Sequence-Based Identification of First-Person Camera Wearers in Third-Person Views

Anonymous CVPR submission

Paper ID XXXXX

Abstract

As immersive education and collaborative environments continue to advance, understanding complex student interactions within these shared spaces has become increasingly important. While large-scale datasets like Ego4D and Ego-Exo4D have advanced egocentric vision research, they lack the rich, multi-user interactions critical for collaborative learning and robotics. To address this gap, we introduce TF2025, an expanded dataset featuring synchronized first- and third-person views of actors, enhanced synchronization, and multiple train-test splits. We also propose a sequence-based approach for identifying first-person camera wearers in broader third-person views. By leveraging motion cues and person re-identification, our method improves robustness and significantly outperforms state-of-the-art approaches. This work advances the analysis of multi-camera interactions in embodied vision and education. The code and dataset will be made publicly available upon acceptance.

1. Introduction

Rapid advancements in virtual reality (VR), augmented reality (AR) and narrative-centered learning technologies [6, 36, 46], have enabled the development of highly interactive applications in immersive education [25, 40]. In these settings, first-person (egocentric) cameras, such as body-worn or head-mounted devices, capture dynamic, close-range perspectives directly from the students' viewpoints. In contrast, traditional third-person (exocentric) cameras, such as those monitoring a classroom, provide a broader, contextual view that complements this egocentric footage. Combining information from these two types of views offers significant potential for analyzing complex student-student or student-teacher interactions in collaborative environments [29, 69].

Datasets like **Ego4D** [30] and **Ego-Exo4D** [32] have been instrumental in advancing egocentric vision. Ego4D offers a large-scale collection of first-person videos capturing daily activities. Building on this, Ego-Exo4D pairs egocentric footage with third-person (exocentric) views, creating a multi-view dataset that enables researchers to explore cross-view relationships—such as how robots or first-person camera wearers can use widely available third-person videos to enhance learning [39].

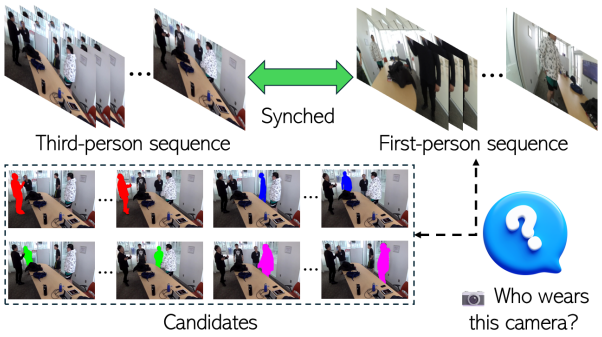


Figure 1. **Problem definition.**

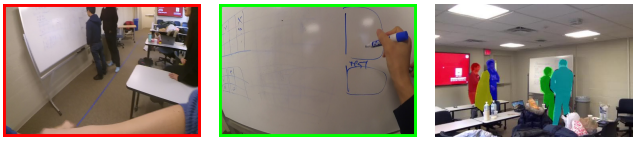
However, Ego-Exo4D is limited to scenarios featuring only a single camera wearer, restricting its applicability in collaborative educational environments—such as multi-user VR classrooms and immersive training [2, 77]—where multiple students frequently interact and operate egocentric devices simultaneously.

In this work, we study the identification of camera wears in third-person views. In multi-user immersive learning environments, systems typically maintain basic identity information for students. However, simply knowing a student's identity leaves a critical disconnect between their first-person perspective and the overarching third-person view. For example, an ID tag does not reveal how a student's egocentric focus correlates with their body actions and surroundings, which is crucial for activity recognition and skill assessment. To create this essential link, we define our task to identify the student's precise segmentation mask in the third-person view using their first-person video as a clue, as shown in Fig. 1. This mask-level cross-view match-

ing provides significantly more value for fine-grained action assessment, though it presents a major technical challenge.

Several methods [26, 79, 83] have been proposed for this task. However, these approaches are frame-based, making predictions using only frame-level information. Due to the inherent challenges of first-person cameras, such as occlusions and motion blur, relying on single frames reduces reliability. Additionally, since camera wearers frequently move, the first-person view can sometimes become uninformative. For instance, when the camera points at blank walls, white boards or other featureless surfaces.

In this work, we introduce a sequence-based approach for identifying first-person camera wearers in third-person views. Compared to frame-based methods, our approach offers several key advantages. First, by using multiple frames, the model reduces reliance on any single uninformative frame. Second, our experiments demonstrate that matching the motion of camera wearers in third-person views with optical flow in first-person views provides strong and reliable evidence about identity. Lastly, using a sequence of frames increases the chance of camera wearers seeing each other, which leads to an improved identification clue.



(a) First-person view 1 (b) First-person view 2 (c) Third-person view

Figure 2. **Data sample from TF2025**. The TF2025 dataset aims to establish a link between students' first-person views and their third-person segmentation masks in classrooms.

To support this objective, we introduce **TF2025 (Third-First 2025)**, an expanded dataset designed to benchmark algorithms that facilitate these collaborative environments, as shown in Fig. 2. TF2025 integrates videos from three source datasets: **TF2023** [83], **IUShareView** [79], and **Ego4D-TF**, a subset of Ego4D enhanced with our additional annotated ground truth masks. TF2025 is 2.3 times the size of TF2023, the current largest dataset for this task, and includes more challenging frames. It also features three new train/test splits designed to evaluate multiple levels of generalizability, directly translating to diverse educational and training applications. Furthermore, we conducted manual synchronization to improve temporal alignment between first- and third-person views.

To summarize, our main contributions are as follows:

1. We introduce TF2025, a larger and more challenging dataset featuring multiple new evaluation benchmarks tailored for complex multi-user interactions.
2. We propose *motion matching*, a novel approach for first- and third-person matching.
3. We present a new framework termed *Motion Appearance Fusion (MAF)* that adaptively fuses information from multiple cues, outperforming state-of-the-art methods.

2. Related Works

Egocentric Video Understanding has advanced computer vision by capturing first-person experiences for applications like activity recognition [17, 28, 30, 55, 57, 70], object interaction [30, 50, 54, 63], and navigation. Recent studies leverage self-supervised learning to enhance action recognition [4, 49, 53, 60, 78], while others focus on gaze prediction [37, 38, 47, 58, 75] and hand-object interactions [74, 80] to infer human intent. Egocentric videos also aid robotics in task and imitation learning [41–43]. In social analysis, they help detect engagement, infer relationships, and identify conversational partners [3, 5, 11, 30, 83]. Large-scale datasets like Ego4D [30] have further driven research, supporting a wide range of applications from human-robot collaboration to social behavior modeling.

Cross-View Matching aligns corresponding elements across different views to enhance spatial understanding. Recent advancements leverage Graph Neural Networks (GNNs) and attention mechanisms for improved feature matching. SuperGlue [64] uses GNNs with positional encodings to refine keypoint descriptors, while LoFTR [72] employs self- and cross-attention for feature alignment. ASpanFormer [15] introduces a hierarchical attention with adaptive attention span adjustment. Other approaches enhance efficiency [14, 66] and unify dense and sparse matching methods [24, 76]. GlueStick [61] refines correspondence estimation by jointly matching points and lines.

Joint First- and Third-person Video Understanding Recent research has made significant strides in connecting first-person (egocentric) and third-person (exocentric) video understanding. Approaches include cross-view action recognition [8, 67, 71], generating egocentric video from third-person footage, aligning frames, and learning viewpoint-invariant representations [48]. Cross-view data has improved pose estimation [1, 16, 54]. Other work focuses on summarizing egocentric videos with third-person views [20, 35], refining egocentric models using third-person data [54, 62], and co-segmentation via attention prediction. Datasets with both views include Ego-Exo4D [31], Assembly101 [65], and Charades-Ego [68]. Other datasets like Epic-Kitchens [17–19] also combine these views, supporting gaze prediction and activity modeling.

