Invisibility Cloak: Personalized Smartwatch-Guided Camera Obfuscation

Xue Wang
University of California, Los Angeles
Los Angeles, CA, USA
xw526@ucla.edu

Yang Zhang
University of California, Los Angeles
Los Angeles, CA, USA
yangzhang@ucla.edu

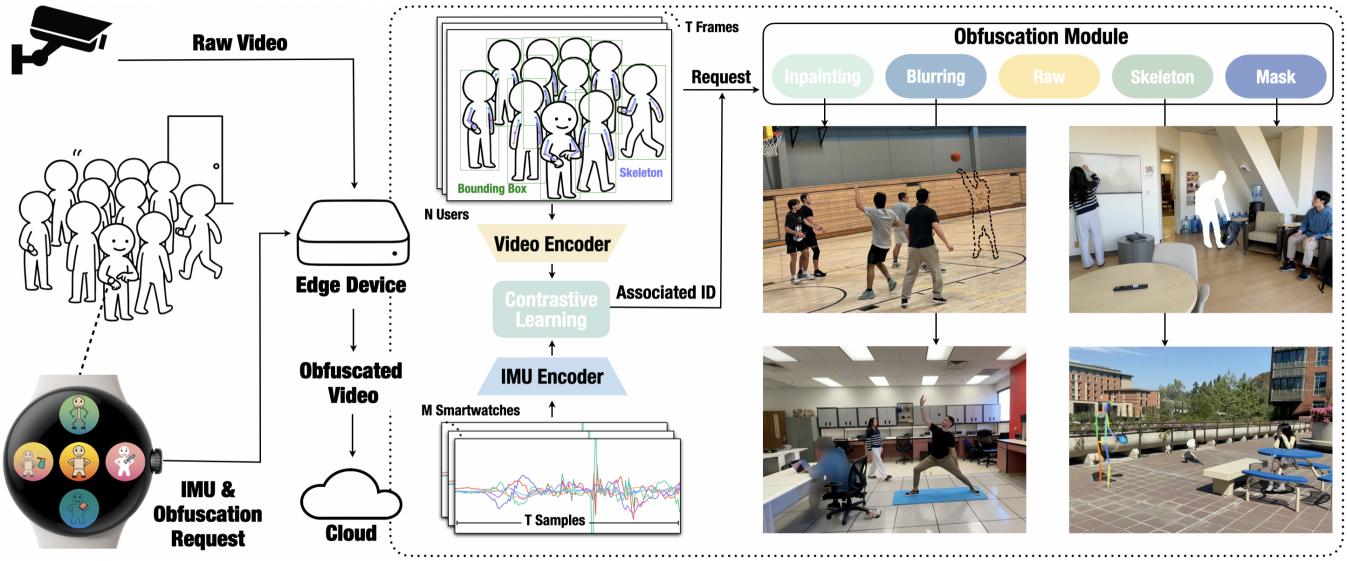

Figure 1: Our system consists of a smartwatch app that streams IMU signals and obfuscation requests to an edge device, which runs our user-device association algorithm. This system operates under the following standard security assumptions: all wearable devices are fully compatible with the local camera system, and data on the edge device is secure from unauthorized access and cyber-attacks prior to transmission to the cloud.

ABSTRACT

Cameras are in their golden age due to recent advances in visual AI techniques that significantly extend the applicability and accuracy of vision-based applications including healthcare, entertainment, and security. In public environments, individuals usually have different and changing privacy preferences against their visual information being shared with other entities. To accommodate these varying user needs for visual privacy, we created *Invisibility Cloak*, a camera obfuscation technique leveraging inertial signals collected from smartwatches to guide an edge device to remove visual information from camera recordings before they are streamed out for cloud-based inferences. Specifically, a smartwatch user can select an obfuscation level that fits their privacy preference in that context and cameras in the environment will use smartwatch signals to identify that user and remove visual information associated with the user. On the conceptual level, our system demonstrates a privacy design rationale which removes information to be shared with a broader internet infrastructure (i.e., cloud) by providing more information to a trusted local camera system (i.e., camera sensor + edge computing device). We developed a custom data-association pipeline and collected data from real-world configurations. Evaluation of our pipeline indicates a user identification accuracy of 95.48% among 10 individuals when our system is provided with only 2 seconds of data.

UIST '25, September 28-October 1, 2025, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/2025/09.
https://doi.org/10.1145/3746059.3747601

CCS CONCEPTS

• **Human-centered computing** → **Mobile devices**; **Interactive systems and tools**.

KEYWORDS

Camera Obfuscation; Smartwatch; User Identification; Privacy

ACM Reference Format:
Xue Wang and Yang Zhang. 2025. Invisibility Cloak: Personalized Smartwatch-Guided Camera Obfuscation. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3746059.3747601

1 INTRODUCTION

Vision AI is becoming an increasingly practical approach to address the workforce shortage by freeing human operators for safer and more creative tasks. Recent advances in large language models (LLM) have extended the applicability as well as lowered the engineering efforts to apply vision AI in a wide array of use scenarios – safety, security, healthcare, entertainment, education, and beyond. We have begun to see a huge wave of vision systems to be deployed, especially in public environments, for their readily available use cases. However, user privacy in the age of vision AI has been a critical problem that impedes AI from being even more useful, due to the lack of tools to support user privacy in vision systems.

People have different privacy preferences as vision AI prevails [66]. One might need a camera to log their activity, recommending the best time to inject insulin, while others do not want their outfit with an irritating propaganda slogan to be possibly leaked and posted on social media. Accommodating various user privacy preferences is difficult. As contexts change, the same user may have varying privacy preferences, making it even more challenging to accommodate personalized privacy configurations. For a long time, camera utilization has been approached in a binary manner — either deploying cameras in vision applications where privacy concerns are minimal or banning them entirely in scenarios where privacy is a major concern.

To address the tension between utility and privacy in camera-enabled environments, recent research has proposed obfuscation techniques that allow cameras to automatically identify and remove privacy-sensitive information from their field of view. For instance, selective obfuscation methods leverage thermal signals from the human body to detect and redact sensitive pixels before any data is transmitted beyond the local device [24, 34]. These systems aim to preserve application-relevant information while eliminating potentially identifying content. Such approaches align with the principle of obscurity — a fundamental privacy mechanism that makes certain information less visible or accessible to unintended audiences. In a world saturated with digital data, obscurity enables individuals to participate in public or digital spaces while minimizing the risk of being tracked, profiled, or exposed. It acts as a subtle cloak, offering protection without requiring users to withdraw entirely. However, maintaining obscurity is increasingly difficult in the face of pervasive IoT systems and AI-driven analytics. Existing obfuscation techniques remain limited in their ability to support personalized privacy configurations that adapt to each individual's context, highlighting the need for user-centric, dynamic approaches to visual privacy.

To support finer-grained support for privacy preference down to an individual level, we propose *Invisibility Cloak*, a visual-inertial system using IMU signals from smartwatches, a popular wearable device, to allow cameras to identify and obfuscate users according to their privacy preferences. Our system acts as a cloak, allowing people to navigate these spaces with a reduced risk of their actions being monitored, analyzed, or exploited. Individual-level privacy accommodation in shared visual space helps individuals maintain privacy without completely withdrawing from digital or public spaces. And thus our system allows individuals to enjoy the benefits of digital interactions without fully exposing their personal data, supporting a level of anonymity that can protect users from identity theft, tracking, or unwanted marketing.

Specifically, smartwatches in our system yield arm movements of their users which oftentimes exhibit uniqueness as users can statically pose their body differently (e.g., standing with both hands down, vs. sitting with arm resting on a table) or engage in different activities (e.g., typing vs. drinking) that lead to different kinetic signatures. Our system identifies these unique biometrics for user identification, with a custom matching pipeline based on cross-modal contrastive learning. We conducted a series of evaluations, including synthesized multi-user data consisting of individually collected data from 10 users scripted with various daily activities at *In-Lab* and *In-the-Wild* locations. We also conducted a real-time multi-user evaluation. Both evaluations indicated promising results. For synthesized data, the average precisions of three evaluation configurations (i.e., within-user, cross-user, and cross-scene), including various activities and locations, exceed 94%, and no specific precision falls below 93%. For real-time multi-user evaluation, average precisions exceed 89% across all configurations, including unseen camera positions and activities. We also conducted investigations on window length, handedness, camera position, seen vs. unseen activities, usability study, and edge computing deployment to drive further insights into our proposed methodology.

Our contributions are as follows:

- Implementation of a novel interaction scenario where smartwatches serve as privacy safeguards, guiding camera obfuscation based on individual privacy preferences.
- New user-device association pipeline based on contrastive learning, which demonstrated superior accuracy to the state of the art.
- Evaluation of system using a data synthesis approach in concert with live multi-user data collection, yielding insightful knowledge about current and future works.

2 RELATED WORK

2.1 Video Obfuscation Techniques

Privacy in video analytics can be protected using obfuscation techniques applied before, during, or after recording. These methods can prevent sensitive data exposure while preserving content utility. Privacy risks can be mitigated at the physical level by preventing unauthorized recording, such as using mechanical lids that automatically/manually cover the camera [49, 53, 61] or by applying optical masking techniques on the sensor level [67]. Several techniques obfuscate sensitive information during recording. For example, [71] proposed a novel Privacy-sensitive Objects Pixelation (PsOP) framework that can automatically perform personal privacy filtering during live video streaming. The specified or sensitive information is directly removed before storage or transmission [51, 59] or only recording the designated target [27]. In wearables, privacy-aware eye tracking [52] uses differential privacy to prevent re-identification while maintaining functional gaze data.

