaluation Results on Temporal Misalignment. We compare HiSync against the time-domain ablation ("w/o Spectral"), pre-alignment methods (Linear, XCorr, DTW) applied to the time-domain backbone, and the VIPL baseline. Temporal misalignment increases progressively from $\mathcal{T}_{clean}$ to $\mathcal{T}_3$. Results demonstrate that HiSync maintains high robustness even under severe synchronization noise ($\mathcal{T}_3$) where time-domain methods degrade sharply.

Methods	Accuracy(%) ± SD			
	T_{clean}	T_1	T_2	T_3
HiSync	97.01(±0.26)	96.47(±0.43)	95.70(±0.95)	95.14(±0.97)
w/o Spectral	76.83(±0.60)	65.19(±0.64)	59.16(±3.11)	50.38(±6.82)
Linear [21]	78.18(±1.17)	73.38(±6.37)	65.81(±3.32)	51.82(±5.98)
XCorr [37]	79.11(±0.58)	74.95(±6.79)	63.82(±3.36)	53.27(±5.91)
DTW [7]	77.98(±1.04)	74.91(±1.32)	72.09(±1.62)	64.74(±4.32)
VIPL [66]	70.49(±1.52)	65.89(±2.23)	65.06(±1.66)	61.05(±2.70)

in RGB mode at 1280×720 resolution (30 fps). Based on the findings in Sec. 10.1, we utilized the wrist placement for the IMU sensor to balance signal quality and usability. IMU data was collected by a host computer and transmitted to the robot via the public network. This setup introduces realistic IOT network latency (measured at ≈ 387,ms on average with ≈ 50,ms jitter), strictly testing the system's robustness under non-ideal transmission conditions. Furthermore, unlike the batched inference used in offline evaluations, the real-robot system performs frame-by-frame inference, resulting in an effective processing rate of approximately 5,Hz due to computational overhead.

8.2 Design and Procedure

The study consisted of two sessions: (1) an Interaction Usability Experiment and (2) a CSI Accuracy Experiment. Both sessions covered new indoor (3 m, 5 m) and outdoor (10 m, 20 m, 30 m) scenarios that were not included in the dataset.

8.2.1 Interaction Usability Experiment. To assess interaction usability, we compared HiSync against standard modalities in a robot summoning task across varying distances. (i) **Remote Controller (RC)**, utilizing the manufacturer's standard joystick; (ii) **Voice Control**, employing the robot's built-in wake-up word and command interface; and (iii) **HiSync**, where participants triggered the approach via a natural right-hand wave mapped to the "Summon" semantic. Notably, Voice Control was exclusively evaluated at the short range (3 m), as participants consistently declined to use voice at longer distances (>5 m) due to the social embarrassment associated with shouting and the system's technical inability to capture distant audio. To minimize learning effects, modality and distance order were counterbalanced across participants.

8.2.2 CSI Accuracy Experiment. To evaluate robustness against active interference, participants worked in pairs at an **adversarial setting**. One participant acted as the **Owner**, issuing the five standard commands defined in Sec. 3.5. The other acted as a **Strong Negative Sample (Mimic)**, instructed to mimic the owner's movements as closely as possible to "trick" the robot into responding to them. Additionally, the experiments were conducted in public spaces with uncontrolled passersby, who served as **Weak Negative Samples** (providing occlusion and background noise).

8.3 Results and Analysis

8.3.1 Objective. As shown in Tab. 3, we evaluated the CSI accuracy and the response time (from gesture onset to robot reaction). Despite the challenging real-world settings involving adversarial mimics (strong negatives) and uncontrolled pedestrians (weak negatives), HiSync maintains high accuracy and operational efficiency. Notably, even at the extended distance of 30 m, HiSync achieves an accuracy of 90.9% with an average response time of 3.88 s, demonstrating its viability for long-range field deployment. These results demonstrate that HiSync remains effective when deployed on a real quadruped robot and with strong bystander interference. Compared to the training dataset, the real-robot study involves new participants, a different camera configuration, and previously unseen indoor and outdoor environments, suggesting that HiSync can generalize from curated offline data to in-the-wild deployments.

8.3.2 Subjective. We evaluated subjective feedback using the NASA-TLX [26] and a subset of the System Usability Scale (SUS) [10]. A Wilcoxon Signed-Rank Test was conducted to determine statistical significance between HiSync and the baselines (Voice and Remote Controller), as shown in Figure 18. Note that specific NASA-TLX metrics were inverted for consistency, so higher scores always indicate better performance.

Quantitatively, HiSync achieved the highest mean scores across all metrics. Participants reported a significantly higher willingness to use HiSync compared to both Voice ($p < 0.00049$) and the Remote Controller (RC) ($p = 0.034$). HiSync was significantly easier to learn ($p < 0.00049$) and use ($p < 0.005$) than the RC, effectively combining the intuitiveness of natural voice interaction (*n.s.* in ease of use) with the reliable performance of the RC (*n.s.* in performance). Regarding task load (TLX), HiSync significantly outperformed both baselines across all sub-scales (all $p < 0.03$). Deeper analysis revealed that interaction distance influenced these preferences (Fig. 19: at close range, participants felt the RC was "overkill" (P4, P7). Conversely, as the distance increased, distinct issues emerged: participants noted they "could not see the robot dog clearly" (P4) and that "control errors with the remote controller increased" (P9). Moreover, specifically for users unfamiliar with robots, the workload of operating at long distances increased significantly ($p < 0.05$).