Person identification has been explored in the context of first- and third-person cross-view understanding. While some work has focused on top-view cameras [7, 9, 10, 81], our work aligns with the use of side-view cameras as third-person views [26, 79, 83]. Although side-view cameras provide a narrower field of view compared to top-view cameras, they offer richer appearance details for the camera

wearer candidates. Recent advancements, such as multi-branch deep networks and semi-Siamese architectures, have improved joint embeddings for aligning first- and third-person views [26, 79], while PEN [83] introduces a dual-branch framework, combining a personal branch for relational analysis and an environment branch for geometric matching. Building on these foundations, our work focuses on identifying camera wearers in third-person views by incorporating temporal motion cues and multi-cue information fusion to achieve better performance.

3. Methodology

3.1. Problem Definition

We define the task of identifying the camera wearer in third-person views as follows: Given a t-frame sequence of first-person frames $Ego : \{Ego_1, Ego_2, \ldots, Ego_t\}$, a sequence of third-person frames $Exo : \{Exo_1, Exo_2, \ldots, Exo_t\}$, and N mask sequences in the third-person view representing N candidate people,

$$\begin{aligned} &\text{Candidate 1: } \{Mask_{11}, Mask_{12}, \ldots, Mask_{1t}\} \\ &\text{Candidate 2: } \{Mask_{21}, Mask_{22}, \ldots, Mask_{2t}\} \\ &\qquad\vdots \\ &\text{Candidate N: } \{Mask_{N1}, Mask_{N2}, \ldots, Mask_{Nt}\}, \end{aligned}$$

the goal is to identify the mask sequence which represents the correct candidate that corresponds to the first-person frame sequence $Ego : \{Ego_1, Ego_2, \ldots, Ego_t\}$. In our setting, we choose $t = 30$ as a balance between having enough frames for meaningful information and avoiding excessive computational cost; however, our introduced method can be applied to sequences of arbitrary length.

This task presents unique and significant challenges. Unlike traditional person identification or re-identification [12, 82, 84], we cannot rely on the appearance of the camera wearers themselves, which is typically the most robust source of information. As egocentric camera wearers are often at the center of interactions, their view presents additional challenges: limited scene coverage due to the camera's closeness to the center of interaction, frequent occlusions from nearby objects or people, and motion blur from quick movements. These issues can substantially affect the quality and reliability of the first-person view, complicating information extraction.

Prior work [83] shows the importance of multi-source integration for this task, introducing a frame-based approach that utilizes geometric cues from image matching and person detection in both views. However, image matching is often too unreliable, as methods like GlueStick [61] rely on significant overlap between views—an assumption that frequently breaks between first-person and third-person views. This limitation highlights the need for more robust techniques to address the differences in these views.

In this section, we present a novel sequence-based framework, *Motion Appearance Fusion (MAF)*. First, we introduce a new matching method based on camera motion, termed *motion matching*. Next, we describe our *appearance matching* model, which computes appearance similarity between individuals in first-person and third-person views. Finally, we propose the *Confidence-Based Adaptive Fusing (CBAF)* module to integrate information from these two sources. The structure of *MAF* is illustrated in Fig. 3.

3.2. Motion Matching

First, we establish a connection between the body motion of camera wearers in the third-person view and the optical flow in the first-person view. While optical flow has been widely used in egocentric vision analysis [22, 59, 79], to our knowledge, we are the first to explicitly link it to third-person body motion for first-third cross-view matching.

We start by discussing the concept of motion within the context of this paper. Some egocentric datasets, such as **Ego4D** and **Ego-Exo4D**, utilize head-mounted and eyewear cameras, while others like **IUShareView** and **TF2023** use chest-worn cameras. Our approach does not restrict the prediction of motion to any predefined body part. Instead, we seek to predict movements from any body part that could induce changes (optical flow) in the first-person view. This design choice enhances the generalizability of our model, as body-worn cameras in real-world scenarios can be mounted on various body parts, such as the shoulder, wrist, or waist.

In addition, we define two values to model the camera motion: T and R. Camera motion can be represented by translational and rotational motions, which have distinct impacts on the optical flow in the first-person view. For translational motion, due to motion parallax, objects closer to the camera exhibit faster apparent motion across the field of view compared to those farther away, resulting in a depth-dependent optical flow pattern. In contrast, for rotational motion, the optical flow is depth-independent.

To model this, we apply a sequence-based backbone, *MViT* [27], to a candidate in the third-person view, which is cropped using the third-person frames and the candidate's segmentation mask. We then use a Multi-Layer Perceptron (MLP) to predict two values, T_{exo} and R_{exo}, which together form a system that models the egocentric camera motion.

For the first-person view, we first compute two values, T_i and R_i for each frame i,

$$T_i = \underset{j \in frame_i}{\text{median}} \left(depth(p_{ij}) \sqrt{fl_x(p_{ij})^2 + fl_y(p_{ij})^2} \right) \quad (1)$$

$$R_i = \underset{j \in frame_i}{\text{median}} \left(\sqrt{fl_x(p_{ij})^2 + fl_y(p_{ij})^2} \right), \quad (2)$$

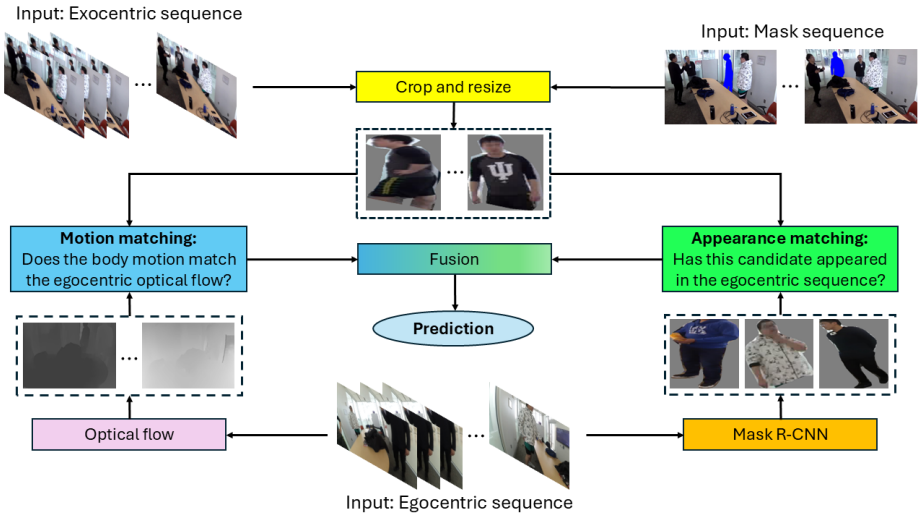


Figure 3. **Overview of the structure of our framework (MAF)** We utilize two models: *motion matching*, which matches the motion of the candidate to the egocentric optical flow, and *appearance matching* that checks if the candidate appeared in the first-person view. Then we integrates information from both sides to make the final prediction.

where p_{ij} denotes the j-th pixel in frame i, and $depth(p_{ij})$, $fl_x(p_{ij})$, and $fl_y(p_{ij})$ represent the depth, x-axis optical flow, and y-axis optical flow, respectively.

We then compute the total motion of the first-person view by aggregating the two motion values T and R across all frames, represented as a tuple of the cumulative motion, as shown in Eq. 3,

$$\text{motion}(\{Ego_1, \ldots, Ego_t\}) = \left(\sum_{i=1}^{t-1} T_i, \sum_{i=1}^{t-1} R_i \right). \quad (3)$$

In our model, we utilize *Raft* [73] to compute optical flow, and *ZoeDepth* [13], a monocular depth estimation model, to predict absolute depth. Additionally, we employ *Mask R-CNN* [33] to filter out foreground pixels where people are detected, as their presence can introduce independent motion not caused by the camera wearer.

For training, we randomly select 16 consecutive frames from a 30-frame sequence. Using our motion model, we predict the motion values T_{exo} and R_{exo} for the selected sequence. The loss function is computed as the sum of squared errors between the predicted motion values and the corresponding ground truth values derived from the respective first-person sequence.