Closer to our work, prior works investigated various post-capture obfuscation techniques. Traditional blurring and pixelation [16, 32, 47] reduce privacy risks but degrade usability. Selective obfuscation [2] allows customizable filtering of sensitive objects, but struggles with dynamic contexts. To enable personalized privacy, Cardea

[50] employs context-aware privacy profiles that adjust in real time, with gesture-based overrides. In IoT environments, [15] enables face anonymization in video streams for protecting personal information. In online conferencing, ZoomP3 [55] protects privacy-sensitive participants while allowing recordings to be shared. AI-driven tools [65] help blind users obfuscate private content in photos. Finally, PrivacyLens [24] integrates RGB and thermal imaging to remove personally identifiable information (PII) before video leaves the device, ensuring privacy without post-processing artifacts.

2.2 Wearable Enabled Identification and Obfuscation

Additionally, the sensing scheme of our work intersects literature leveraging wearables to enable user identification. Wearable devices offer a promising platform for privacy-aware user identification by sensing unique physiological and behavioral traits.

Motion-based approaches, such as Nod to Auth [60] and HCR-Auth [20], authenticate users based on biometric data gathered through head gestures via IMU sensors, offering intuitive interactions in VR/AR environments. Electrical sensing methods, like those proposed by Cornelius et al. [12], exploit body impedance to capture internal biometric features without relying on visual input. Ear-based techniques such as EarEcho [19] and Voice In Ear [18] utilize the anatomical structure of the ear and body-conducted vocal vibrations for in-ear or bone-conduction authentication. Gait-based approaches [3] using ultra-wideband wearables enable passive identification by measuring inter-device body distances, requiring no explicit user input.

Recent work also explores how wearables support sensor obfuscation and privacy protection. Blinder [63] proposes a federated learning-based approach for sensor data anonymization, using variational autoencoders and discriminators to obscure private attributes while retaining public utility. Similarly, [64] introduces diffusion-based obfuscation frameworks by adding synthetic sensor data to the original one. In this case, the useful data is reserved and the sensitive information is obscured. Moore et al. [38] proposed an approach that anonymizes sensitive information captured by wearable cameras to enable fall detection while mitigating ethical and privacy concerns. These diverse strategies reflect the growing emphasis on unobtrusive, user-centric, learning-based, and context-aware mechanisms that dynamically balance privacy and utility.

2.3 Data Association Across Modalities

Many approaches leverage vision tracking, IMU, and wireless signals (e.g., WiFi) to facilitate fast and accurate localization of users, with trade-offs in accuracy, availability, and ease of deployment. One of the most common approaches is to integrate vision tracking with wireless signals. ViTag [9] and Vi-Fi [33] link bounding box sequences from vision trackers with IMU and Wi-Fi FTM data. RFCam [11] improves robustness by fusing Wi-Fi CSI with video analytics to match mobile devices to users in video footage. Similarly, EyeFi [17] aligns Wi-Fi motion trajectories with camera data, enabling rapid, non-intrusive identification. IMU2CLIP [37] further enhances user recognition by aligning IMU motion data with egocentric video and text narrations through contrastive learning. Who Goes There [35] identifies users by matching the silhouette video clips with accelerometers in a multi-person environment.

Our data-association approach is related to prior motion-based user-device association techniques. IDIoT [6] identifies wearable devices attached to different parts of the user's body. It associated camera-based pose data with IMU signals from wearables on limbs such as the arms and legs. Their algorithm, based on pairwise alignment optimization, was evaluated using a public dataset and demonstrated promising results. The Martini Synch [28] presents a pairing mechanism also using IMU signals when two devices perform the same movements. Closest to our work are systems that identify users using their IMU signals under a camera. Tong et al. [58] propose a multi-camera user identification system in which smartphones are worn on users' chests as wearable IMU devices. Henschel et al. [22] present a system that uses visual-inertial data to match and track multiple users in dynamic outdoor environments (e.g., during soccer games or cross-walking), with the IMU sensor attached at the users' hip height. Sun et al. [54] introduce a visual-inertial fusion method to identify the single user among multiple users who is holding a smartphone embedded with IMU sensors. These systems assume that users follow distinct trajectories and have been primarily evaluated in outdoor environments where users have enough space for separate and continuous movement. They may struggle in indoor scenarios where movements are more free-form, subtle, or similar among users. In contrast, our system uniquely leverages a smartwatch for user identification, achieving superior performance through a novel learning-based data association pipeline despite nuanced user movements. To the best of our knowledge, no prior work in data association has leveraged smartwatches for camera obfuscation or investigated their utility and trade-offs across real-world scenarios, highlighting the novelty and contribution of our system.

3 SYSTEM DESIGN

Our system consists of two stages. The first stage is user reidentification in shared environments, which establishes the association between users and their devices while also determining which user sent which obfuscation requests. The second stage is personalized video obfuscation, which introduces hierarchical obfuscation options, allowing users to customize their privacy settings based on their preferences and specific situations. Both stages happen on an edge AI device located near the camera, before the processed data is streamed out of the user's proximity or stored for longer terms. These two stages are implemented through three parallel threads: one receives IMU signals from smartwatches, the second listens for obfuscation requests from individuals, and the third computes user-device associations and applies video obfuscation based on the processed results and received requests. Further details of these threads can be found in Section 4. Below we list our design goals which have been carefully considered in the selection of signals to utilize and devices that generate these signals when developing our system:

- **D1:** our system should apply to the most common vision devices and use commodity devices for scalability.
- **D2:** our system should differentiate nuanced differences in user motions, allowing swift and accurate user re-identification.

- **D3:** personalization of obfuscation should be accommodated by our system providing users with various levels for camera obfuscations.
- **D4:** obfuscation level should be easily adjustable by the user to quickly adapt privacy protection across user contexts.

3.1 User Re-Identification in Shared Environments

Various modalities have been explored for associating individuals with their corresponding data in shared environments. Some previous work [9, 33, 44] utilizes depth cameras, which provide 3D information and can significantly improve user localization. However, most surveillance infrastructure only uses RGB cameras, and depth information is often difficult to obtain. Other studies leverage WiFi signals for identity association and localization [1, 17, 29, 46]. These systems, however, typically require specific hardware setups, such as WiFi signal receivers installed in the environment. Moreover, indoor spaces with complex layouts or furniture can introduce multipath effects in WiFi signals, which degrade association accuracy. GPS has also been used to associate devices with individuals [36], but it struggles to provide high resolution in small or crowded areas. Bluetooth, RF, and audio-based approaches [25, 40–42] offer finer granularity than GPS but are limited in noisy or cluttered environments.

Given these limitations, we adopt a camera-IMU association strategy, leveraging inertial signals from wearables, particularly smartwatches, to support user identification and localization. As demonstrated in prior work [10, 21, 22, 57], IMUs provide a reliable source of movement data that naturally aligns with visual observations. Unlike external signal-based systems, IMUs are the most common sensor on smartwatches, an increasingly popular and ubiquitous commodity device (**D1**). These IMUs are physically worn on the user's wrist, capturing motion signatures that precisely correspond to individual activities and body dynamics, even for those with minute movements. This tight coupling between the device and the user ensures a higher-fidelity identity association signal (**D2**).

3.2 Personalized Data Obfuscation

In shared environments such as offices, campuses, or public buildings, users often hold varying expectations about how their data should be processed and protected. Unlike in isolated settings, shared spaces require negotiation and compromise to ensure fairness among all users [70]. Although some systems, such as [39], provide multiple levels of video obfuscation, they lack the ability to identify users within shared environments, limiting their effectiveness in enforcing user-specific privacy preferences. To address this limitation, surveillance systems must support flexible, individualized privacy configurations that can accommodate diverse user needs without introducing conflicts among co-located individuals [14]. According to Section 3.1, once users and devices are correctly associated, our system should accurately identify and localize individuals who initiate requests. Building on this foundation, we propose a hierarchical video obfuscation framework that allows users to customize how their visual presence is represented, while preserving Vision AI's utility for situational awareness and analytics for other users (**D3**).

In our proposed system, we define five hierarchical levels of video obfuscation, enabling users to choose a privacy configuration that aligns with their needs and environmental context. Figure 1 (right) illustrates the visual effects of each obfuscation level, demonstrating how the system dynamically adjusts the privacy granularity based on user preference. These levels balance the trade-off between data utility and personal privacy:

Raw The video is captured, stored, and transmitted without any modification. This level is suitable for fully consenting environments where high-fidelity signals are necessary.

Masking Whole body regions are detected using YOLO-based segmentation and overlaid with solid masks. This conceals identity-revealing features while maintaining spatial and movement context.

Blurring Whole body regions are blurred to obscure identifiable characteristics while preserving motion and activity patterns as well as some color information from clothes.

Inpainting The individual is removed from the video frame and the background is synthetically filled in. This provides the highest level of privacy by eliminating visual indicators of a user's presence.

Skeleton Overlay After inpainting, a skeletal pose representation is rendered in place of the user. This enables functional activity monitoring without exposing identifiable appearance features.

3.3 Privacy–Utility Trade-offs and Operational Validity

User privacy needs are often dynamic, changing with context, time, and location. In this case, individuals may have different requirements for privacy protection depending on the trade-offs between privacy and vision-based AI functionality. For instance, a user might pick *Skeleton* in a yoga class where their body postures need to be logged for educational purposes. In contrast, the same user may prefer *Inpainting* for stronger identity anonymity when entering a grocery store using vision AI to track inventories. *Invisibility Cloak* is built on the principle that personal privacy protection and AI-driven functionality can coexist, allowing vision-based AI to maintain core functionality while respecting individual privacy preferences. For instance, skeleton-level obfuscation only keeps essential structural information required for applications like fall detection or activity monitoring, while removing sensitive features like nudity or facial expressions.