Table 3: Quantitative Results of Real-World Experiments. We report the mean CSI accuracy and response time (from gesture onset to robot reaction) under adversarial settings with strong mimics. Data are presented as Mean ± Standard Error (SE).

(a) Performance vs. Interaction Distance. This table illustrates the effectiveness of HiSync under different real-world distance conditions.

Distance(m)	CSI Accuracy (%) ±*SE*	Time(s) ±*SE*
3	96.4 ± 2.55	3.46 ± 0.159
5	94.5 ± 4.03	3.23 ± 0.130
10	92.7 ± 3.53	3.07 ± 0.150
20	92.7 ± 3.53	3.09 ± 0.130
30	90.9 ± 3.91	3.88 ± 0.175

(b) Performance vs. Gesture Category. This table demonstrates the generalization of the proposed gestures under HiSync.

Gesture	CSI Accuracy (%) ±*SE*	Time(s) ±*SE*
Right	92.7 ± 3.534	3.12 ± 0.132
Left	96.4 ± 2.547	3.47 ± 0.175
Approach	90.9 ± 3.912	3.33 ± 0.157
Retreat	90.9 ± 3.912	3.43 ± 0.161
Summon	96.4 ± 2.547	3.38 ± 0.176

Participants praised the system's **naturalness** and **social presence**, noting it "looks like a real dog" (P5, P8, P11) and is "suitable for public spaces" (P7, P12), with emphasis that "It was excellent, and I felt very confident" (P4, P12).

9 APPLICATIONS

9.1 Public-venue service-robot interactions

In public indoor settings such as airports, train stations, and shopping malls, our CSI pipeline enables users to summon and direct service robots from a distance using brief hand gestures (for example, *Summon*, *Approach*, *Left*, *Right*). *Imagine entering a shopping mall, making a short wrist wave with a smartwatch and within moments a nearby service robot recognizes the gesture, navigates to the user for further instruction.* Technically, the system aligns far-field optical-flow spectra captured by robot-mounted cameras with IMU traces from the user's wearable, which lets the pipeline identify who issued a command in crowded scenes without explicit per-user enrollment, fiducial markers, or spoken input. This maintains socially appropriate interaction in reverberant or quiet environments.

9.2 Robotaxi summons

In outdoor parking lots and similar open environments, a user can invoke a *summon→approach* gesture from tens of meters away to request a vehicle or low-speed carrier. Operating in the frequency domain, cross-modal spectral matching is inherently more tolerant to long-range visual degradation and variable illumination. The proposed setup that requires only a monocular camera on the vehicle and a wearable IMU minimizes infrastructure requirements while still supporting explicit confirmation cues and escalation procedures when the initial signal is ambiguous.

9.3 "Hi and Sync" Paradigm

CSI can serve as a lightweight selection mechanism to bind an individual to a specific robot and subsequently enable behaviors such as *Follow-Me* and *Approach & Align*. The same spectral alignment that disambiguates commands in multi-user scenes also reduces interference from nearby bystanders, supporting reliable selection, re-binding, and handover among co-located robots, making far-range "Hi and Sync" a feasible primitive for applications such as guided tours, patrols, and parcel delivery.

10 DISCUSSION

10.1 Choice of IMU position

As illustrated in Fig. 10, we compared three placements on the dominant hand. Beyond sensing fidelity, these choices reflect applicability in practice: smartwatches are the most widely adopted wearables, smart rings promise the least obtrusive "always-on" wear in the near future, and hand-held phones require no additional hardware.

Performance results (Fig. 22) show that the palm IMU achieves the highest CSI accuracy, likely due to stronger distal motion and firmer sensor-gesture coupling that preserves high-frequency components exploited by our spectral encoder. The ring underperforms, plausibly due to higher micro-motion noise, while the relatively poor performance for wrist-mounted IMU may be attributed to the wrist's less discriminative motion patterns, as well as noise introduced by forearm movements.

Design Implications. When robustness at long range is paramount, prefer palm-held. Smartwatches trade some accuracy for ubiquity and convenience, and rings favor wearability but require tighter stabilization and higher-quality sensing; phones offer strong accuracy without extra devices when hands are occupied.

10.2 Maximum Interaction Range

We set the maximum interaction range to 34 m. At this distance, a person occupies too few pixels for reliable motion estimation; consequently, our optical-flow pipeline (VideoFlow) no longer produces stable hand-motion cues (Fig. 21). Although a stronger flow estimator or a super-resolution pre-processing step could, in principle, extend this range slightly, we deliberately retain 34 m as a conservative upper bound for two reasons. First, beyond this range, users themselves find it difficult to visually acquire and keep the intended interlocutor in view. Second, in a brief interview study (N=9), participants consistently reported that, under everyday conditions, they would not attempt to initiate interaction at or beyond 34 m. This observation aligns with proxemics-based HCI research: engagement willingness drops sharply in the far public zone and typically requires additional signifiers or explicit invitation [4, 73].