During inference, we employ a sliding window approach to evaluate motion similarity across overlapping sequences. Specifically, we calculate the sum of squared errors for the motion values across three overlapping sequences: frames 1–16, 8–23, and 15–30. This sum of squared errors is denoted as "motion score". A detailed illustration of the *motion matching* model is provided in the supplementary materials.

3.3. Appearance Matching

The presence of individuals in the first-person view provides additional cues for matching, as camera wearers should not see themselves in their own first-person views. This information has been utilized in prior work on third/first-person matching. *PEN* [83] introduces a personal branch that implicitly learns this information through cross-attention between third- and first-person views. Their model takes as input the third-person candidate, the individuals visible in the corresponding first-person view, and a binary training label indicating whether the pair is a match.

However, we empirically discovered that this approach is suboptimal, as the visibility of the camera wearer in the first-person view does not strictly correlate with their identity. For instance, even if a camera wearer is not visible in their own first-person view, it does not necessarily imply that the view belongs to them. In this work, we demonstrate empirically that training a person re-identification (re-ID) model directly is more effective for this task.

We implement our person re-ID model, denoted *appearance matching*, using a *Vision Transformer (ViT)* [23] backbone. During training, the backbone outputs a feature vector, and we attach an MLP head to generate classification labels. During inference, we use the feature vector to compute the L2 distance between pairs of individuals. Following existing person re-ID works [34], we use three data augmentation techniques: random horizontal flip, random cropping, and random patching, each applied with a 50% probability. We apply two loss functions: cross-entropy loss for classification and triplet loss for metric learning.

While it is natural to use the camera wearer as the anchor

and another person in their view as the negative sample, this approach may result in non-camera wearers only appearing as negatives during training. To address this imbalance, we introduced "pseudo" first/third-person pairs by using non-camera wearers as anchors and the camera wearer as the negative sample, as depicted in Fig. 4.

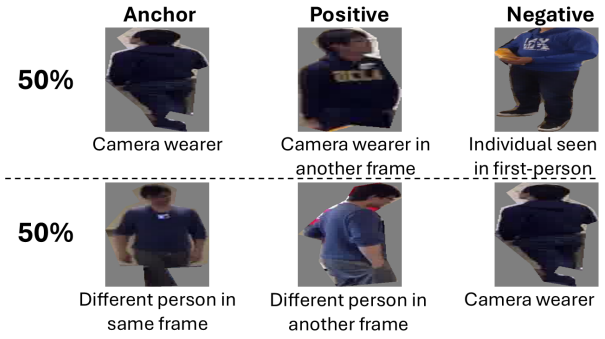


Figure 4. **Triplet selection.**

During inference, we select frames 1, 16, and 30 from the first-person sequence and apply *Mask R-CNN* to crop out all detected individuals. Each detected individual and third-person candidates are passed through the same *ViT* backbone, and we compute its average L2 distance from the third-person candidate across all frames, denoted as "appearance score". These three frames were chosen as a balance between capturing most individuals appearing in the first-person sequence and avoiding excessive detections of the same person, which could complicate the fusion process. We provide a detailed illustration of the *appearance matching* model in the supplementary materials.

3.4. Confidence-Based Adaptive Fusing

After *motion matching* and *appearance matching*, we have the following results for N candidates in third-person and M detected individuals in the first-person sequence, where motion scores and appearance scores represent the results from the two models:

$$\text{Motion scores: } \{Motion_1, \ldots, Motion_N\}$$
$$\text{Appear. scores 1: } \{Appearance_{11}, \ldots, Appearance_{1N}\}$$
$$\vdots$$
$$\text{Appear. scores M: } \{Appearance_{M1}, \ldots, Appearance_{MN}\}$$

The goal of this task is to identify the correct candidate among $\{1, \ldots, N\}$ that matches the first-person sequence. This poses several challenges for the fusion method:

1. Both M and N are different for each query.
2. The same individual can be detected multiple times in the 3 chosen frames.
3. Motion scores and appearance scores behave differently. A lower motion score reflects better matching between the third/first-person motion, while a lower appearance score means the candidate is seen in the first-person view, so unlikely to be the camera wearer.

We propose *Confidence-Based Adaptive Fusing (CBAF)*, a novel module designed to integrate results from the two preceding modules. *CBAF* adaptively selects which source of information to use for candidate elimination or final prediction, requires no training for different datasets, and can handle arbitrary values of M and N.

We first define the concept of confidence in the context of this paper:

$$\text{Confidence} = \lambda_{\text{Re-ID}} \cdot \alpha_{\text{mask}} \cdot \frac{x_{(2)}}{x_{(1)}}$$

where:

- $x_{(1)}$ is the smallest value among N values from the same source,
- $x_{(2)}$ is the second smallest value among N values from the same source,
- $\lambda_{\text{Re-ID}}$ is a constant representing our trust in the *appearance matching* model, which depends on prior knowledge about the dataset (e.g., how clearly individuals are visible in the first-person view). For motion scores, $\lambda_{\text{Re-ID}} = 1$,
- α_{mask} is the confidence from *Mask R-CNN* when detecting this individual. For motion scores, $\alpha_{\text{mask}} = 1$.

Essentially, this calculates the ratio between the two smallest items from each score source. A higher confidence value indicates a clearer distinction between the most prominent candidate and the runner-up, while a lower confidence value suggests ambiguity between the two candidates. This ratio helps us evaluate the reliability of each source of information, guiding how much trust we place in their respective results.

The *CBAF* algorithm is presented in Algorithm 1. It compares the confidence of motion scores against that of all appearance confidence scores. If the motion scores exhibit the highest confidence, the candidate with the lowest motion score is selected as the prediction. Conversely, if an appearance source yields the highest confidence, the corresponding candidate is removed from all sources, and that source is subsequently eliminated. Examples illustrating this algorithm are provided in the supplementary materials.

4. Experiments

4.1. Dataset

We introduce **TF2025 (Third-First 2025)**, an expanded dataset built on the foundation of three datasets: **TF2023**, **IUShareView**, and **Ego4D**.

We begin by merging TF2023 with IUShareView, as these two datasets were both recorded in classroom settings, utilize the same camera models, and feature overlapping actors. IUShareView [26, 79] was designed for the task of

Algorithm 1 Confidence-Based Adaptive Fusing (CBAF)

Require: Motion scores $\{Motion_1, \ldots, Motion_N\}$, appearance scores $\{Appearance_{i1}, \ldots, Appearance_{iN}\}$ for $i = 1, \ldots, M$, $Candidates$ $\{1, \ldots, N\}$

while len$(Candidates) > 1$ **do**
  Compute motion confidence: $C_{\text{motion}} \leftarrow \frac{Motion_{(2)}}{Motion_{(1)}}$
  Initialize best source $s \leftarrow 0$, best confidence $C_{\text{best}} \leftarrow C_{\text{motion}}$
  for each appearance source $i = 1, \ldots, M$ **do**
    Compute appearance confidence: $C_{\text{Re-ID}_i} \leftarrow \lambda_{\text{Re-ID}} \cdot \alpha_{\text{i}} \cdot \frac{Appearance_{i(2)}}{Appearance_{i(1)}}$
    if $C_{\text{Re-ID}_i} > C_{\text{best}}$ **then**
      Update $s \leftarrow i$, $C_{\text{best}} \leftarrow C_{\text{Re-ID}_i}$
    end if
  end for
  if $s = 0$ **then**
    return $Candidates[\arg\min(Motion)]$
  else
    Remove the candidate with index $\arg\min(s)$ from all sources and $Candidates$
    Remove $\{Appearance_{s1}, \ldots, Appearance_{sN}\}$ from all scores
  end if
end while
return $Candidates[0]$

identifying first-person camera wearers, containing 9 sets of frame sequences. Later, TF2023 [83] was introduced and is currently the largest dataset featuring synchronized first- and third-person views, along with segmentation masks for every individual in third-person views. Both datasets were designed to study multi-camera wearer interactions in social scenarios, such as presentations, discussions and snack breaks. To support this focus, the official splits of both datasets include only those frames where actors are actively engaged in interactions. In TF2025, we address a more realistic scenario by extracting all frames from the source videos of these two datasets and constructing continuous 30-frame sequences. The newly incorporated frames introduce significantly greater challenges compared to existing ones, as examples shown in Fig. 5.