To support real-time responsive and personalized control, we decided to accommodate users with an intuitive smartwatch interface (Figure 1, left), designed and implemented on the Google Pixel Watch 2. This smartwatch serves a dual purpose: it transmits IMU signals for identification and localization, and it allows users to send personalized privacy requests to nearby IoT devices. Once users send their desired obfuscation request, the edge device applies the selected obfuscation to their visual representation in the video in real-time (**D4**). Note that each individual is segmented and processed independently, enabling personalized obfuscation even when multiple users appear in the same scene.

Each obfuscation level inherently involves trade-offs between utility and privacy. For example, while *Raw* data provides the highest fidelity for behavior analysis, it exposes all visual identifiers.

On the other hand, *Inpainting* and *Skeleton* provide strong privacy guarantees for individuals while still retaining adequate contextual information to support many vision-based AI tasks. *Invisibility Cloak* thus offers a flexible mechanism for managing visual information disclosure. By addressing both individual and contextual privacy needs, it facilitates broader adoption of obfuscation techniques within existing surveillance infrastructures, supporting both functionality and privacy in real-world deployments.

4 SYSTEM IMPLEMENTATION

4.1 Self-Supervised Cross-Modality Contrastive Learning

In the previous Section 3.1, we determined to use IMU recordings and RGB video streams to perform user-device associations. However, these two modalities originate from fundamentally different sensory inputs — one capturing motion through inertial measurements and the other through visual observations. The disparity in data representation poses a challenge in aligning them within a shared feature space. To address this, we introduce a self-supervised cross-modality contrastive learning framework to facilitate the association. Note that contrastive learning has emerged as a powerful approach for aligning unlabeled data from different modalities [13, 37, 45], and has been first utilized for user-device association through our system. The details of the feature encoders, contrastive loss function, and the cross-modal association mechanism are discussed subsequently.

4.1.1 Feature Encoding for IMU and RGB Sequences. As a first step, we employ feature encoders to extract representations for each modality. The encoder structures for each modality are shown in Figure 2. For IMU data, we utilize linear accelerometer and gyroscope recordings as input to capture the user's motion patterns. For RGB streams (i.e., video), we first apply the state-of-the-art YOLO11x [26] to extract bounding boxes and body keypoints, specifically focusing on the wrists, elbows, and shoulders. These bounding boxes, along with the corresponding arm keypoints, are then processed by the video encoder, which transforms the spatial pose information into feature representations.

IMU Encoder In *Invisibility Cloak*, the IMU input contains the linear acceleration and gyroscope data recorded via smartwatches. The IMU Encoder consists of three main components: group normalization, convolutional blocks, and a bidirectional GRU network. Figure 2 (left) illustrates the IMU Encoder architecture. We set the number of groups in group normalization to 2, corresponding to the two types of signals—linear acceleration and gyroscope. Each convolutional block applies a 2D convolution with a kernel size of 3, followed by batch normalization, a ReLU activation function, and dropout for regularization. The extracted spatial features are then passed through a bidirectional GRU to model temporal dependencies.

Video Encoder The video encoder extracts movement-based features from RGB video frames by detecting individuals present. Each frame in the video stream is processed using a YOLO11x pose detection model to identify keypoints corresponding to body joints and the whole body bounding boxes. Since the smartwatch is worn on the wrist, keypoints from other parts of the body (e.g., legs or head) may introduce confusion. Therefore, we specifically focus on tracking the wrist, elbow, and shoulder joints, as these are most aligned with the smartwatch's movement.

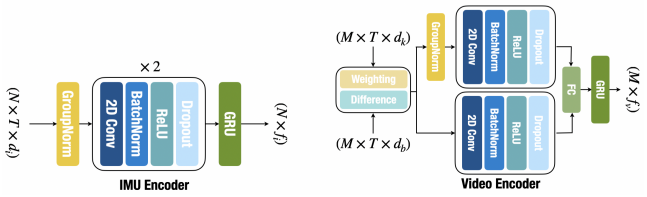

Figure 2: Feature encoder architectures for IMU and video modalities. T represents the number of frames in each tracklet. N denotes the number of individuals transmitting IMU signals to the edge device, and M denotes the number of individuals detected in the last frame of the current tracklet. The dimensions of the input vectors are as follows: d_i for IMU signals, d_k for arm keypoints, and d_b for bounding boxes. The output feature dimensionalities are f_i for the IMU encoder and f_v for the video encoder.

To capture dynamic motion information from each individual, the video encoder first computes motion vectors by measuring how the positions of joints change between consecutive frames. Since pose estimation can be susceptible to noise from shaking, occlusion, or misdetections, some keypoints are not as reliable. To address this, we designed a confidence-weighting mechanism that prioritizes keypoints with higher confidence scores to mitigate this issue. The same strategy is also applied to bounding boxes, ensuring that only trustworthy detections influence the motion tracking. After this processing, the concatenated motion data calculated with keypoints and bounding boxes is fed as input for further processing.

Following that, we adopt a video encoder architecture similar to the one used for IMU signals but with separate branches for bounding boxes and keypoints. The architecture of the video encoder is displayed in Figure 2 (right). For the keypoints, we set the number of groups in group normalization to 3, corresponding to the three types of keypoints (wrist, elbow, and shoulder). Group normalization is not applied to the bounding box branch, as the bounding box features are treated as a single, unified entity rather than grouped components. Each branch then passes through its own set of convolutional blocks. The extracted features from the keypoints and bounding boxes are fused using a fully connected layer before being fed into GRU layers for temporal modeling. The final video embeddings are a compact representation for each detected person in the video, summarizing their visual and motion characteristics, and these embeddings are then used for association with IMU embeddings.

4.1.2 Contrastive Loss. Data association requires aligning features encoded from different modalities within a shared latent space. To achieve this, we adopt a linear projection strategy inspired by CLIP [37] that maps both IMU and video features onto a unit hypersphere, where the cosine similarity can be directly computed for efficient association. As illustrated in Figure 3, we implement an IMU projection layer and a video projection layer that transform the IMU and video embeddings into aligned representations. Each projected

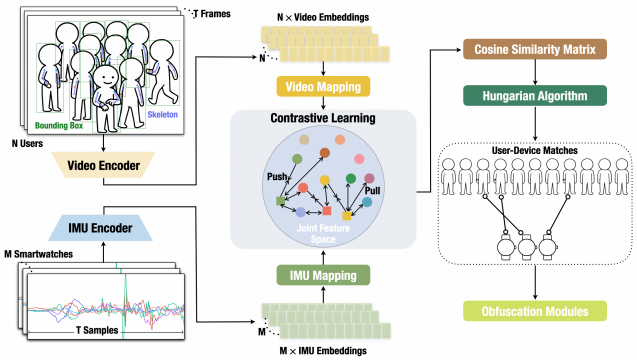

Figure 3: User–device association model architecture.

IMU-video feature pair is labeled as either positive (i.e., matched IMU and video signals from the same user) or negative (unmatched signals). To train the model, we apply contrastive learning: positive pairs are drawn from aligned IMU and pose embeddings, while all other combinations serve as negative samples. During the model training, we employ the symmetric InfoNCE loss [43] as the objective function. It encourages the model to learn robust cross-modal associations by pushing matched IMU and video embeddings closer together while pulling apart unmatched ones.

4.2 User-Device Association via Bipartite Graph

Once the cross-modal representations from the IMU and video encoders are aligned in the joint latent space, we establish associations between them using a bipartite graph matching approach.

Let N be the number of IMU streams and M be the number of video tracklets detected within the current time window. Of note that N may not necessarily equal M, as some individuals may not be transmitting IMU, or users who are transmitting data may be temporarily out of the camera's field of view. We first compute the cosine similarity matrix $S \in \mathbb{R}^{N \times M}$ between each IMU embedding and video embedding, where each element in S quantifies how well each IMU embedding matches with each video embedding. The association problem is then formulated as a bipartite graph matching problem [56]. In this graph $G = \{I, V, E\}$, the nodes I correspond to the IMU embeddings and the nodes V represent video embeddings, with edge E weights given by the similarity scores. Our objective is to find an optimal one-to-one matching that maximizes the total similarity across all paired nodes. This matching can be efficiently solved using the Hungarian algorithm (also known as the Kuhn-Munkres algorithm) [30]. The algorithm finds the set of pairs that maximizes the overall similarity score as the sum of the similarity scores of all pairs in the matching.

4.3 Video Obfuscations

As detailed in the previous section, our system supports 5 levels of video obfuscation options—*Raw*, *Blurring*, *Masking*, *Inpainting with Skeleton Overlay*, and *Inpainting only*—to accommodate diverse privacy preferences for people in different environments. In this section, we describe how these obfuscation approaches are implemented to enable real-time processing in edge devices in detail.

For lightweight and efficient obfuscation, *Blurring* and *Masking*, we leveraged traditional computer vision techniques that operate directly on the segmented human body areas obtained from the YOLO11x model. For scenarios requiring more advanced privacy protection, such as *Inpainting* and *Skeleton Overlay*, we employ deep learning-based techniques that provide better context preservation.

Masking In the masking approach, a solid occlusion color (e.g., white or a user-defined hue) is applied to the segmented region. This is achieved by replacing the pixel values within the mask with the occlusion color, thereby concealing the user's identity (including face, body shapes, and clothing information) while preserving the overall scene context. The operation is implemented via OpenCV [7] built-in functions to ensure minimal latency.

Blurring Blurring is implemented by applying a box filter over the segmented region. Typically, a kernel size of 101×101 is used to effectively smooth out fine details while retaining coarse motion and structural information. This filter convolves the masked area with a uniform kernel, ensuring that identifying features are obscured while the motion patterns are still preserved.