10.3 Amplitude over Time

Despite some participants using low-amplitude gestures even at very long ranges, we observed a reliable self-adaptation: when the robot did not respond, all participants amplified their gestures as in

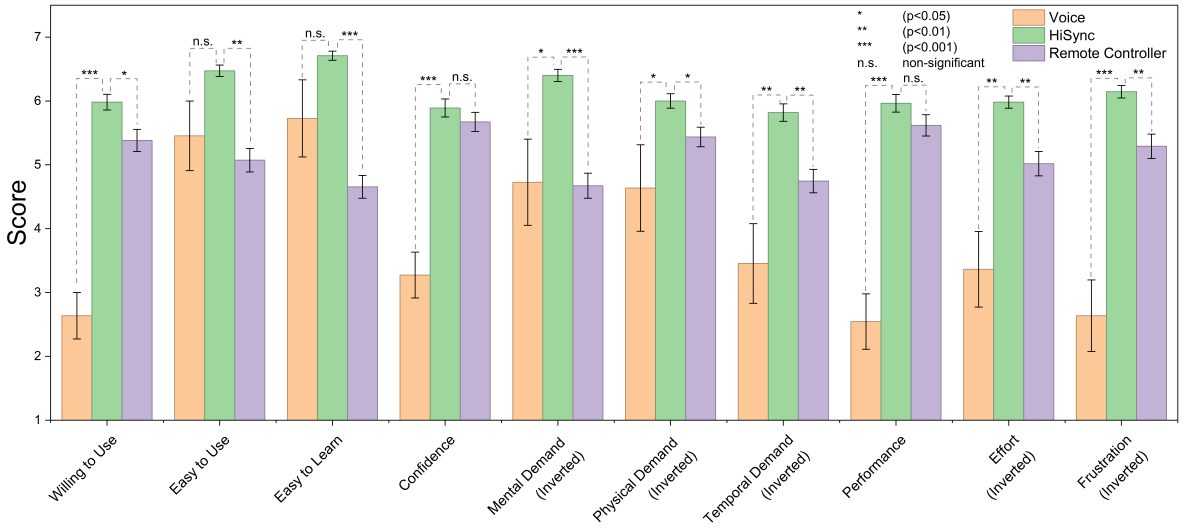


Figure 18: Objective Results of Real Robot Experiment. Mean subjective scores with standard error for the three interaction methods. Statistical significance is marked based on the Wilcoxon Signed-Rank Test. Note: Scores for NASA-TLX subscales are inverted, meaning higher values consistently represent a better user experience across all indicators. Overall, HiSync consistently outperforms both baselines across all metrics, demonstrating superior usability and significantly reduced user workload.

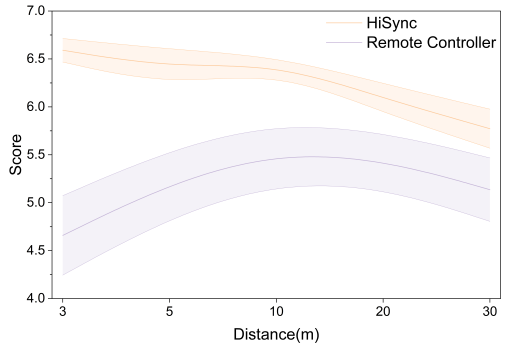


Figure 19: Aggregated subjective scores as a function of distance. This figure illustrates the average of SUS and NASA-TLX results. The higher the score, the better the system. HiSync maintains a significant ($p < 0.0007$) advantage over the Remote Controller. Shaded areas denote standard error.

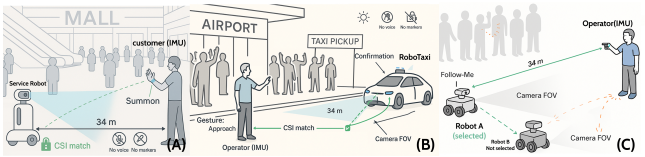


Figure 20: Applications. (A) Service-robot interactions at a mall. (B) Summoning a robotaxi. (C) "Hi and Sync" paradigm.

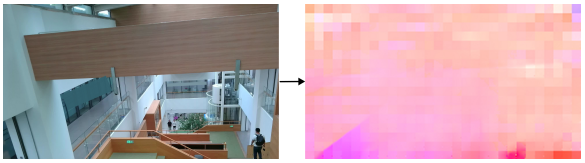


Figure 21: Illustration of VideoFlow Failure Case. At this extreme distance, human movements have become so minute that they are obscured by noise.

Fig. 23. This points to a confidence-contingent, graceful-escalation strategy for HRI systems: *when recognition confidence drops below a calibrated threshold, the system should deliver cues (e.g., a brief LED blink or ring vibration) that explicitly invites the user to repeat the command with a larger or more distinctive motion.* Such graded feedback preserves natural interaction in the default case while reducing misrecognition at long range, without requiring users to learn a new gesture vocabulary or make burdensome movements unless needed.