Ego4D [30], launched in 2021, is a large-scale egocentric dataset with over 3,670 hours of first-person video captured using wearable cameras. Ego4D does not inherently support this task, as it primarily consists of egocentric videos featuring solo camera wearers. To address this, we created and annotated **Ego4D-TF**, a subset for cross-view matching. We identified scenes in Ego4D containing at least three people and at least two synchronized egocentric videos. We then sampled one of every six frames to match the frame rate of TF2023 and IUShareview and selected one of the

Figure 5. Examples of highly challenging first-person frames including camera wearers looking at a whiteboard, ceiling, or being occluded by their own arms.

first-person views as a "pseudo-third-person" view. Using YOLOv11 [45], we extracted segmentation masks for visible individuals and tracked them across frames for each 30-frame sequence. We provide a more comprehensive explanation and examples of our annotation method in the supplementary materials.

Manual Synchronization: We performed manual calibration to improve the synchronization accuracy of TF2023 and IUShareView using a technique inspired by the manual synchronization procedure used in **Ego-Exo4D**. By identifying clear single-frame actions visible across multiple cameras, such as moments when a hand touches specific objects, we aligned their timestamps across different camera views. This process reduced the temporal alignment error by up to 4 frames (0.8 seconds).

Cross-dataset Evaluation: Prior work [83] identified a phenomenon in this task where some methods may unintentionally exploit dataset biases in third-person views. For instance, in TF2023, camera wearers often appear more frequently near the center of frames, leading models to learn visual cues that are not transferable to other datasets or applications. To address this issue, we constructed a test split using Ego4D-TF, enabling cross-dataset evaluation to better assess generalizability.

Dataset Splits: Following prior works in first- and third-person learning [32, 54], we created three splits designed to evaluate methods at different levels of generalizability. We believe this is crucial for evaluating diverse potential applications. For instance, in an immersive learning environment where multiple students engage in VR/AR group activities, the model should be trainable using data from the classroom and the appearances of the participants. In contrast, in public settings such as detecting unauthorized camera wearers, there may be no prior knowledge of the location or even the camera models used.

We designed three levels of splits, as outlined below:

1. **Seen:** We allocated the first 80% of each source video from TF2023 and IUShareView for training and the remaining 20% for testing.
2. **Unseen:** We split the videos from TF2023 and IUShareView such that neither the camera wearers nor the locations overlap between training and testing.
3. **Cross-Dataset:** We used data from TF2023 and IUShareView for training and Ego4D-TF for testing.

Methodology	TF2023 Unseen(Acc↑)	TF2025 Seen(Acc↑)	TF2025 Unseen(Acc↑)	TF2025 Cross-dataset(Acc↑)
Random guess	32.5%	28.7%	33.1%	46.4%
Sequence baseline	34.7%	39.4%	34.1%	47.0%
ThirdFirstNet [79]	35.4%	31.1%	33.8%	47.4%
PEN [83]	67.2%	85.9%	58.8%	52.9%
Motion matching(ours)	84.6%	88.8%	80.5%	67.7%
Appearance matching(ours)	79.5%	66.1%	63.4%	72.2%
MAF(ours)	**90.3%**	**93.8%**	**84.4%**	**75.7%**

Table 1. **Experiment Results on TF2025 and TF2023.** Models were evaluated by matching a first-person sequence to the correct third-person candidate, given a third-person sequence and multiple candidate masks. Results are reported in accuracy. We converted the original TF2023 dataset to 30-frame sequences in order to compare with sequence-based methods.

This ensures no shared individuals, locations, or camera models between the training and testing sets.

A detailed comparison of the sizes of TF2023 and each split of TF2025 is provided in Tab. 2.

Dataset	TF2023	TF2025		
Split	Unseen	Seen	Unseen	Cross-Dataset
Number of Frames				
Train	47335	104760	80490	131010
Test	21689	26250	50520	32880
Total	**69024**	131010	131010	**160890**
Number of Sequences				
Train	N/A	3492	2683	4367
Test	N/A	875	1684	1096
Total	**N/A**	4367	4367	**5363**

Table 2. **Detailed Comparison of dataset sizes.** TF2025 is approximately 2.3 times larger than TF2023, offering multiple evaluation splits and providing full 30-frame sequences rather than sporadic frame sets.

4.2. Implementation Details and Results

We implemented three baseline methods for comparison: two frame-based methods and one sequence-based method.

PEN [83] is the current state-of-the-art method for this task. We applied it to each of the 30 frames and summed its prediction scores to generate a sequence-level prediction.

ThirdFirstNet [79] originally requires segmentation masks from the previous frame, making it unsuitable for this task. Following the adaptation in [83], we modified it to use the current frame's masks and the model's identification branch. Frame-level predictions were summed to generate sequence-level results.

Since no sequence-based methods exist for this task, we implemented a *two-stream* baseline using two sequence-based backbones (*MViT*): one for third-person and one for first-person sequences. The third-person backbone was modified to accept 4-channel inputs by concatenating the candidate mask sequence with the third-person sequence. Both sequences are fed into their respective backbones and trained with a contrastive loss: matching sequences minimize the L2 distance between outputs, while non-matching sequences maximize it. During inference, we also employed the sliding window approach and selected the candidate mask sequence with the minimal cumulative L2 distance to the first-person sequence.

For our proposed method (*MAF*), we trained our *motion matching* and *appearance matching* models separately using the AdamW [52] optimizer with a learning rate of 1e-4. Both models were trained to process input images and sequences with a resolution of 224×224 pixels. The *motion matching* model was trained with a batch size of 32 for 50 epochs, while the *appearance matching* model used a batch size of 64 for 100 epochs. The *MViT* backbone for *motion matching* was pretrained on Kinetics-400 [44], and the *ViT* backbone for *appearance matching* was pretrained on ImageNet [21]. All experiments were conducted on two Nvidia RTX A6000 GPUs. We set the hyperparameter $\lambda_{\text{Re-ID}}$ to 1.75, 1.75 and 3 for the three splits, supported by sensitivity study in supplement materials.

The experimental results are shown in Tab. 1. *MAF* significantly outperforms frame-based methods across all splits, particularly in the unseen and cross-dataset splits, where higher levels of generalizability are required. On the original TF2023 dataset, *PEN* achieved an accuracy of 67.2%. However, its performance dropped to 58.8% on the unseen split of TF2025 due to the inclusion of more challenging frames.

We observed that while *PEN* achieved significant improvements over random guessing on both the seen and unseen splits, it failed to outperform random guessing by a large margin on the cross-dataset split. We attribute this to *PEN*'s reliance on geometric matching, which requires a

Dataset	TF2025		
Methodology	Seen	Unseen	Cross-dataset
T only	85.1%	73.8%	64.2%
R only	83.0%	75.3%	66.7%
Both (ours)	**88.8%**	**80.5%**	**67.7%**

Table 3. Ablation study on motion calculation.

Dataset	TF2025		
Methodology	Seen	Unseen	Cross-dataset
PEN-Personal	**83.0%**	50.3%	48.3%
Appearance (ours)	66.1%	**63.4%**	**72.2%**

Table 4. Ablation study on appearance matching.

	Dataset	TF2025		
Methodology	Frames	Seen	Unseen	Cross-dataset
No fusing	N/A	88.8%	80.5%	67.7%
Direct	16	77.6%	66.7%	67.3%
CBAF(ours)	16	91.5%	82.2%	72.9%
Subtraction	1,16,30	89.9%	79.7%	73.1%
CBAF(ours)	1,16,30	**93.8%**	**84.4%**	**75.7%**

Table 5. Ablation study on fusing methods.

portion of the view to be visible in both the first-person and third-person views. Essentially, it assumes that the camera wearer is facing away from the third-person camera. However, in the Ego4D-TF setting, where participants often face each other in board games, this assumption does not hold, showing its lack of robustness.