Inpainting The system runs a lightweight generative model called *Mobile Inpainting GAN* (MI-GAN) [48] when a user requests inpainting-based options. This model was pre-trained on the Places2 dataset [69] and is optimized for real-time, high-quality, and computationally efficient image inpainting. In the first stage, the segmentations of target users are determined for each frame. These segmentations serve as masks and are fed into the model together with the original frames. The model then erases the target users and reconstructs the background that was previously obscured by them. By processing frames individually and continuously, the system generates a video in which the selected individual is entirely erased, creating an effect similar to an invisibility cloak reminiscent of that depicted in *Harry Potter*, which inspired our system name.

Skeleton Overlay For this obfuscation option, the system first applies inpainting to remove the user from the frame. Then, a skeletal structure is overlaid on the inpainted image with pose keypoints extracted during the tracking process. The skeleton is rendered with color-coded lines, preserving the dynamics of the user's movement but removing any identifiable features.

4.4 Modality Synchronization and Truncation

Invisibility Cloak involves two modalities: RGB videos and IMU signals. However, they are recorded at different sampling rates. Video is recorded at 30 frames per second (FPS), while the IMU sensors operate at 50 Hz. To align these rates, we first extract the timestamp for each video frame from its metadata and then identify the closest corresponding IMU measurement based on its timestamp. After this process, the IMU recordings are downsampled from 50 Hz to 30 Hz, aligning with the video FPS for further processing. The durations of each user's performance under the same scenario are different. Therefore, merging all users' recordings can lead to some segments where only one user and one smartwatch are present, especially in the last several frames. In such cases, association is unnecessary. To address this, we only keep the synthesized multi-user video and IMU streams where at least three users are present in the scene. Then the IMU recordings and synthesized video frames are segmented into tracklets - sliding windows of fixed length T

Figure 4: Example activity scenes in different environments.

with an overlap of $T-1$ frames. Within each tracklet, we estimate the association at the final frame (frame T) by aggregating information from the preceding $T-1$ frames. The association is performed frame-by-frame as the window moves along the sequences. Intuitively, a smaller T allows the system to respond more quickly when a user initiates a video obfuscation request. Here, the value of T was set to 60, corresponding to a delay of 2 seconds. In practice, to prevent the leakage of privacy-sensitive video information, an edge device could withhold video for this short duration before streaming it to the cloud for further processing. For storage purposes, this brief delay has no impact. We further explore the effect of varying T on system performance in the next section.

5 DATA COLLECTION

This section details the data collection process, including participant recruitment, experimental setup, and data acquisition procedures in both fixed (*In-Lab*) and real-world (*In-the-wild*) environments. We also describe the data collection protocol and the synthetic dataset constructions for model training and evaluation.

5.1 Experiment Setup

The dataset includes two types of environmental conditions. The *In-Lab* study was conducted in a laboratory area of approximately 80 m². The *In-the-Wild* study involved a variety of real-world environments with natural, lived-in settings. Specifically, these environments were the participants' homes, including kitchens, dining rooms, and living rooms. In the *In-Lab* environment, the camera was mounted on a tripod, which was 1.5 meters above the ground. These camera angles were kept consistent across all participants. In the *In-the-Wild* environments, one camera was installed in each location. Since these were personal living spaces, the camera angle varied depending on the room layout. Across all environments, cameras were placed at an approximate height of 1.5 meters from the floor.

In the *In-Lab* environment, we recorded activities including ones commonly found in office spaces, classrooms, physical movement sequences, and general environment-independent gestures such as adjusting glasses and touching hair; in the *In-the-Wild* environments, we recorded activities natural to these environments such as cooking, cleaning, and watching TV. Figure 4 displays example scenes from both environmental conditions. Participants were instructed to perform the activities as naturally as they would in their daily lives. This setup aimed to capture real-world variations in surveillance perspectives and user behaviors in natural settings. Detailed scenario and activity descriptions are provided in Appendix A.

5.2 Participants and Procedures

We recruited 10 participants (4 females), each of whom performed the predefined activities following given instructions. During the sessions, the movements of the participants were recorded simultaneously using wearable IMU sensors and RGB cameras. The collected IMU data and RGB videos were stored in the smartwatch and camera, respectively. All 10 participants attended the In-Lab study. However, for the *In-the-Wild* study, 5 participants (all male) from the original group chose not to participate due to space constraints and personal preferences.

During data collection, participants wore a Google Pixel Watch 2 on both wrists, continuously writing IMU data (linear accelerometer and gyroscope signals) with a sampling rate of 50 Hz into internal memory. Visual data was captured using a tripod-mounted iPhone 12 Mini, recording video footage in 0.5x zoom mode at 30 fps. The distance between the camera and the participant is at least one meter. Each session lasted 30 minutes. Participants followed verbal instructions provided by the researchers and were encouraged to perform each action naturally. Each activity within a scenario was performed only once per participant, and the order of instructions was randomized.

5.2.1 Synthetic Dataset. One of the primary challenges in multi-user settings is accurate person tracking and labeling, which is both computationally intensive and time-consuming. Instead of manually annotating identities across frames, we create synthetic multi-user sequences by combining recordings from different individuals performing under the same scenarios. We estimated pose keypoints from each participant then overlaid multiple sequences of participants' keypoints onto a single timeline. Specifically, we first synthesize the video and IMU sequences from multiple users under the same scenario, resulting in matched video–IMU streams for the intended number of users, with each frame containing data ranging from 1 to 10 users. This approach enables us to train and test the model with various user and device combinations without requiring tedious, labor-intensive tracking annotations, or repeated data collections. For the *In-Lab* environment, we collected 125.91 minutes of data from 10 participants, and 26.31 minutes from 5 participants in the *In-the-Wild* setting. After synthesizing the multi-user sequences, the *In-Lab* and *In-the-Wild* datasets contain synchronized IMU–video sequences of 14.25 minutes and 5.1 minutes, respectively. This corresponds to 25,634 frames and 256,340 user–device association pairs in the *In-Lab* dataset, and 9,189 frames with 45,945 user–device association pairs in the *In-the-Wild* dataset. Figure 5 illustrates examples of synthesized multi-user video streams.

5.2.2 Multi-User Dataset. To validate the performance of *Invisibility Cloak* in real-world settings, we conducted an additional Multi-User data collection at the same place as the *In-Lab* settings. In this study, we invited 8 participants where 4 of whom had previously participated in the synthesized user study. 4 smartwatches were randomly assigned to 4 participants, and all devices were worn on their right wrists. Three cameras were set up to capture

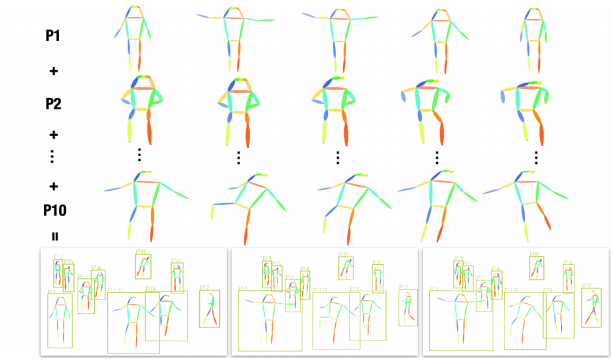

Figure 5: Illustration of the synthetic multi-user data generation process. The bottom row shows examples of synthesized frames where multiple users' skeletons are combined to simulate realistic multi-user scenes.

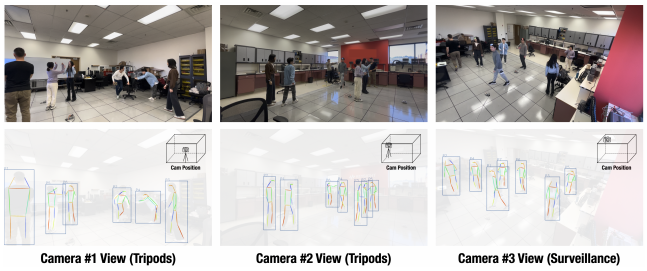

Figure 6: Multi-user data collection setup and camera views. Round colored stickers on participants' heads guided our later ground truth annotations. The lower row shows the corresponding camera positions and angles in the room with detected poses and bounding boxes.

the activities in this space: two were positioned similarly to the synthesized dataset (referred to as *Tripods*), and one simulated a surveillance camera placed high near the ceiling corners (referred to as *Surveillance*), illustrated in Figure 6. During data collection, a screen displayed all activities gathered from the synthesized dataset, and participants were instructed to randomly pick and perform one activity for approximately 30 seconds. Notifications were sent every 30 seconds to remind participants to switch to a new activity. This session lasted approximately 5 minutes. Following this, we conducted a second round of data collection using a similar protocol but with a new set of activities displayed on the screen for participants to select from, as detailed in Section A. Before this round, the smartwatches were retrieved and randomly reassigned among the participants. This second session lasted approximately 5 minutes. For the ground truth labeling, we first applied the YOLOv11 tracking model to detect and locate each individual's position and movements automatically. Due to the complexity of multi-user scenarios and potential tracking inaccuracies, we then manually reviewed and corrected the associations frame by frame to ensure accurate ground truth annotations.

6 EVALUATION

We assess the overall performance of *Invisibility Cloak* in this section. Given that we used state-of-the-art algorithms for human detection and pose estimation [31, 68], which have already achieved high precision, our evaluation primarily examines the accuracy of the associations between smartwatches and the corresponding individuals in the videos. Accurate user-device association is the foundation for camera-based obfuscations and is thus critical to preserving individual privacy. To comprehensively evaluate association performance, we conducted experiments across three distinct datasets: a synthesized dataset with a 10-user/10-smartwatch in lab configuration, a 5-user/5-smartwatch in the wild dataset collected in real-world home settings, and a real-time multi-user dataset associating 8-user/4-smartwatch to simulate dynamic, real-world conditions. In addition to evaluating association accuracy, we also report results on video obfuscation performance and conduct a real-time usability study to understand user experience and system effectiveness in practice.