10.4 Practical Deployment and Efficiency

While our work focuses on the specific challenge of Command Source Identification (CSI), real-world integration also requires the ability of gesture classification, which is already a mature field for single IMU systems [18, 64, 70]. Although HiSync processes continuous streaming by matching motion energy correlation independent of specific gesture classes, running the full visual-inertial fusion continuously is computationally redundant for idle periods. We therefore advocate for a cascaded processing pipeline where HiSync operates as a downstream binding module, triggered only after a valid gesture candidate is detected. In this setup, the wearable IMU operates in a low-power mode using acceleration thresholds or lightweight on-device classifiers to filter background noise and

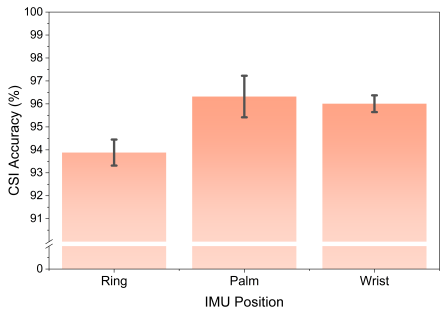


Figure 22: CSI Accuracy vs. IMU Position. While the Palm and Wrist provide slightly stronger motion cues than the Ring, the system achieves high accuracy (> 93%), validating the system's usability across diverse wearable form factors. Error bar denotes standard deviation.

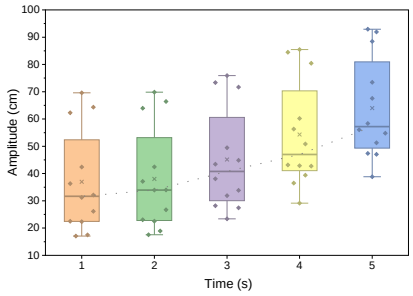


Figure 23: Gesture Amplitude vs. Time. Under non-response conditions, participants gradually amplify their gesture amplitude, suggesting a natural "graceful escalation" strategy for interaction recovery. Error bar shows standard deviation.

establish communicative intent. This strategy leverages the maturity of IMU-based recognition for intent detection while reserving HiSync's spectral capabilities for the distinct task of resolving the command source in multi-user environments, thereby balancing responsiveness with operational duration.

10.5 The Necessity of Visual Binding

Human-robot interaction in long-range scenarios relies heavily on localizing the user who issues a command. Commands such as "Come here" are semantically incomplete without knowing the user's location. In our setting, users convey intent through IMU-based gestures, which do not encode positional information, while user location is captured separately by the robot's onboard cameras. This separation introduces ambiguity in identifying the command source when multiple visual distractors are present. HiSync addresses this challenge by reframing user localization as an intent grounding problem across heterogeneous sensing modalities. By synchronizing motion-based intent signals with visual observations, HiSync enables the robot to disambiguate command sources and attribute commands to a specific user in the environment.

11 LIMITATIONS and FUTURE WORK

Limitations. Our study was conducted in semi-controlled corridors and atriums with two concurrent participants and only occasional passersby; the resulting low crowd density (about 3 people per frame) falls short of the congestion and occlusion found in malls and airports, limiting ecological validity. The participant pool underrepresents older adults and children, constraining generalizability across age groups. Moreover, the current pipeline is sensitive to robot ego-motion and often benefits from the platform pausing before identification, which may restrict responsiveness and practical applicability. Finally, we assumed a setup where only the target user transmits inertial data to the robot. While this setup allows us to validate the system's ability to spatially localize a *known command source*, it does not cover multi-user scenarios where multiple IMU-equipped users simultaneously signal to a shared robot (e.g., in public spaces where a communal robot is controlled by nearby users). Although one could, in principle, run the HiSync procedure independently for each IMU stream, we have not yet validated this extension experimentally.

Future Work. Next steps include leveraging on-board multi-sensor fusion (e.g., robot IMU/odometry/visual-inertial SLAM) to compensate ego-motion and sustain CSI while the platform is moving; extending the approach to capture command-source identification from natural, unconstrained, and small-amplitude gestures rather than a fixed gesture set and even infer from user intentions other than gesture; exploring human-prior-free, pixel-level sync to better handle occlusion and relax assumptions about fixed IMU placement; conducting systematic evaluations in multi-IMU settings and refining the pipeline to mitigate unforeseen failures that emerge during validation; and conducting in-situ deployments in high-density venues (e.g., large shopping malls) to assess performance, latency, and failure modes under real-world crowding.

12 CONCLUSION

We introduced HiSync, an optical-inertial fusion framework for Command Source Identification (CSI) that binds "who is commanding whom" at long range without voice, markers, or user appearance prior. By matching a wearable IMU's motion spectrum to per-candidate flow spectra from a robot-mounted RGB camera, HiSync enables robust hand gesture interactions across embodiments (service robots, drones, quadrupeds) and crowded scenes, expanding interaction range to 34 m. Limitations include low-density test sites, constrained demographics, and sensitivity to robot ego-motion. Future work will (i) fuse on-robot sensors to operate reliably while moving, (ii) report camera-agnostic, distance-aware metrics tied to human pixel footprint, (iii) broaden to natural micro-gestures, and (iv) evaluate in situ in malls and transit hubs. Reframing long-range summoning as cross-modal matching, HiSync advances far-range, many-to-many human-robot interaction.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2024YFB4505500 & 2024YFB4505501, Beijing Key Lab of Networked Multimedia, Institute for Artificial Intelligence, Tsinghua University (THUAI).