The introduced *two-stream* baseline achieved a slight improvement over random guessing on the seen split but did not surpass random guessing on the unseen and cross-dataset splits by any notable margin. This demonstrates that the significant improvement of our model stems from its design of visual cues, rather than the sequence-based backbones themselves.

4.3. Ablation studies

We conducted several ablation studies to justify the choice of our model design in each modules.

First, we validated our design to predict two motion values separately. We compared this approach to predicting only one of the two motion values, i.e., T or R. As shown in Tab. 3, our method outperformed single-source prediction across all three splits.

Second, we compared our *appearance matching* model to the *personal branch* introduced in *PEN* [83], as both methods utilize the appearances of individuals in first-person views. The results are shown in Tab. 4. We observed that the *personal branch* of *PEN* significantly outperformed our method on the seen split. This is because the *personal branch* model takes both the third-person candidate and the individual in the first-person view as inputs, learning their differences implicitly through cross-attention. As a result, for the seen split, it can memorize the appearance of third-person camera wearers and make predictions even when no individual appears in the first-person sequence. In contrast, our method takes in the L2 distance between individuals from both views. If no individual appears in the first-person sequence, our *appearance matching* method defaults to random guessing. However, this does not negatively impact our overall performance, as our fusion method make predictions based on the *motion matching* results when this happens.

On the other two splits, where camera wearers did not appear during training (unseen and cross-dataset), our method significantly outperformed *PEN*, demonstrating much better generalizability. This finding also underscores the benefit of having multiple levels of splits in the TF2025 dataset for evaluation.

Finally, we evaluated the performance of our fusion method (*CBAF*). We compared it against two other methods: *subtraction fusing*, which subtracts appearance scores from motion scores, and *direct fusing*, which uses individuals detected in the first-person view to eliminate candidates based on the order of *Mask R-CNN* confidence scores and then selects the best candidate among the remaining, using motion scores. Since *direct fusing* cannot handle cases where the same person appears multiple times, we also modified our *CBAF* method to use only frame 16 for comparison. The results are shown in Tab. 5.

5. Conclusion

We explore the task of identifying first-person camera wearers in third-person views. We introduce a new dataset, **TF2025**, which integrates two existing datasets (**IUShareView** and **TF2023**) with additional frames, along with **Ego4D-TF**, an annotated subset of **Ego4D**. The dataset also includes three new splits to evaluate different levels of generalizability. We propose a novel sequence-based framework, *MAF*, which incorporates a new *motion matching* method and an *appearance matching* model. Additionally, we introduce a novel method to integrate results from these two models. Experiments show that our approach significantly outperforms existing works, achieving state-of-the-art performance. Future work could incorporate more egocentric datasets to increase the diversity of camera wearer appearances and camera models. This could also lead to improved models for measuring camera motions and better generalization across various camera models.

References

[1] Adnan Abdullah, Ruo Chen, Ioannis Rekleitis, and Md Jahidul Islam. Ego-to-exo: Interfacing third person visuals from egocentric views in real-time for improved rov teleoperation. *arXiv preprint arXiv:2407.00848*, 2024. 2

[2] Laure Abensur Vuillaume, Jonathan Goffoy, Nadège Dubois, Nathacha Almoyner, Cécile Bardet, Evelyne Dubreucq, Sophie Klenkenberg, Anne-Françoise Donneau, Camille Dib, Alexandre Ghuysen, et al. Collaborative virtual reality environment in disaster medicine: moving from single player to multiple learners. *BMC medical education*, 24(1):422, 2024. 1

[3] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva. With whom do i interact? detecting social interactions in egocentric photo-streams. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2959–2964. IEEE, 2016. 2

[4] Peri Akiva, Jing Huang, Kevin J Liang, Rama Kovvuri, Xingyu Chen, Matt Feiszli, Kristin Dana, and Tal Hassner. Self-supervised object detection from egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5225–5237, 2023. 2

[5] Stefano Alletto, Giuseppe Serra, Simone Calderara, and Rita Cucchiara. Understanding social relationships in egocentric vision. *Pattern Recognition*, 48(12):4082–4096, 2015. 2

[6] Asmaa Saeed Alqahtani, Lamya Foaud Daghestani, and Lamiaa Fattouh Ibrahim. Environments and system types of virtual reality technology in stem: A survey. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(6), 2017. 1

[7] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 253–268. Springer, 2016. 2

[8] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171:61–68, 2018. 2

[9] Shervin Ardeshir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 285–300, 2018. 2

[10] Shervin Ardeshir, Sandesh Sharma, and Ali Broji. Egoreid: Cross-view self-identification and human re-identification in egocentric and surveillance videos. *arXiv preprint arXiv:1612.08153*, 2016. 2

[11] Sophia Bano, Tamas Suveges, Jianguo Zhang, and Stephen J Mckenna. Multimodal egocentric analysis of focused interactions. *IEEE Access*, 6:37493–37505, 2018. 2

[12] Apurva Bedagkar-Gala and Shishir K Shah. A survey of approaches and trends in person re-identification. *Image and vision computing*, 32(4):270–286, 2014. 3

[13] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4

[14] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6301–6310, 2021. 2

[15] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. *European Conference on Computer Vision (ECCV)*, 2022. 2

[16] Feng Cheng, Mi Luo, Huiyu Wang, Alex Dimakis, Lorenzo Torresani, Gedas Bertasius, and Kristen Grauman. 4diff: 3d-aware diffusion model for third-to-first viewpoint translation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024. 2

[17] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021.

[19] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2

[20] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2016. 2

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[22] Ameya Dhamanaskar, Mariella Dimiccoli, Enric Corona, Albert Pumarola, and Francesc Moreno-Noguer. Enhancing egocentric 3d pose estimation with third person views. *Pattern Recognition*, 138:109358, 2023. 3

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[24] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 2

[25] Noel Enyedy, Joshua A Danish, and David DeLiema. Constructing liminal blends in a collaborative augmented-reality learning environment. *International Journal of Computer-Supported Collaborative Learning*, 10(1):7–34, 2015. 1

[26] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017. 2, 3, 5

[27] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 3

[28] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *2011 international conference on computer vision*, pages 407–414. IEEE, 2011. 2

[29] Joyce Horn Fonteles, Clayton Cohn, Efrat Ayalon, Mengxi Zhou, Ashwin TS, Eduardo Davalos, Zhijian Li, Surya Rayala, Divya Mereddy, Austin Coursey, et al. Analyzing embodied learning in classroom settings: A human-in-the-loop ai approach for multimodal learning analytics. *Learning and Instruction*, 103:102274, 2026. 1

[30] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 1, 2, 6

[31] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2024. 2

[32] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 1, 6

[33] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[34] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9664–9667, 2023. 4

[35] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European conference on computer vision (ECCV)*, pages 70–85, 2018. 2

[36] Miao Hu, Xianzhuo Luo, Jiawen Chen, Young Choon Lee, Yipeng Zhou, and Di Wu. Virtual reality: A survey of enabling technologies and its applications in iot. *Journal of Network and Computer Applications*, 178:102970, 2021. 1

[37] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 2

[38] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 2

[39] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 1

[40] Lasse Jensen and Flemming Konradsen. A review of the use of virtual reality head-mounted displays in education and training. *Education and information technologies*, 23 (4):1515–1529, 2018. 1

[41] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360, 2022. 2

[42] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024.