6.1 Evaluation Metrics

To assess the performance of *Invisibility Cloak*, we adopt several metrics commonly used in multi-object tracking and person re-identification. Specifically, we evaluate the user-device association component using Identification Precision (IDP), Identification Recall (IDR), and the Identification F1 (IDF1) score.

In our setting, an association is considered as the True Positive (IDTP) when the predicted match between a smartwatch (IMU device) and a user in the video exactly matches the ground truth. The False Positive (IDFP) is the case where a smartwatch is incorrectly associated with a user. The False Negative (IDFN) is marked when a valid association exists in the ground truth, but the system fails to predict it. Then the Identification Precision (IDP) is defined as:

$$IDP = \frac{IDTP}{IDTP + IDFP}$$

It computes the fraction of detected individuals in the current frame that are correctly matched with their corresponding wireless devices. And the Identification Recall (IDR) and the Identification F1 score are defined as:

$$IDR = \frac{IDTP}{IDTP + IDFN}, \quad IDF1 = \frac{2 \times IDP \times IDR}{IDP + IDR}$$

IDR measures the proportion of true associations that are correctly retrieved, while the IDF1 score reflects the balance between precision and recall. It provides a comprehensive measure of the association module's performance.

6.2 Evaluation on Synthesized Dataset

6.2.1 User-Device Associations. In this section, we introduce the system performance on the synthesized dataset across different aspects. According to Section 5, the synthesized dataset comprises two parts: the *In-Lab* user study involving 10 users and the *In-the-Wild* study collected from 5 users. For each study, we designed 4 scenarios that include multiple common activities. Note that the *Independent* activity was featured in both studies. To best leverage the collected data, we configured all users wearing the devices to require obfuscation to generate the maximum number of user-device pairs to be tested. This configuration made false negatives

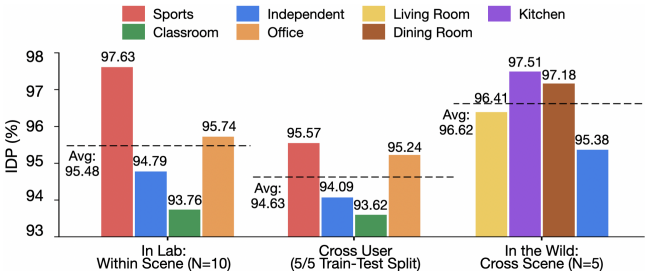

Figure 7: IDP across the three evaluation configurations. Left: Within-user evaluation results on the *In-Lab* dataset across four scenarios. Middle: Cross-user evaluation on the *In-Lab* dataset, trained on data from 5 participants and tested on the remaining 5. Right: Cross-scene evaluation on the *In-the-Wild* dataset with different activity sets.

rare to happen, thus, our analysis in this section primarily focuses on the precision score. Recall and the F1 scores will be evaluated with the multi-user real-time dataset in Section 6.3.

Within User Evaluation. Within-user evaluations were conducted using the *In-Lab* dataset to assess the system's performance in associating 10 users with 10 devices. We split each synthesized and synchronized recording stream (described in Section 5.2.1) into training (80%) and testing (20%) parts. To ensure an unbiased evaluation with no overlap between the training and testing sets, the 20% test samples were selected as multiple non-continuous 3-second segments across the whole recording streams, starting at random positions that did not intersect with any part of the training data. Subsequently, we applied a sliding window of 60 frames (2 s) with a step size of 1 frame moving along the subset recordings to predict the association between users and devices frame by frame. In the training process, we used Adam as the optimizer with a learning rate of 1×10^{-2} and employed a cosine annealing learning rate scheduler, training the model for 100 epochs.

The results are shown in Figure 7 (left). We observed that the average precision for the *In-Lab* study (10 users/10 devices associations) reached 95.48% (SD=1.425%). Among the four *In-Lab* scenarios, the highest association precision was achieved in the *Sports* scenario, where participants' large and dynamic movements facilitated accurate associations and obfuscation request matching. In contrast, relatively lower performance was observed in the *Independent* and *Classroom* scenarios. In the *Independent* scenario, participants only performed small, subtle arm activities without full-body motions such as touching face or adjusting glasses; also in the *Classroom* scenario, participants often remained almost still for some activities (e.g., while reading or checking emails on laptops), which makes IMU signals less distinguishable and thus association more challenging. Besides the precision, the average IDF1 score for the in-lab within-user evaluation was 97.64%, with a standard deviation of 0.757%. Overall, the association precision was consistently high, with an average precision of over 95% and no specific precision falling below 93%.

Cross User Evaluation. We also conducted cross-user evaluations to examine whether *Invisibility Cloak* can generalize to unseen users within the same environment. Specifically, we set the number of testing users N to 5, randomly selecting these users from the pool of 10 users available in the *In-Lab* dataset. Recordings from the remaining $10 - N$ users composed the training set. Then we conducted the 5-user/5-device association experiment. The sliding window size was also set to 60 frames with a step size of 1 frame, and the training procedure remained identical to that described in the within-user evaluations above, except the learning rate was set to 1×10^{-4} to prevent overfitting. The results are shown in Figure 7 (middle). The average association precision among 5 unseen users was 94.63% (SD = 0.801%). Consistent with the within-user evaluations, the *Sports* scenario exhibited the highest association accuracy, while the *Classroom* scenario still showed relatively low precision. However, we did not observe any significant drops in precision when associating and matching unseen participants with their smartwatch IMU data. These findings provide insights into the system's adaptability to effectively associate and track users, even when encountering previously unseen individuals performing similar actions under identical scenes.

Cross Scene Evalution. To assess the generalizability of *Invisibility Cloak* across different scenes (variations in camera positions, environments, and user activities), we trained the model using data from all 10 users in the *In-Lab* study and tested it on data from 5 users in the *In-the-Wild* study. The test data, collected from multiple users' homes, features distinct scenes, diverse camera angles, and a broader range of activities compared to the training data. The training procedure was identical to that used in the cross-user evaluations. The results are presented in Figure 7 (right). In this experiment, the study locations were each user's kitchen and living room, which differ substantially from the *In-Lab* settings. Moreover, the activities were entirely different from those in the *In-Lab* study, except for the *Independent* scenario. The average association precision achieved was 96.62% (SD = 0.818%), and the IDF1 score was 96.75% (SD = 0.562%) across the four scenarios. Although the activities differed, we did not observe any significant performance drops in the new scenes and activities, such as those in the *Kitchen*, *Living Room*, and *Dining Room* scenarios. We also noted that the *Independent* scenario still yielded the lowest performance, indicating that larger, more dynamic activities provide more information and facilitate more accurate associations. Overall, these results demonstrate the robustness and adaptability of *Invisibility Cloak* under realistic and diverse real-world conditions.

6.2.2 Window Length. In this experiment, we repeated the *Within User*, *Cross User*, and *Cross Scene* studies described above but with different window lengths for user-device associations. During this experiment, the training set consistently employed a fixed window length of 15 frames (0.5 second) to extract features while the testing set was evaluated using various window lengths – 15 (0.5 second), 30 (1 second), 45 (1.5 seconds), 60 (2 seconds), and 75 (2.5 seconds) – to determine how the amount of temporal context influences the accuracy of our association process. Figure 8 (A) shows the experimental results. We noticed that longer windows provide richer temporal context for better association, but excessively long windows may introduce noise or redundancy. Considering all these factors, we selected a 2-second window (60 frames) for all experiments conducted in the rest of this evaluation.

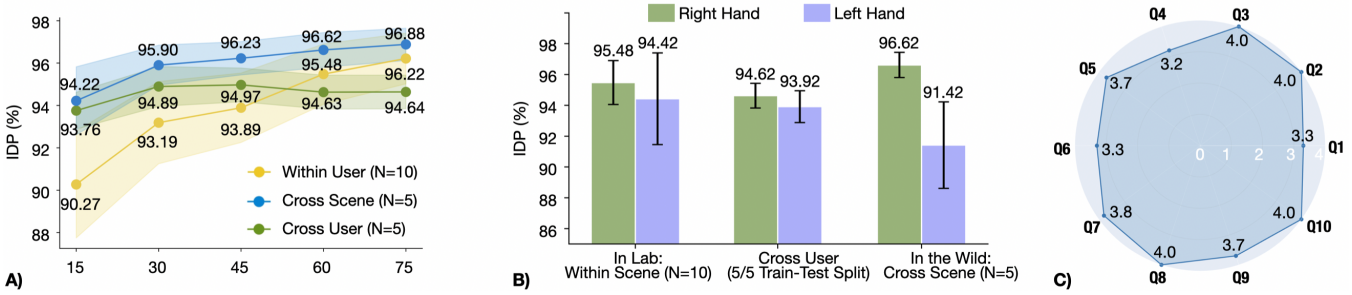

Figure 8: A) Effect of window length on user-device association. B) Effect of handedness on user-device association. C) Adjusted SUS Item Scores (0-4 scaled).

6.2.3 Handedness. In the synthesized data collection, all users were required to wear smartwatches on both their left and right wrists. The experiments described above were conducted using data from the right hand. To examine the influence of handedness on system performance, we conducted an additional experiment using data from the left hand. Among the 10 participants in the *In-Lab* study, only one participant is left-handed, while the remaining nine are all right-handed. For the cross-user and cross-scene experiments, all test participants' dominant hands are their right hands. We repeated the experiment described in Section 6.2.1, and the observations are displayed in Figure 8 (B). We found that the associations with right-hand IMU data are consistently better compared to the associations with left-hand IMU signals. This difference can be explained by the predominance of right-handed users in our training and testing dataset, which results in the model becoming better optimized for right-hand associations. Second, right-handed individuals generally exhibit more distinguishable movements with their dominant hand. They typically prefer to use their right hand for actions such as grabbing, dragging, and picking up objects, which provides clearer motion patterns for the system to learn and match. In contrast, left-hand IMU signals often capture smaller or less frequent movements, which increases ambiguity in the association with visual skeleton tracking. However, despite these differences with right or left hands, our system maintains robust performance regardless of which wrist the user wears the smartwatch on, indicating its overall resilience to variations in handedness.