References

[1] Fengyan An and Bilong Liu. 2022. Narrowband Active Noise Control Using Decimated Controller for Disturbance with Close Frequencies. *Symmetry* 14, 3 (2022), 607.

[2] Shannon K. T. Bailey and Cheryl I. Johnson. 2020. A Human-Centered Approach to Designing Gestures for Natural User Interfaces. In *Human-Computer Interaction. Multimodal and Natural Interaction*. Lecture Notes in Computer Science, Vol. 12182. Springer Cham, 3–18. doi:10.1007/978-3-030-49062-1_14

[3] Simon Baker, Daniel Scharstein, James P Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. 2011. A database and evaluation methodology for optical flow. *International journal of computer vision* 92, 1 (2011), 1–31.

[4] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. 2010. Proxemic interaction: designing for a proximity and orientation-aware environment. In *ACM International Conference on Interactive Tabletops and Surfaces*. 121–130.

[5] Eran Bamani, Eden Nissinman, Inbar Meir, Lisa Koenigsberg, and Avishai Sintov. 2024. Ultra-range gesture recognition using a web-camera in human–robot interaction. *Engineering Applications of Artificial Intelligence* 132 (2024), 108443.

[6] Chaitanya Bandi and Ulrike Thomas. 2025. Action Recognition via Multi-View Perception Feature Tracking for Human–Robot Interaction. *Robotics* 14, 4 (2025), 53.

[7] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*. 359–370.

[8] Rafael Berral-Soler, Rafael Muñoz-Salinas, Rafael Medina-Carnicer, and Manuel J Marín-Jiménez. 2024. DeepArUco++: Improved detection of square fiducial markers in challenging lighting conditions. *Image and Vision Computing* 152 (2024), 105313.

[9] Shubhang Bhatnagar, Sharath Gopal, Narendra Ahuja, and Liu Ren. 2023. Long-distance gesture recognition using dynamic neural networks. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1307–1312.

[10] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[11] Julia Cambre and Chinmay Kulkarni. 2019. One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–19.

[12] Jessica R Cauchard, Jane L E, Kevin Y Zhai, and James A Landay. 2015. Drone & me: an exploration into natural human-drone interaction. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 361–365.

[13] Xiang Chang, Zihe Chen, Xiaoyan Dong, Yuxin Cai, Tingmin Yan, Haolin Cai, Zherui Zhou, Guyue Zhou, and Jiangtao Gong. 2024. " It Must Be Gesturing Towards Me": Gesture-Based Interaction between Autonomous Vehicles and Pedestrians. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.

[14] Jialin Chen, Armaghan Moemeni, and Praminda Caleb-Solly. 2023. Comparing a graphical user interface, hand gestures and controller in virtual reality for robot teleoperation. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 644–648.

[15] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. 2019. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8351–8361.

[16] Seyyed Saeid Cheshmi, Buyao Lyu, Thomas Lisko, Rajesh Rajamani, Robert A McGovern, and Yogatheesan Varatharajah. 2025. Improving Out-of-distribution Human Activity Recognition via IMU-Video Cross-modal Representation Learning. *arXiv preprint arXiv:2507.13482* (2025).

[17] Torben Cichon and Jürgen Roßmann. 2018. Digital twins: assisting and supporting cooperation in human-robot teams. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 486–491.

[18] Agastasya Dahiya, Dhruv Wadhwa, Rohan Katti, and Luigi G Occhipinti. 2024. Efficient Hand Gesture Recognition using Artificial Intelligence and IMU based Wearable Device. *IEEE Sensors Letters* (2024).

[19] Minh Duc Dang, Samira Pulatova, and Lawrence H Kim. 2025. User-Defined Co-Speech Gesture Design with Swarm Robots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.

[20] Joseph DelPreto and Daniela Rus. 2020. Plug-and-play gesture control using muscle and motion sensors. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 439–448.

[21] Paul Furgale, Joern Rehder, and Roland Siegwart. 2013. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1280–1286.

[22] Michael Fürst, Shriya TP Gupta, René Schuster, Oliver Wasenmüller, and Didier Stricker. 2021. HPERL: 3d human pose estimation from RGB and lidar. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 7321–7327.

[23] Eyal Ginosar and Jessica R Cauchard. 2023. At first light: Expressive lights in support of drone-initiated communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[24] Farid Gulmammadov. 2009. Analysis, modeling and compensation of bias drift in MEMS inertial sensors. In *2009 4th International conference on recent advances in space technologies*. IEEE, 591–596.

[25] Ayush Gupta and Rama Chellappa. 2025. MimicGait: A Model Agnostic approach for Occluded Gait Recognition Using Correlational Knowledge Distillation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 4757–4766.

[26] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[27] Zhe He, Zixuan Wang, Chun Yu, Chengwen Zhang, Xiyuan Shen, and Yuanchun Shi. 2025. WritingRing: Enabling Natural Handwriting Input with a Single IMU Ring. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.

[28] Roberto Henschel, Timo von Marcard, and Bodo Rosenhahn. 2019. Simultaneous identification and tracking of multiple people using video and imus. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.