[43] Haresh Karnan, Garrett Warnell, Xuesu Xiao, and Peter Stone. Voila: Visual-observation-only imitation learning

for autonomous navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2497–2503. IEEE, 2022. 2

[44] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7

[45] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 6, 15

[46] James C Lester, Hiller A Spires, John L Nietfeld, James Minogue, Bradford W Mott, and Eleni V Lobene. Designing game-based learning environments for elementary science education: A narrative-centered learning perspective. *Information Sciences*, 264:4–18, 2014. 1

[47] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013. 2

[48] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 2

[49] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 2

[50] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2

[51] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 13

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[53] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[54] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *European Conference on Computer Vision*, pages 407–425. Springer, 2024. 2, 6

[55] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016. 2

[56] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016. 13

[57] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 2

[58] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021. 2

[59] Keisuke Ogaki, Kris M Kitani, Yusuke Sugano, and Yoichi Sato. Coupling eye-motion and ego-motion features for first-person activity recognition. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7. IEEE, 2012. 3

[60] A Emin Orhan, Wentao Wang, Alex N Wang, Mengye Ren, and Brenden M Lake. Self-supervised learning of video representations from a child's perspective. *arXiv preprint arXiv:2402.00300*, 2024. 2

[61] Rémi Pautrat, Iago Suárez, Yifan Yu, Marc Pollefeys, and Viktor Larsson. Gluestick: Robust image matching by sticking points and lines together. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9706–9716, 2023. 2, 3

[62] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, 132(11):4880–4936, 2024. 2

[63] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021. 2

[64] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[65] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 2

[66] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12517–12526, 2022. 2

[67] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018. 2

[68] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos, 2018. 2

[69] Nithin Sivakumaran, Chia-Yu Yang, Abhay Zala, Shoubin Yu, Daeun Hong, Xiaotian Zou, Elias Stengel-Eskin, Dan Carpenter, Wookhee Min, Cindy Hmelo-Silver, et al. A multimodal classroom video question-answering framework for automated understanding of collaborative learning. In *Proceedings of the 27th International Conference on Multimodal Interaction*, pages 516–525, 2025. 1

[70] Sibo Song, Vijay Chandrasekhar, Bappaditya Mandal, Liyuan Li, Joo-Hwee Lim, Giduthuri Sateesh Babu, Phyo Phyo San, and Ngai-Man Cheung. Multimodal multi-stream deep learning for egocentric activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 24–31, 2016. 2

[71] Bilge Soran, Ali Farhadi, and Linda Shapiro. Action recognition in the presence of one egocentric and multiple static cameras. In *Asian Conference on Computer Vision*, pages 178–193. Springer, 2014. 2

[72] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[73] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4

[74] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019. 2

[75] Sanket Kumar Thakur, Cigdem Beyan, Pietro Morerio, and Alessio Del Bue. Predicting gaze from egocentric social interaction videos and imu data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 717–722, 2021. 2

[76] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10247–10266, 2023. 2

[77] Nesse Van der Meer, Vivian van der Werf, Willem-Paul Brinkman, and Marcus Specht. Virtual reality and collaborative learning: A systematic literature review. *Frontiers in Virtual Reality*, 4:1159905, 2023. 1

[78] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021. 2

[79] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–652, 2018. 2, 3, 5, 7

[80] Yue Xu, Yong-Lu Li, Zhemin Huang, Michael Xu Liu, Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. Egopca: A new framework for egocentric hand-object interaction understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5273–5284, 2023. 2

[81] Liang Yang, Hao Jiang, Jizhong Xiao, and Zhouyuan Huo. Ego-downward and ambient video based person location association. *arXiv preprint arXiv:1812.00477*, 2018. 2

[82] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 3

[83] Ziwei Zhao, Yuchen Wang, and Chuhua Wang. Fusing personal and environmental cues for identification and segmentation of first-person camera wearers in third-person views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16477–16487, 2024. 2, 3, 4, 6, 7, 8

[84] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 3

Sequence-Based Identification of First-Person Camera Wearers in Third-Person Views

Supplementary Material

6. Additional Visualization of Models

We provide detailed visualization of our *motion matching* and *appearance matching* methods in Fig. 6 and Fig. 7.

7. Additional Ablation Studies

We evaluated the choice of backbones and the effectiveness of the sliding window approach. For this analysis, we selected two transformer-based video methods, *MViT* and *Swin3D* [51], using their default weights from torchvision [56], both pretrained on **Kinetics-400**. We compared two input strategies: (1) direct prediction using the full 30-frame sequence (frames 1 to 30), and (2) cumulative predictions from three overlapping 16-frame sequences (frames 1–16, 7–23, and 14–30). This comparison was conducted using *Swin3D*, as the pretrained *MViT* weights do not support 30-frame inputs. The results of this ablation study are presented in Tab. 6.

Dataset	TF2025		
Methodology	Seen	Unseen	Cross-dataset
Swin3d w/o sliding window	74.7%	63.2%	57.7%
Swin3d w/ sliding window	82.2%	72.9%	64.5%
MViT w/ sliding window	**88.8%**	**80.5%**	**67.7%**

Table 6. Ablation study on backbones and sliding window.

We validated our choice of optical flow computation. Specifically, we compared our approach of calculating the hypotenuse of the optical flow components along the x and y axes against using optical flow from a single direction, as shown in Tab. 7.

Dataset	TF2025		
Methodology	Seen	Unseen	Cross-dataset
X axis only	83.1%	72.8%	63.3%
Y axis only	74.7%	70.9%	63.5%
Hypotenuse of x and y	**88.8%**	**80.5%**	**67.7%**

Table 7. Ablation study on optical flow directions.

We also performed a sensitivity study on the hyperparameter $\lambda_{\text{Re-ID}}$ in CBAF, which balances between motion and appearance matching. As shown in Tab. 8, optimal values are 1.75 for the seen split, 1.75 for the unseen split, and 3 for the cross-dataset split, with minimal differences from non-optimal values. This demonstrates the adaptability of our method to different scenarios. For example, the higher value in the cross-dataset scenario reflects the greater reliability of appearance matching compared to motion matching due to the difference in camera models. In addition, CBAF exhibits robustness to changes in this hyperparameter, with performance varying only by $\sim 1\%$ in the seen split for values ranging from 1.25 to 2.5.

Dataset	TF2025		
$\lambda_{\text{Re-ID}}$	Seen	Unseen	Cross-dataset
0(motion)	88.8%	80.5%	67.7%
0.25	88.9%	80.6%	67.7%
0.5	89.6%	81.5%	68.0%
0.75	90.5%	82.5%	69.0%
1	91.4%	82.6%	71.4%
1.25	93.0%	83.6%	73.0%
1.5	93.5%	84.3%	73.6%
1.75	**93.8%**	**84.4%**	73.7%
2	93.3%	83.9%	74.7%
2.25	92.9%	83.6%	74.8%
2.5	93.3%	83.2%	75.4%
3	92.3%	82.4%	**75.7%**
3.5	91.9%	81.8%	75.5%
4	91.9%	81.4%	75.2%

Table 8. Sensitivity analysis for the hyperparameter of CBAF.

To further demonstrate the robustness of our model to the selection of λ, we conducted validation by splitting each test set into three subsets. In each run, we used one subset to select λ, and the remaining two for testing, averaging the results across all three combinations. As shown in Tab. 9, this results in only a negligible reduction ($\leq 0.5\%$) while we significantly outperform state-of-the-art ($\sim$20%).

Selection of λ	Seen	Unseen	Cross
w/ validation	93.7%	84.3%	75.2%
w/o validation	93.8%	84.4%	75.7%