6.3 Evalution on the Multi-User Dataset

We also evaluated the performance of *Invisibility Cloak* using the multi-user dataset, where the synthesized dataset served as the training set and the multi-user dataset was used as the test set. The experiments utilized a window length of 2 seconds (60 frames), with all other training protocols consistent with those described in Section 6.2.1. To address real-world scenarios where individuals requesting obfuscation are not always within the camera's field of view, we introduced an additional heuristic layer to enhance the practical deployment performance of *Invisibility Cloak*.

6.3.1 Unassociating Out-of-sight Users. Users who stream IMU data and send obfuscation requests may occasionally move out of the camera's field of view for a while (e.g., answering the phone outside the room). Since *Invisibility Cloak* employs the Hungarian algorithm for one-to-one matching, requests from users who are temporarily out-of-view could inadvertently be assigned to users currently visible in the camera's frame. This incorrect assignment can negatively affect the IDR and IDF1 scores. However, an out-of-sight user who requests obfuscation does not require immediate attention because their image is already naturally obscured. The primary concern is to ensure that obfuscation requests are not incorrectly assigned to visible users who either did not issue a request or issued a different request. To address this, we introduce an "unlikely" matching threshold. If any cost score derived from negating the similarity score for a given IMU-video pair exceeds this threshold, the system considers it highly improbable that this pair represents a valid match, thus preventing incorrect associations. Here, we set the threshold to 3.0 for the following multi-user experiment.

6.3.2 Association Results. Table 1 presents the detailed evaluation results after we applied the heuristic layers mentioned above. In the first round, where participants performed activities identical to those in the synthesized dataset and the cameras were similarly positioned on tripods, the precision achieved was comparable to the synthesized dataset results reported in Section 6.2. This indicates that the system successfully adapted to real-world scenarios where multiple users coexist in an environment. In addition, this high association precision confirms that our system can reliably fulfill obfuscation requests in realistic environments. However, the IDF1 score was slightly lower than the synthesized dataset results. This reduction was primarily due to cases where participants requesting obfuscation temporarily moved out of the camera's field of view, as the tripod-mounted cameras could not cover the entire space.

For the surveillance-camera scenario, we observed a slight performance drop due to the significant viewpoint differences from the training data. Nevertheless, the system maintained a high precision of approximately 92.31%, indicating that it remains reliable even with significant changes in camera viewpoints. During the second round of data collection, participants were asked to perform previously unseen activities. Of note that, the system's performance still remained comparable to that observed with seen activities. Therefore, our approach generalizes well across novel user activities without requiring fine-tuning with camera- or activity-specific training data.

Metric	IDP (%)	IDR (%)	IDF1 (%)
Round #1: Seen Activities			
Tripods	94.85	93.88	94.36
Surveillance	92.31	90.32	91.30
Round #2: Unseen Activities			
Tripods	91.18	89.61	90.39
Surveillance	89.41	88.67	89.04

Table 1: Evaluation results from the multi-user dataset.

6.3.3 Obfuscation Results. Additionally, we examined the offline-processed obfuscation performance from the Multi-User videos. We applied *Inpainting* obfuscation to all smartwatch-wearing participants (4 out of 8). Note that our pipeline first estimates poses and segments humans, then performs user association, and finally applies the corresponding obfuscation algorithm. For the robust performance of the obfuscation algorithms we adopted, we expected the final obfuscation performance to remain consistent across obfuscation options once segmentation and association have been determined, making these two steps more significant in the performance of obfuscation.

We defined an error frame as one in which a smartwatch-wearing participant was either not obfuscated at all or experienced significant pixel leakage of personally identifiable features (e.g., face, upper body shape). Minor visible regions (e.g., fingernails, shoe soles, or garment edges) were not considered as obfuscation errors, since these elements typically do not carry meaningful identity information. We identified a total of 42.36 seconds of obfuscation errors across all three camera position recordings, corresponding to 2.30% of the total recorded video. These errors primarily came from two sources: *user segmentation*, which was implemented using the YOLO11m Segmentation model. This component was not the focus of our contribution and was therefore not extensively evaluated; and *user-device association*, which is the core innovation of our system, was enabled via smartwatch-based camera-IMU association. Errors in this stage directly affected obfuscation performance.

6.4 Real-time Usability Study

6.4.1 Setup and Procedures. In addition to evaluating system performance, we conducted a supplemental usability study with 10 participants (5 female) to assess real-world applicability. The study followed the same experimental setup described in Section 5.1. Before the session, participants were shown an introductory video and a brief tutorial explaining the obfuscation system.

Each participant was asked to wear a smartwatch on their preferred wrist, as they normally would in daily life. The smartwatch ran a custom application (Figure 1 bottom left) that served as both an interface and a control tool, streaming IMU data and allowing participants to select their desired obfuscation level. The video stream was captured using the rear camera of an iPhone 12 mini. Both the video and IMU signals were synchronized and transmitted to a MacBook Pro with an M2 chip, which acted as the edge device. All data processing, computation, and video obfuscation were performed locally on this laptop.

During the study, each participant interacted with all obfuscation options at least twice. There were no constraints on user behavior. Participants were free to stand, sit, walk quickly, or run. They were also not instructed to avoid occlusions caused by objects or other people. Meanwhile, background activities from others in the same space continued naturally throughout the study environment. The association model used in this study was pre-trained on the *In-Lab* dataset with a 2-second time window. Real-time obfuscated video was projected onto a TV screen, allowing participants to observe the results of their interactions immediately. And these videos were recorded for further examination. After experiencing the system, each participant completed the System Usability Scale (SUS) questionnaire [8] to evaluate perceived usability.

6.4.2 Results. Figure 8 (C) shows the average adjusted SUS scores from 10 participants. Overall, *Invisibility Cloak* achieved a mean SUS score of 92.5 out of 100 (SD = 5.12). According to the adjective rating scale established in [4, 5], this score places *Invisibility Cloak* in the *"Best Imaginable"* category and within the *"Acceptable"* range for interface usability. These results indicate that participants found the system not only technically effective but also highly usable, intuitive, and well-suited for real-world applications.

After analyzing individual questions in the SUS questionnaire, we found that Q2, Q3, Q8, and Q10 received the highest scores (4.0). These questions relate to perceived ease of use and minimal learning effort. The results indicate strong agreement among participants that *Invisibility Cloak* is intuitive and easy to use without requiring extensive instruction. Additionally, we reviewed a total of 33 minutes of obfuscated video recorded during the study and manually verified the system's obfuscation performance. Each frame contained between 2 and 5 individuals. A frame was considered successfully obfuscated if the participant's body was masked according to their selected obfuscation level and if either no pixels from the user were visible or any visible pixels did not correspond to personally identifiable features. Across the entire recorded video, we identified 6 incorrect obfuscation clips totaling 3.7 seconds in duration, which represents 0.18% of the total recording. These findings further demonstrate the robustness and reliability of *Invisibility Cloak* in dynamic, real-world, multi-user environments.

7 DISCUSSION

7.1 Edge Implementation

A key objective of our system is to enable real-time, privacy-aware user identification and data obfuscation without uploading sensitive raw video to remote cloud infrastructure. In this section, we deployed the whole system pipeline, from video and IMU signal receiving, signal encoder, cross-modal association, and obfuscation components, to assess real-time performance and deployment feasibility on common edge devices. Specifically, we selected a desktop equipped with an NVIDIA RTX 4060 GPU and an Apple Mac Mini M4 (featuring 10 CPU cores, 10 GPU cores, 16 GB of memory, and a 256 GB SSD) [23] as the edge devices for our deployment study. We also conducted tests on a workstation with an NVIDIA RTX A5500 GPU as a performance baseline. To maximize performance on each platform, we deployed the models in hardware-optimized formats. Models running on Apple Silicon were converted to CoreML, while those on NVIDIA GPUs were saved in TorchScript format. We then ran the whole pipeline on these devices using a window length T of

Device	Raw	Mask	Blur	Inpaint	Skeleton
Mac Mini	21.29	18.40	18.46	12.16	11.48
Desktop	87.26	32.12	30.32	12.51	11.04
Workstation	80.91	30.95	30.19	17.07	14.47

Table 2: FPS for five obfuscation modes across three edge devices.

60 frames, matching 10 users with 10 devices based on predefined obfuscation requests. In this evaluation, we measured the processing speed over a continuous 1-minute deployment and results in frames per second (FPS) that are summarized in Table 2.

According to the results, we observed that the *Masking* and *Blurring* obfuscation methods, both non-deep learning-based, were the fastest, consistently achieving over 30 FPS. In contrast, *Inpainting* and *Skeleton*, which rely on generative models, were more computationally intensive and required longer processing times. On the workstation, the *Inpainting* obfuscation processing reached an average of 17.07 FPS, and the *Skeleton* ran at around 15 FPS. Although the Mac Mini is less powerful than the workstation, performance remained acceptable. Both *Inpainting* and *Skeleton* operated at around 12 FPS, and *Masking* and *Blurring* are all about 18 FPS. In this case, the real-time deployment of *Invisibility Cloak* is feasible on both high-end and edge-class hardware, with performance suitable for practical privacy-aware applications in real-world scenarios.