[29] Yutaka Hiroi and Akinori Ito. 2016. Influence of the height of a robot on comfortableness of verbal interaction. *IAENG International Journal of Computer Science* 43, 4 (2016), 447–455.

[30] Masoumehsadat Hosseini, Tjado Ihmels, Ziqian Chen, Marion Koelle, Heiko Müller, and Susanne Boll. 2023. Towards a consensus gesture set: A survey of mid-air gestures in HCI for maximized agreement across domains. In *proceedings of the 2023 CHI conference on human factors in computing systems*. 1–24.

[31] Damian Hostettler, Simon Mayer, Jan Liam Albert, Kay Erik Jenss, and Christian Hildebrand. 2025. Real-Time Adaptive Industrial Robots: Improving Safety And Comfort In Human-Robot Collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[32] Panjian Huang, Yunjie Peng, Saihui Hou, Chunshui Cao, Xu Liu, Zhiqiang He, and Yongzhen Huang. 2024. Occluded gait recognition with mixture of experts: an action detection perspective. In *European Conference on Computer Vision*. Springer, 380–397.

[33] Krzysztof Jaskot and Artur Babiarz. 2010. The inertial measurement unit for detection of position. *Przeglad Elektrotechniczny* 86, 11A (2010), 323–333.

[34] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. 2023. *Ultralytics YOLO*. https://github.com/ultralytics/ultralytics

[35] Michail Kalaitzakis, Brennan Cain, Sabrina Carroll, Anand Ambrosi, Camden Whitehead, and Nikolaos Vitzilaios. 2021. Fiducial markers for pose estimation: Overview, applications and experimental comparison of the artag, apriltag, aruco and stag markers. *Journal of Intelligent & Robotic Systems* 101, 4 (2021), 71.

[36] Rie Kamikubo, Seita Kayukawa, Yuka Kaniwa, Allan Wang, Hernisa Kacorri, Hironobu Takagi, and Chieko Asakawa. 2025. Beyond Omakase: Designing Shared Control for Navigation Robots with Blind People. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[37] Charles Knapp and Glifford Carter. 2003. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing* 24, 4 (2003), 320–327.

[38] Martin Knoche, Stefan Hörmann, and Gerhard Rigoll. 2021. Susceptibility to image resolution in face recognition and trainings strategies. *arXiv preprint arXiv:2107.03769* (2021).

[39] Ji-Wook Kwon, Hyoujun Lee, Jongdeuk Lee, Na-Hyun Lee, Jong Chan Kim, Taeyoung Uhm, and Young-Ho Choi. 2024. UWB-based human-following system with obstacle and crevasse avoidance for polar-exploration robots. *Applied Sciences* 14, 16 (2024), 6918.

[40] Christine P Lee, Bengisu Cagiltay, and Bilge Mutlu. 2022. The unboxing experience: Exploration and design of initial interactions between children and social robots. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–14.

[41] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. 2023. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5715–5724.

[42] Jan Leusmann, Steeven Villa, Thomas Liang, Chao Wang, Albrecht Schmidt, and Sven Mayer. 2025. An approach to elicit human-understandable robot expressions to support human-robot interaction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[43] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. 2019. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3958–3967.

[44] Yu-Jhe Li, Zhengyi Luo, Xinshuo Weng, and Kris M Kitani. 2020. Learning shape representations for clothing variations in person re-identification. *arXiv preprint arXiv:2003.07340* (2020).

[45] Zhuojun Li, Chun Yu, Yizheng Gu, and Yuanchun Shi. 2023. ResType: Invisible and Adaptive Tablet Keyboard Leveraging Resting Fingers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[46] Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. 2025. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1769–1782.

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

[48] Hanfang Lyu, Xiaoyu Wang, Nandi Zhang, Shuai Ma, Qian Zhu, Yuhan Luo, Fugee Tsung, and Xiaojuan Ma. 2025. Signaling Human Intentions to Service Robots: Understanding the Use of Social Cues during In-Person Conversations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[49] Sebastian OH Madgwick, Andrew JL Harrison, and Ravi Vaidyanathan. 2011. Estimation of IMU and MARG orientation using a gradient descent algorithm. In *2011 IEEE international conference on rehabilitation robotics*. Ieee, 1–7.

[50] James Male and Uriel Martinez-Hernandez. 2021. Recognition of human activity and the state of an assembly task using vision and inertial sensor fusion methods. In *2021 22nd IEEE International Conference on Industrial Technology (ICIT)*, Vol. 1. IEEE, 919–924.

[51] Kayla Matheus, Rebecca Ramnauth, Brian Scassellati, and Nicole Salomons. 2025. Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations. *ACM Transactions on Human-Robot Interaction* 14, 3 (2025), 1–42.

[52] Krista Merry and Pete Bettinger. 2019. Smartphone GPS accuracy study in an urban environment. *PloS one* 14, 7 (2019), e0219890.

[53] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2CLIP: language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 13246–13253.

[54] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. 2018. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In *Asian conference on computer vision*. Springer, 562–576.

[55] Edwin Olson. 2011. AprilTag: A robust and flexible visual fiducial system. In *2011 IEEE international conference on robotics and automation*. IEEE, 3400–3407.