Table 9. Validation for the selection of λ

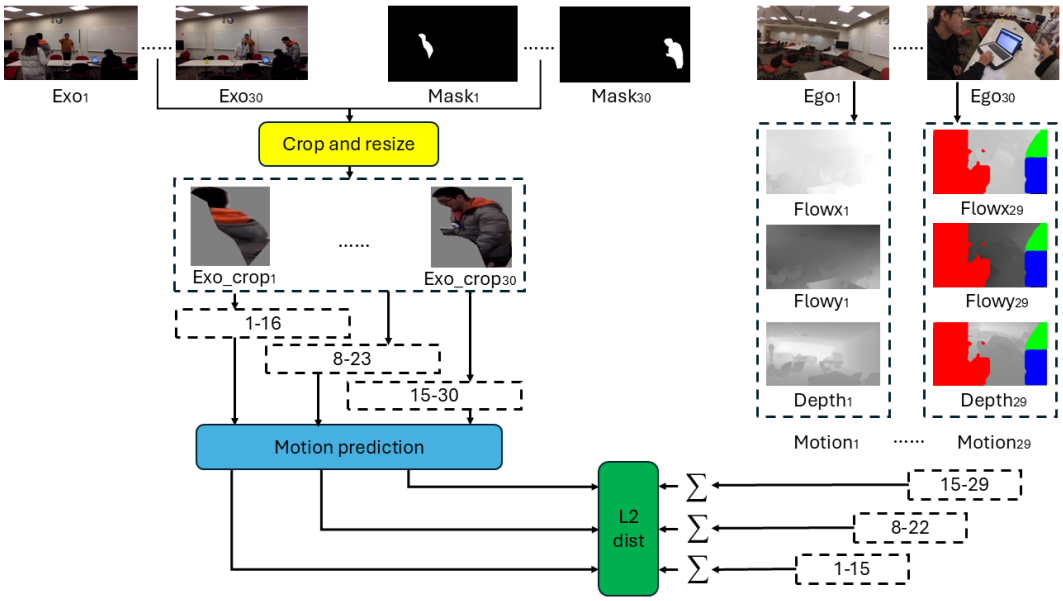


Figure 6. **Overview of the motion matching model.** We employ a sequence-based backbone (*MViT*) on the cropped third-person candidate and align it with the motion estimated from the first-person view. For visualization purposes, we rescale the optical flows such that black represents negative flow, white represents positive flow, and colored masks indicate foreground pixels excluded from motion estimation due to person detection. There are only 29 frames for first-person because optical flow requires 2 frames to calculate.

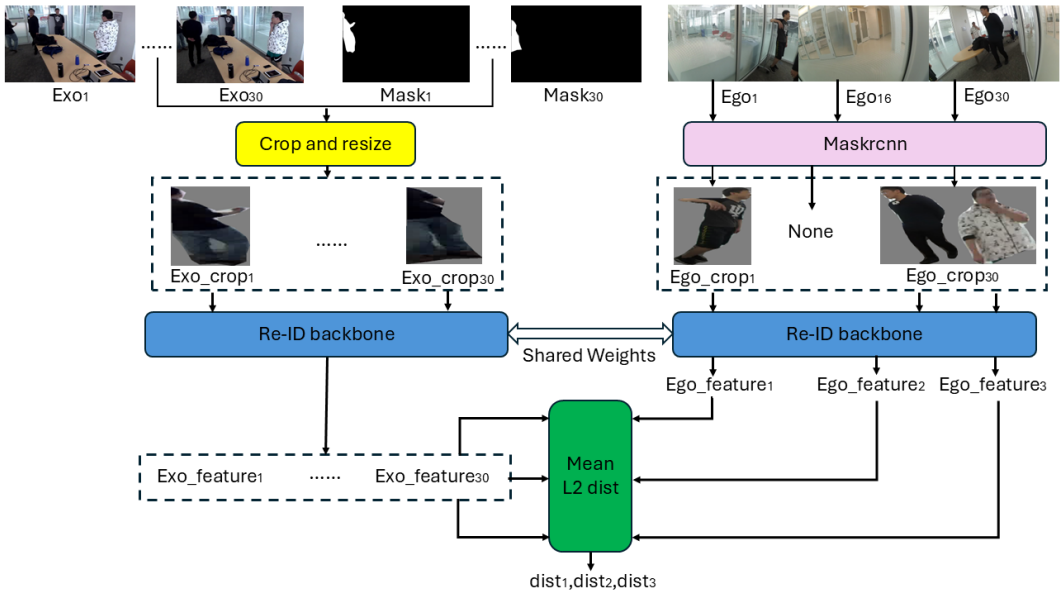


Figure 7. **Overview of appearance matching.** We apply *Mask R-CNN* on three selected frames from the first-person view and pass each of the M detected individuals, along with the third-person candidate, through the re-ID backbone. We then compute the average L2 distance between each detected individual and the third-person candidate across all 30 frames. This process results in M distance scores.

8. Examples of Fusing Algorithm

To demonstrate our fusion method (**CBAF**), we provide step-by-step examples in Figs. 8–11. For clarity in visualization, we set $\lambda_{\text{Re-ID}} = 1$ and $\alpha_{\text{mask}} = 1$ for each mask, enabling us to omit these terms during confidence calculations.

9. Annotation Details and Data Samples

To create **Ego4D-TF**, a subset of **Ego4D** annotated to support our task, we used the following annotation pipeline:

Candidates:	1	2	3	4
Motion scores:	12.4	22.3	1.2	5.8
Re-ID scores: (Frame 1)	None			
Re-ID scores: (Frame 16)	None			
Re-ID scores: (Frame 30)	None			

Figure 8. **Example 1**: No individuals are detected in the first-person view, leading to the prediction of candidate 3 (with the smallest motion score).

Candidates:	1	2	3	4	
Motion scores:	3.7	16.7	21.4	15.8	Conf.=15.8/3.7=4.3
Re-ID scores: (Frame 1)	12.4	11.6	5.8	14.7	Conf.=11.6/5.8=2
Re-ID scores: (Frame 16)	14.3	7.4	13.8	16.1	Conf.=13.8/7.4=1.9
Re-ID scores: (Frame 30)	11.9	9.2	12.3	17.5	Conf.=11.9/9.2=1.3

Figure 9. **Example 2**: Detected several individuals in first-person view, but motion has the highest confidence, predicting candidate 1.

First, we identified group activity videos in Ego4D containing at least three people, with at least two videos already synchronized. From these, we selected one video with a broad enough view to frequently capture at least two other people. We then sampled one out of every six frames to align the frame rate with **IUShareView** and **TF2023**, and applied YOLO-v11 [45] to 30-frame sequences to detect and track individuals. For sequences where more than two people were detected, we manually verified the masks of the first and last frames to ensure quality. Sequences were rejected if the mask quality was low, or the masks corresponded to different individuals in the first and last frames (lost tracking). The likelihood of losing tracking during a sequence was minimal, as all five groups of videos we used involved board games, where participants rarely made large movements. We show two examples of accepted annotations in Fig. 12 and two examples of rejected annotations in Fig. 13.

We also present sample data from the three source datasets used in **TF2025**: **TF2023**, **IUShareView**, and **Ego4D-TF**, illustrating the similarities and differences in their settings, as visualized in Fig. 14.

Candidates:	1	2	3	
Motion scores:	2.4	18.7	1.9	Conf.=2.4/1.9=1.3
Re-ID scores: (Frame 1)	11.3 17.1	11.7 6.7	3.6 10.7	Conf.=11.3/3.6=3.1 Conf.=10.7/6.7=1.6
Re-ID scores: (Frame 16)	16.5 14.0	4.2 13.3	19.4 9.8	Conf.=16.5/4.2=3.9 Conf.=13.3/9.8=1.6
Re-ID scores: (Frame 30)	11.4	9.9	13.1	Conf.=11.4/9.9=1.2

(a) Calculate confidence score for each source.

Candidates:	1	3	
Motion scores:	2.4	1.9	
Re-ID scores: (Frame 1)	11.3 17.1	3.6 10.7	
Re-ID scores: (Frame 16)	14.0	9.8	
Re-ID scores: (Frame 30)	11.4	13.1	

(b) Remove the appearance source with highest confidence, and the candidate with the smallest score.

Candidates:	1	3	
Motion scores:	2.4	1.9	Conf.=2.4/1.9=1.3
Re-ID scores: (Frame 1)	11.3 17.1	3.6 10.7	Conf.=11.3/3.6=3.1 **Conf.=17.1/10.7=1.6**
Re-ID scores: (Frame 16)	14.0	9.8	**Conf.=14.0/9.8=1.4**
Re-ID scores: (Frame 30)	11.4	13.1	**Conf.=13.1/11.4=1.1**

(c) Recalculate confidence scores. The updated scores are highlighted in **bold**.

Candidates:	1
Motion scores:	2.4
Re-ID scores: (Frame 1)	17.1
Re-ID scores: (Frame 16)	14.0
Re-ID scores: (Frame 30)	11.4

(d) Only 1 candidate 1 remaining, predicting candidate 1.