7.2 Power and Computational Cost

The system *Invisibility Cloak* performs associations between users and devices relying on three deep learning models: the YOLO model for pose estimation, a customized association model for matching, and a lightweight generative model for inpainting. Both the YOLO model and the inpainting models we leveraged are pretrained, and their computational costs are detailed in the works [26, 48]. Therefore, we mainly focus on the computational cost of the customized model here. For the tracklet length T of 60 frames and 10 users per frame matched with 10 active smartwatches, the customized model requires 0.458G floating-point operations (FLOPs) with 0.381M parameters. In the real-world deployment study described in Section 7.1, the power consumption varied across different edge devices. We measured the power usage when running the full system pipeline (from pose estimation to inpainting) using an electricity usage monitor [62]. To isolate the system's actual power draw, we also recorded baseline power consumption when the system was not running the system code then subtracted this value to find out about our system's real-world power consumption. The power consumption recorded for the Mac Mini was about 25.9 W, and for the desktop with an RTX 4060 GPU was around 79 W, indicating feasibility for long-term deployment scenarios.

7.3 Accuracy and Robustness

Although our system demonstrates promising results in multi-user tracking and identification, it does not yet achieve perfect accuracy (i.e., 100%) in real-world scenarios, as no system can guarantee flawless performance. The association algorithm sometimes misidentifies users, particularly when individuals exhibit similar movement patterns or gesture features. Our system currently relies on commercial models such as YOLO for pose estimation and human segmentation. While these models are widely adopted and generally robust, the performance may degrade under extreme conditions, such as severe occlusions, poor lighting, or rapid movements. However, in the supplemental usability study, participants engaged in various natural activities, including rapid movement and occlusion by objects or people, but we did not observe significant performance degradation. This demonstrates the system's resilience in real-world scenarios.

To enhance accuracy and robustness further, future work will focus on developing a multi-camera distributed system and integrating additional modalities, such as thermal and depth imaging. We also plan to incorporate confusion minimization and long-term correction mechanisms as supplementary strategies to improve system performance under diverse and dynamic environmental conditions.

7.4 System Scalability and Adaptability

Although our experiments have demonstrated effective multi-user locating and tracking in moderately crowded settings (e.g., 10 participants in an indoor space), system performance may degrade when the number of users increases substantially in very crowded spaces (e.g., bustling malls with over 100 pedestrians). In such scenarios, issues like occlusions, tracking ambiguities, frequent obfuscation requests, and increased computational load can affect both obfuscation precision and system responsiveness. However, in large spaces that are typically monitored by multiple cameras (since a single camera would rarely accommodate and record so many individuals), the workload can be dispersed across different cameras and integrated at the edge server using distributed processing strategies.

We also acknowledge that the participant pool in this study, while sufficient for feasibility validation, was limited in both size and diversity. This may affect the system's generalizability across broader user populations and application contexts. To improve scalability, generalizability, and robustness, future work will focus on recruiting more participants and evaluating the system in a broader range of environments. We also plan to explore more efficient and optimized algorithms, as well as strategies for dynamic resource allocation, to enhance the system's scalability and maintain robust performance in locating and tracking target users under congested conditions. Additionally, we aim to conduct long-term real-world deployments and testing (e.g., over several months) to gather deeper insights and guide future system refinements.

8 CONCLUSION

We proposed *Invisibility Cloak*, a visual-inertial system that leverages wearable IMU signals to enable personalized video obfuscation in shared environments. Our system associates IMU streaming from smartwatches with human activities captured in surveillance videos to dynamically identify users and apply privacy-preserving requests according to individual preferences. We conducted comprehensive evaluations on both synthesized and real-time multi-user scenarios. We also designed hierarchical obfuscation levels to accommodate diverse user privacy preferences across different real-world scenarios, ranging from raw video transmission to advanced inpainting

with skeleton overlays. In addition, our system demonstrated low-latency performance on edge devices, ensuring real-time processing in practical applications. We believe that *Invisibility Cloak* uniquely addresses the challenge of balancing utility and privacy in vision AI, paving the way for practical, privacy-aware applications in surveillance and IoT systems.

REFERENCES

[1] Alexandre Alahi, Albert Haque, and Li Fei-Fei. 2015. RGB-W: When vision meets wireless. In *Proceedings of the IEEE International Conference on Computer Vision*. 3289–3297.

[2] Rawan Alharbi, Mariam Tolba, Lucia C Petito, Josiah Hester, and Nabil Alshurafa. 2019. To mask or not to mask? balancing privacy with visual confirmation utility in activity-oriented wearable cameras. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 3 (2019), 1–29.

[3] Adriano Arra, Alessio Bianchini, Joana Chavez, Pietro Ciravolo, Fatjon Nebiu, Martina Olivelli, Gabriele Scoma, Simone Tavoletta, Matteo Zagaglia, and Alessio Vecchio. 2019. Personalized gait-based authentication using UWB wearable devices. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 206–210.

[4] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[5] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[6] Adeola Bannis, Shijia Pan, Carlos Ruiz, John Shen, Hae Young Noh, and Pei Zhang. 2023. IDIoT: Multimodal framework for ubiquitous identification and assignment of human-carried wearable devices. *ACM Transactions on Internet of Things* 4, 2 (2023), 1–25.

[7] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

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

[9] Bryan Bo Cao, Abrar Alali, Hansi Liu, Nicholas Meegan, Marco Gruteser, Kristin Dana, Ashwin Ashok, and Shubham Jain. 2022. ViTag: online WiFi fine time measurements aided vision-motion identity association in multi-person environments. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 19–27.

[10] Siyuan Cao and He Wang. 2018. Enabling public cameras to talk to the public. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–20.

[11] Hongkai Chen, Sirajum Munir, and Shan Lin. 2022. Rfcam: Uncertainty-aware fusion of camera and wi-fi for real-time human identification with mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–29.

[12] Cory Cornelius, Ronald Peterson, Joseph Skinner, Ryan Halter, and David Kotz. 2014. A wearable system that knows who wears it. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. 55–67.

[13] Gaole Dai, Huatao Xu, Hyungjun Yoon, Mo Li, Rui Tan, and Sung-Ju Lee. 2024. ContrastSense: Domain-invariant Contrastive Learning for In-the-Wild Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–32.

[14] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. 2018. Personalized privacy assistants for the internet of things: Providing users with notice and choice. *IEEE Pervasive Computing* 17, 3 (2018), 35–46.

[15] Anupam Das, Martin Degeling, Xiaoyou Wang, Junjue Wang, Norman Sadeh, and Mahadev Satyanarayanan. 2017. Assisting users in a world full of cameras: A privacy-aware infrastructure for computer vision applications. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1387–1396.

[16] Mariella Dimiccoli. 2018. J. Mar ın, and E. Thomaz,"Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation,". *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol* 1 (2018).

[17] Shiwei Fang, Tamzeed Islam, Sirajum Munir, and Shahriar Nirjon. 2020. Eyefi: Fast human identification through vision and wifi-based trajectory matching. In *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 59–68.

[18] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. 2021. Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[19] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.

[20] Zhixiang He, Jing Chen, Cong Wu, Kun He, Ruiying Du, Ju Jia, Yangyang Gu, and Xiping Sun. 2024. HCR-Auth: Reliable Bone Conduction Earphone Authentication with Head Contact Response. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–27.

[21] 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.

[22] Roberto Henschel, Timo Von Marcard, and Bodo Rosenhahn. 2020. Accurate long-term multiple people tracking using video and body-worn imus. *IEEE Transactions on Image Processing* 29 (2020), 8476–8489.

[23] Apple Inc. 2024. Mac Mini with M4 Chip. https://www.apple.com/shop/buy-mac/mac-mini/apple-m4-chip-with-10-core-cpu-and-10-core-gpu-16gb-memory-256gb

[24] Yasha Iravantchi, Thomas Krolikowski, William Wang, Kang G Shin, and Alanson Sample. 2024. PrivacyLens: On-Device PII Removal from RGB Images using Thermally-Enhanced Sensing. *Proceedings on Privacy Enhancing Technologies YYYY (X)* 1 (2024), 20.

[25] Tatsuya Ishihara, Kris M Kitani, Chieko Asakawa, and Michitaka Hirose. 2018. Deep radio-visual localization. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 596–605.

[26] Glenn Jocher and Jing Qiu. 2024. *Ultralytics YOLO11*. https://github.com/ultralytics/ultralytics

[27] Youssef Khazbak, Junpeng Qiu, Tianxiang Tan, and Guohong Cao. 2020. TargetFinder: A privacy preserving system for locating targets through IoT cameras. *ACM Transactions on Internet of Things* 1, 3 (2020), 1–23.

[28] Darko Kirovski, Michael Sinclair, and David Wilson. 2007. The martini synch: joint fuzzy hashing via error correction. In *European Workshop on Security in Ad-hoc and Sensor Networks*. Springer, 16–30.

[29] Belal Korany, Chitra R Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for through-wall person identification from candidate video footage. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–15.

[30] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

[31] Gongjin Lan, Yu Wu, Fei Hu, and Qi Hao. 2022. Vision-based human pose estimation via deep learning: A survey. *IEEE Transactions on Human-Machine Systems* 53, 1 (2022), 253–268.

[32] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–24.

[33] Hansi Liu, Abrar Alali, Mohamed Ibrahim, Bryan Bo Cao, Nicholas Meegan, Hongyu Li, Marco Gruteser, Shubham Jain, Kristin Dana, Ashwin Ashok, et al. 2022. Vi-fi: Associating moving subjects across vision and wireless sensors. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 208–219.