[56] OpenCV. 2013. opencv. https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_frontalface_default.xml.

[57] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. 2021. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12046–12055.

[58] Jeongeun Park, Seungwon Lim, Joonhyung Lee, Sangbeom Park, Minsuk Chang, Youngjae Yu, and Sungjoon Choi. 2023. Clara: classifying and disambiguating user commands for reliable interactive robotic agents. *IEEE Robotics and Automation Letters* 9, 2 (2023), 1059–1066.

[59] 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.

[60] Irene Rae, Leila Takayama, and Bilge Mutlu. 2013. The influence of height in robot-mediated communication. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1–8.

[61] Lala Shakti Swarup Ray, Qingxin Xia, Vitor Fortes Rey, Kaishun Wu, and Paul Lukowicz. 2025. Improving IMU based Human Activty Recognition Using Simulated Multimodal Representations and a MoE Classifier. *Frontiers in Computer Science* 7 (2025), 1569205.

[62] Ángel-Gabriel Salinas-Martínez, Joaquín Cunillé-Rodríguez, Elías Aquino-López, and Angel-Iván García-Moreno. 2024. Multimodal Human–Robot Interaction Using Gestures and Speech: A Case Study for Printed Circuit Board Manufacturing. *Journal of Manufacturing and Materials Processing* 8, 6 (2024), 274.

[63] Davide Scaramuzza and Zichao Zhang. 2019. Visual-inertial odometry of aerial robots. *arXiv preprint arXiv:1906.03289* (2019).

[64] Xiyuan Shen, Chun Yu, Xutong Wang, Chen Liang, Haozhan Chen, and Yuanchun Shi. 2024. MouseRing: Always-available Touchpad Interaction with IMU Rings. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[65] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. 2023. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12469–12480.

[66] Xi Sun, Xinshuo Weng, and Kris Kitani. 2020. When we first met: Visual-inertial person localization for co-robot rendezvous. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10408–10415.

[67] Vladimir Suvorkin, Miquel Garcia-Fernandez, Guillermo González-Casado, Mowen Li, and Adria Rovira-Garcia. 2024. Assessment of noise of mems imu sensors of different grades for gnss/imu navigation. *Sensors* 24, 6 (2024), 1953.

[68] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. 2022. Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–33.

[69] Daria Trinitatova and Dzmitry Tsetserukou. 2023. Study of the effectiveness of a wearable haptic interface with cutaneous and vibrotactile feedback for VR-based teleoperation. *IEEE Transactions on Haptics* 16, 4 (2023), 463–469.

[70] Radu-Daniel Vatavu, Lisa Anthony, and Jacob O Wobbrock. 2018. $ Q: a super-quick, articulation-invariant stroke-gesture recognizer for low-resource devices. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[71] Ella Velner, Paul PG Boersma, and Maartje MA De Graaf. 2020. Intonation in robot speech: Does it work the same as with people?. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 569–578.

[72] Ana M Villanueva, Ziyi Liu, Zhengzhe Zhu, Xin Du, Joey Huang, Kylie A Peppler, and Karthik Ramani. 2021. Robotar: An augmented reality compatible teleconsulting robotics toolkit for augmented makerspace experiences. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[73] Daniel Vogel and Ravin Balakrishnan. 2004. Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*. 137–146.

[74] John H Wall and David M Bevly. 2006. Characterization of inertial sensor measurements for navigation performance analysis. In *Proceedings of the 19th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2006)*. 2678–2685.

[75] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained person perception using WiFi. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5452–5461.

[76] Tian Wang, Pai Zheng, Shufei Li, and Lihui Wang. 2024. Multimodal human–robot interaction for human-centric smart manufacturing: a survey. *Advanced Intelligent Systems* 6, 3 (2024), 2300359.

[77] Xian Wang, Luyao Shen, and Lik-Hang Lee. 2025. A Systematic Review of XR-Enabled Remote Human-Robot Interaction Systems. *Comput. Surveys* 57, 11 (2025), 1–37.

[78] Zeyu Wang, Ruotong Yu, Xiangyang Wang, Jiexin Ding, Jiankai Tang, Jun Fang, Zhe He, Zhuojun Li, Tobias Röddiger, Weiye Xu, et al. 2025. Computing with Smart Rings: A Systematic Literature Review. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–54.

[79] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 79–88.

[80] Oliver J Woodman. 2007. *An introduction to inertial navigation*. Technical Report. University of Cambridge, Computer Laboratory.

[81] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. 2024. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 25456–25467.

[82] Yu Xianjia, Li Qingqing, Jorge Peña Queralta, Jukka Heikkonen, and Tomi Westerlund. 2021. Applications of UWB networks and positioning to autonomous robots and industrial systems. In *2021 10th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, 1–6.

[83] Ruowei Xiao, Aleksi Vianto, Asif Shaikh, Juho Hamari, Johanna Virkki, et al. 2022. Exploring the application of rfid for designing augmented virtual reality experience. *IEEE Access* 10 (2022), 96840–96851.

[84] Guangsheng Xu, Guoyi Zhang, Lejia Ye, Shuwei Gan, Xiaohu Zhang, and Xia Yang. 2025. Optimizing local-global dependencies for accurate 3d human pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology* (2025).