Figure 10. **Example 3**: Two candidates (1 and 3) have similar motion scores, a scenario that can occur when both candidates remain idle during the sequence. Our method resolves this by leveraging appearance information to eliminate candidates.

Candidates:	1	2	3	4	
Motion scores:	12.3	1.5	22.7	2.4	Conf.=2.4/1.5=1.6
Re-ID scores: (Frame 1)	None				
Re-ID scores: (Frame 16)	2.1 19.6	11.4 6.8	13.2 17.5	12.7 16.3	Conf.=11.4/2.1=5.4 Conf.=16.3/6.8=2.4
Re-ID scores: (Frame 30)	13.1 1.5	12.7 16.7	7.3 20.1	15.0 14.2	Conf.=12.7/7.3=1.7 Conf.=14.2/1.5=9.5

(a) Calculate confidence score for each source.

Candidates:	2	3	4
Motion scores:	1.5	22.7	2.4
Re-ID scores: (Frame 1)			
Re-ID scores: (Frame 16)	11.4 6.8	13.2 17.5	12.7 16.3
Re-ID scores: (Frame 30)	12.7	7.3	15.0

(b) Remove the appearance source with highest confidence, and the candidate with the smallest score.

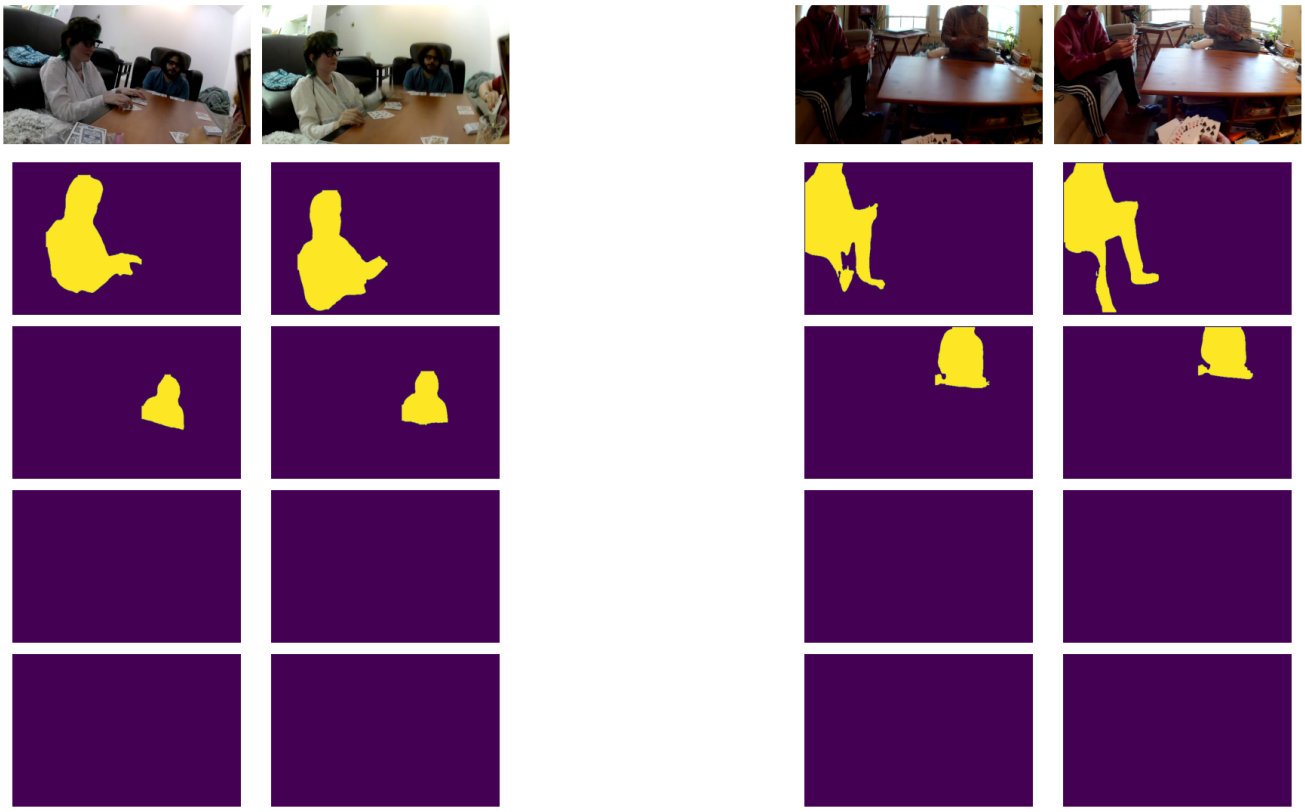
Candidates:	2	3	4	
Motion scores:	1.5	22.7	2.4	Conf.=2.4/1.5=1.6
Re-ID scores: (Frame 1)				
Re-ID scores: (Frame 16)	11.4 6.8	13.2 17.5	12.7 16.3	**Conf.=12.7/11.4=1.1** Conf.=16.3/6.8=2.4
Re-ID scores: (Frame 30)	12.7	7.3	15.0	Conf.=12.7/7.3=1.7

(c) Recalculate confidence scores. The updated scores are highlighted in **bold**.

Candidates:	3	4	
Motion scores:	22.7	2.4	**Conf.=22.7/2.4=9.5**
Re-ID scores: (Frame 1)			
Re-ID scores: (Frame 16)	17.5	16.3	**Conf.=17.5/16.3=1.1**
Re-ID scores: (Frame 30)	7.3	15.0	**Conf.=15.0/7.3=2.1**

(d) Motion score has the highest confidence, predicting candidate 4.

Figure 11. **Example 4**: A more complex scenario with two candidates (2 and 4) with similar motion scores. After eliminating two candidates, the updated motion scores exhibit the highest confidence, leading to the final prediction. This example demonstrates how our method adaptively selects which source to trust, effectively integrating information from both sources.

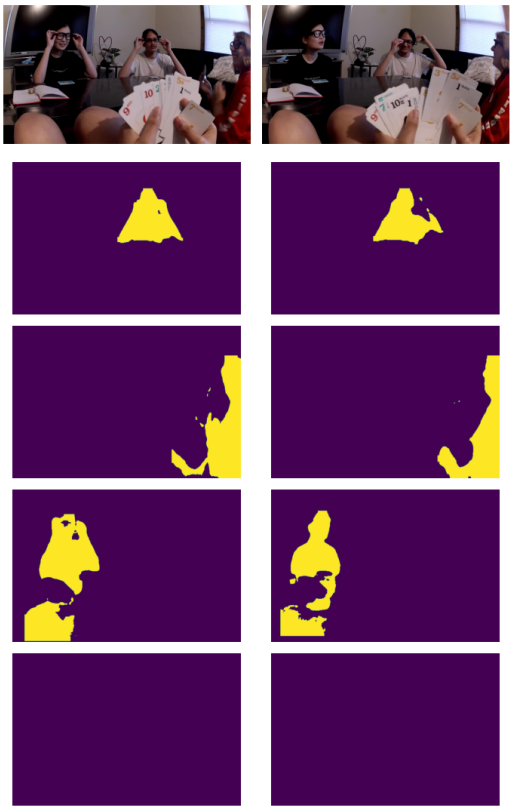


(a) Accepted annotation.

(b) Accepted annotation.)

Figure 12. Examples of accepted annotations for **Ego4D-TF**.

(a) Rejected due to tracking lost.



(b) Rejected due to low quality masks. (The individual to the left was merged with the legs of the camera wearer.)

Figure 13. Examples of rejected annotations for **Ego4D-TF**.

(a) **TF2023** and **IUShareView** (b) **Ego4D-TF**

Figure 14. **Data Comparison** We present sample data from **TF2023**, **IUShareView** (which share actors and settings), and our labeled subset of **Ego4D**, referred to as **Ego4D-TF**. A key difference between these datasets is that in **Ego4D-TF**, we treat first-person views as "pseudo" third-person views. Additionally, **Ego4D-TF** utilizes head-mounted cameras, whereas **TF2023** and **IUShareView** use chest-mounted cameras.