[34] John Mamish, Rawan Alharbi, Sougata Sen, Shashank Holla, Panchami Kamath, Yaman Sangar, Nabil Alshurafa, and Josiah Hester. 2024. NIR-sighted: A programmable streaming architecture for low-energy human-centric vision applications. *ACM Transactions on Embedded Computing Systems* 23, 6 (2024), 1–26.

[35] Alessandro Masullo, Tilo Burghardt, Dima Damen, Toby Perrett, and Majid Mirmehdi. 2019. Who goes there? exploiting silhouettes and wearable signals for subject identification in multi-person environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[36] Filippo LM Milotta, Antonino Furnari, Sebastiano Battiato, Maria De Salvo, Giovanni Signorello, Giovanni Maria Farinella, et al. 2019. Visitors Localization in Natural Sites Exploiting EgoVision and GPS.. In *VISIGRAPP (5: VISAPP)*. 556–563.

[37] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2022. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395* (2022).

[38] Jason Moore, Peter McMeekin, Thomas Parkes, Richard Walker, Rosie Morris, Samuel Stuart, Victoria Hetherington, and Alan Godfrey. 2024. Contextualizing remote fall risk: Video data capture and implementing ethical AI. *NPJ digital medicine* 7, 1 (2024), 61.

[39] Tamara Mujirishvili, Anton Fedosov, Kooshan Hashemifard, Pau Climent-Pérez, and Francisco Florez-Revuelta. 2024. "I Don't Want to Become a Number": Examining Different Stakeholder Perspectives on a Video-Based Monitoring System for Senior Care with Inherent Privacy Protection (by Design).. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[40] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8427–8436.

[41] Le T Nguyen, Yu Seung Kim, Patrick Tague, and Joy Zhang. 2014. IdentityLink: user-device linking through visual and RF-signal cues. In *Proceedings of the*

2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 529–539.

[42] Takayuki Nishio and Ashwin Ashok. 2016. High-speed mobile networking through hybrid mmWave-camera communications. In *Proceedings of the 3rd Workshop on Visible Light Communication Systems*. 37–42.

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[44] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 324–337.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[46] Carlos Ruiz, Shijia Pan, Adeola Bannis, Ming-Po Chang, Hae Young Noh, and Pei Zhang. 2020. IDIoT: Towards ubiquitous identification of IoT devices through visual and inertial orientation matching during human activity. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 40–52.

[47] Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. 2017. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[48] Andranik Sargsyan, Shant Navasardyan, Xingqian Xu, and Humphrey Shi. 2023. MI-GAN: A Simple Baseline for Image Inpainting on Mobile Devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7335–7345.

[49] Sujay Shalawadi, Christopher Getschmann, Niels van Berkel, and Florian Echtler. 2024. Manual, Hybrid, and Automatic Privacy Covers for Smart Home Cameras. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 3453–3470.

[50] Jiayu Shu, Rui Zheng, and Pan Hui. 2016. Cardea: Context-aware visual privacy protection from pervasive cameras. *arXiv preprint arXiv:1610.00889* (2016).

[51] Yoko Sogabe, Shiori Sugimoto, Ayumi Matsumoto, and Masaki Kitahara. 2024. Pre-capture Privacy via Adaptive Single-Pixel Imaging. *arXiv preprint arXiv:2407.00991* (2024).

[52] Julian Steil, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling. 2019. Privacy-aware eye tracking using differential privacy. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–9.

[53] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. 2019. Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Proceedings of the 11th ACM symposium on eye tracking research & applications*. 1–10.

[54] 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.

[55] Yuanyi Sun, Sencun Zhu, and Yu Chen. 2022. ZoomP3: Privacy-preserving publishing of online video conference recordings. *Proceedings on Privacy Enhancing Technologies* (2022).

[56] Steven L Tanimoto, Alon Itai, and Michael Rodeh. 1978. Some matching problems for bipartite graphs. *Journal of the ACM (JACM)* 25, 4 (1978), 517–525.

[57] Thiago Teixeira, Deokwoo Jung, and Andreas Savvides. 2010. Tasking networked cctv cameras and mobile phones to identify and localize multiple people. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 213–222.

[58] Kit-Lun Tong, Kun-Ru Wu, and Yu-Chee Tseng. 2021. The device–object pairing problem: Matching IoT devices with video objects in a multi-camera environment. *Sensors* 21, 16 (2021), 5518.

[59] Junjue Wang, Brandon Amos, Anupam Das, Padmanabhan Pillai, Norman Sadeh, and Mahadev Satyanarayanan. 2017. A scalable and privacy-aware IoT service for live video analytics. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. 38–49.

[60] Xue Wang and Yang Zhang. 2021. Nod to auth: Fluent ar/vr authentication with user head-neck modeling. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[61] Zhiwei Wang, Yihui Yan, Yueli Yan, Huangxun Chen, and Zhice Yang. 2022. CamShield: Securing Smart Cameras through Physical Replication and Isolation. In *31st USENIX Security Symposium (USENIX Security 22)*. 3467–3484.

[62] P3 Kill A Watt. 2025. Electricity Usage Monitor. https://www.amazon.com/P3-P4400-Electricity-Usage-Monitor/dp/B00009MDBU Last accessed 14 July 2025.

[63] Xin Yang and Omid Ardakanian. 2023. Blinder: End-to-end privacy protection in sensing systems via personalized federated learning. *ACM Transactions on Sensor Networks* 20, 1 (2023), 1–32.

[64] Xin Yang and Omid Ardakanian. 2024. Guided Diffusion Model for Sensor Data Obfuscation. *arXiv preprint arXiv:2412.14499* (2024).

[65] Lotus Zhang, Abigale Stangl, Tanusree Sharma, Yu-Yun Tseng, Inan Xu, Danna Gurari, Yang Wang, and Leah Findlater. 2024. Designing Accessible Obfuscation Support for Blind Individuals' Visual Privacy Management. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[66] Shikun Zhang, Yuanyuan Feng, Lujo Bauer, Lorrie Faith Cranor, Anupam Das, and Norman Sadeh. 2021. "Did you know this camera tracks your mood?": Understanding Privacy Expectations and Preferences in the Age of Video Analytics. *Proceedings on Privacy Enhancing Technologies* (2021).

[67] Yupeng Zhang, Yuheng Lu, Hajime Nagahara, and Rin-ichiro Taniguchi. 2014. Anonymous camera for privacy protection. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 4170–4175.

[68] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *Comput. Surveys* 56, 1 (2023), 1–37.

[69] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.

[70] Haozhe Zhou, Mayank Goel, and Yuvraj Agarwal. 2024. Bring Privacy To The Table: Interactive Negotiation for Privacy Settings of Shared Sensing Devices. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–22.

[71] Jizhe Zhou, Chi-Man Pun, and Yu Tong. 2020. Privacy-sensitive objects pixelation for live video streaming. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3025–3033.

A USER STUDY SCENARIOS

A.1 In Lab

Office Space.

(1) Randomly draw something on a whiteboard.
(2) Try to explain the drawing.
(3) Erase the drawing.
(4) Carry a stack of books in your arms, walk into the room, and put the stack down on a desk.
(5) Return the books to the shelf, placing each one back in its place.
(6) Pull a book from the shelf, open it, and read.
(7) Sit down and take notes with pens.
(8) Close the book and place it back on the shelf.

Classroom.

(1) Walk over to the desk and drop the backpack onto it.
(2) Wipe the desk clean, then toss the wipe into the trash bin.
(3) Take the laptop out of the backpack and open it.
(4) After a moment, shut the laptop and put it back into the backpack.
(5) Stand up, wear the backpack, and step out of the room.

Sports.

(1) ROM (a range of motion sequence - followed by video).
(2) Walking around the Room.
(3) Freestyle (except standing still).

Environment Independent.

(1) Adjust smart glasses.
(2) Touch hair.
(3) Touch face.
(4) Stretch arms.
(5) Scratch the other hand/arm.
(6) Adjust clothes.

A.2 In the Wild

Kitchen (Cooking&Eating).

(1) Head to the fridge, open it, and take out something.
(2) Place it on the counter.
(3) Pick up the cutting board and knife and peel an apple.
(4) Open the microwave, place a dish inside, and close the door.
(5) Open the microwave and take out the dish.
(6) Return to the counter and start eating the apple.

Dining Room (Cleaning).

(1) Roll up the sleeves.
(2) Move the chair.
(3) Clean the floor.
(4) Open a drawer, put something inside, then close it.
(5) Open the vacuum and clean the desk.

Living Room (Watching TV).

(1) Head to the couch, grab a cushion, and pad it with your hands.
(2) Put the cushion on the couch and sit down.
(3) Pick up the remote and press the button to turn on the TV.
(4) Put the remote down on the table and lean back on the couch.

Environment Independent.

(1) Adjust smart glasses.
(2) Touch hair.
(3) Touch face.
(4) Stretch arms.
(5) Scratch the other hand/arm.
(6) Adjust clothes.

B MULTI-USER SCENARIOS

(1) Pick up a mug
(2) Hand the mug to another user
(3) Place the mug on the table
(4) Waving/pointing using hands and arms
(5) Using Smartphone
(6) Passing and stacking books
(7) Rearranging papers on a table
(8) Drinking water
(9) Shaking hands
(10) Talk with someone
(11) High-five
(12) Fist bumping
(13) Wiping hands
(14) Throwing an object with others
(15) Rolling wrists
(16) Unboxing
(17) Mimicking guitar strumming
(18) Tapping on others' shoulders
(19) Rock-paper-scissors games with multiple users
(20) Rubbing hands
(21) Spinning an imaginary pen
(22) Giving a thumbs-up
(23) Crossing arms and then extending outward
(24) Salute
(25) Swinging an imaginary tennis racket
(26) Mimicking a baseball throw
(27) Lifting an object overhead
(28) Open the door
(29) Pressing the button
(30) Making a dramatic "Welcome" wave