[85] Naoya Yoshimura, Hironori Yoshida, Fabrice Matulic, and Takeo Igarashi. 2019. Extending discrete verbal commands with continuous speech for flexible robot control. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[86] Jargalbaatar Yura, Bat-Erdene Byambasuren, and Donghan Kim. 2025. Multimodal, multi-step human–robot interaction method for natural interaction of service robots. *Intelligent Service Robotics* (2025), 1–13.

[87] Xinyi Zhang, Qiqi Bao, Qinpeng Cui, Wenming Yang, and Qingmin Liao. 2025. Pose magic: Efficient and temporally consistent human pose estimation with a hybrid mamba-gcn network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10248–10256.

[88] Yan Zhang, Tharaka Sachintha Ratnayake, Cherie Sew, Jarrod Knibbe, Jorge Goncalves, and Wafa Johal. 2025. Can you pass that tool?: Implications of indirect speech in physical human-robot collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[89] Jingxing Zhong, Xingyou Fang, Zukang Yang, Zhen Tian, and Chuangqi Li. 2025. Skybound Magic: Enabling Body-Only Drone Piloting Through a Lightweight Vision–Pose Interaction Framework. *International Journal of Human–Computer Interaction* (2025), 1–31.

[90] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3702–3712.

[91] Ran Zhou, Zachary Schwemler, Akshay Baweja, Harpreet Sareen, Casey Lee Hunt, and Daniel Leithinger. 2023. Tactorbots: a haptic design toolkit for out-of-lab exploration of emotional robotic touch. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–19.

A Quantitative Results of Formative Study of Example User

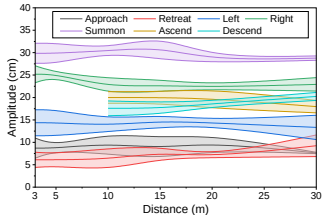


Figure 24: Gesture Amplitude Consistency across Distances. Data from User 1 illustrates a distance-invariant strategy, where the amplitude of gestures does not significantly correlate with the interaction distance. Error bar denotes STD over different robots.

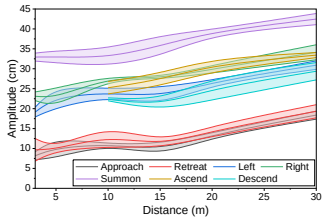


Figure 25: Gesture Amplitude Scaling across Distances. Data from User 2 illustrates a distance-adaptive strategy, where the amplitude of gestures positively correlates with the interaction distance. Error bar denotes STD over different robots.

B Spectral Feature.

B.1 Detailed Spectral Feature Definition

$$\boldsymbol{\phi}(x) = [\text{PSD}; \text{Feat}] = \left[\text{PSD},\ p,\ f,\ \kappa,\ H,\ \Delta f,\ \text{SNR},\ P_{\text{avg}}\right]. \quad (7)$$

- **Peak Height (p):** The maximum magnitude in the power spectral density. This metric captures the intensity of the dominant rhythmic component, facilitating the matching of motion strength between the visual and inertial modalities.
- **Peak Frequency (f):** The frequency at which the spectral peak occurs. It indicates the fundamental tempo of the gesture, serving as the primary cue for synchronizing the speed of the user's hand movement across different sensors.
- **Clarity (κ):** Defined as the ratio of the peak magnitude to the average spectral power. This quantifies the periodicity of the signal, effectively distinguishing deliberate, rhythmic command gestures from irregular or non-periodic background noise.
- **Spectral Entropy (H):** A measure of the concentration of spectral energy distribution. Lower entropy corresponds to a sharper peak, indicating a "purer" or more focused single-frequency movement, which helps in filtering out complex or chaotic motion artifacts.
- **Frequency Spacing (Δf):** The frequency difference between the two most significant peaks. This characterizes the harmonic structure of the motion, acting as a kinematic signature to differentiate between users who may be gesturing at similar fundamental speeds but with different motion styles.
- **Signal-to-Noise Ratio (SNR):** The ratio of the in-band peak energy to the estimated residual noise floor. This evaluates the signal quality, ensuring that the system prioritizes clean, high-confidence motion signals over sensor noise or environmental interference.
- **Band-Average Power (P_{avg}):** The mean power calculated across the passband. This represents the overall energy expenditure of the movement, aiding in the discrimination between high-energy (e.g., large arm waves) and low-energy (e.g., subtle wrist flicks) gestures.

B.2 Spectral Feature Ablation Results

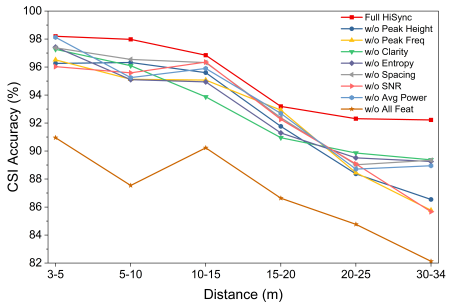


Figure 26: Spectral Feature Ablation Results

As shown in the Fig. 26, the experimental results demonstrate the effectiveness of the Spectral Feature design, with a particularly notable performance improvement observed at longer distances.