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

Supplementary Material

7. Additional Visualization of Models

We provide detailed visualization of our *motion matching* and *appearance matching* methods in Fig. 5 and Fig. 6.

8. 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

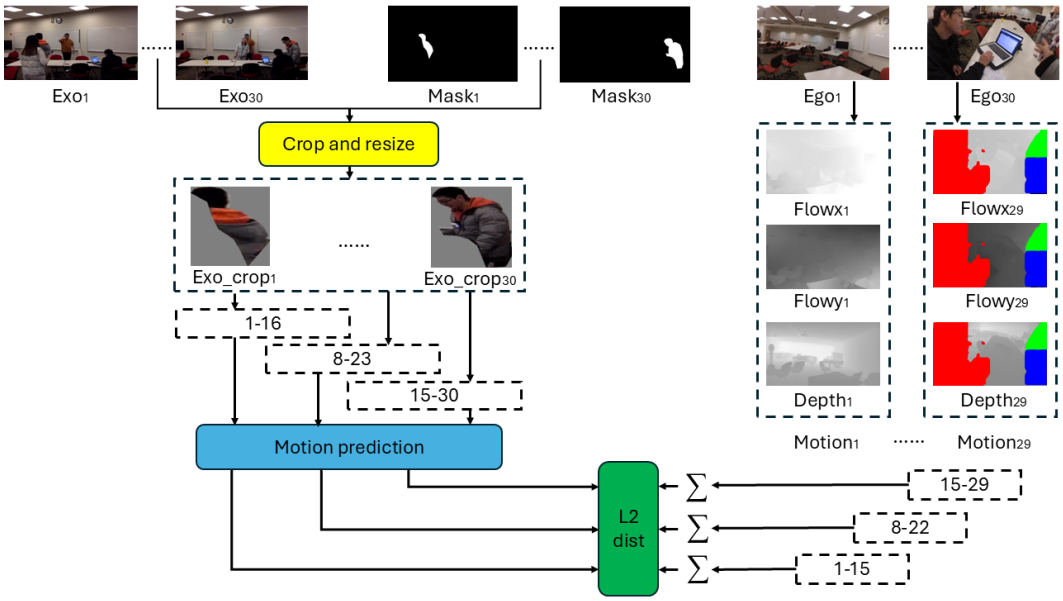


Figure 5. **Overview of the motion matching model.** We employ a sequence-based backbone (*MViT*) on the cropped third-person candidate and align it with the motion estimated from the first-person view. For visualization purposes, we rescale the optical flows such that black represents negative flow, 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.

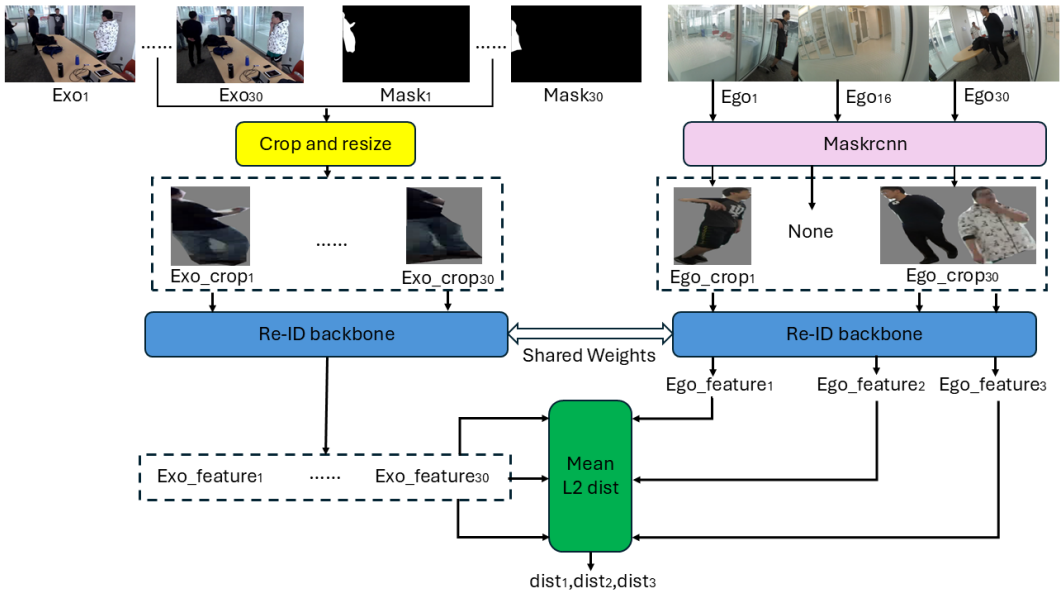


Figure 6. **Overview of appearance matching.** We apply *Mask R-CNN* on three selected frames from the first-person view and pass each of the M detected individuals, along with the third-person candidate, through the re-ID backbone. We then compute the average L2 distance between each detected individual and the third-person candidate across all 30 frames. This process results in M distance scores.

9. Examples of Fusing Algorithm

To demonstrate our fusion method (**CBAF**), we provide step-by-step examples in Figs. 7–10. 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.

10. 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 7. **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 8. **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. 11 and two examples of rejected annotations in Fig. 12.

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. 13.

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 9. **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.

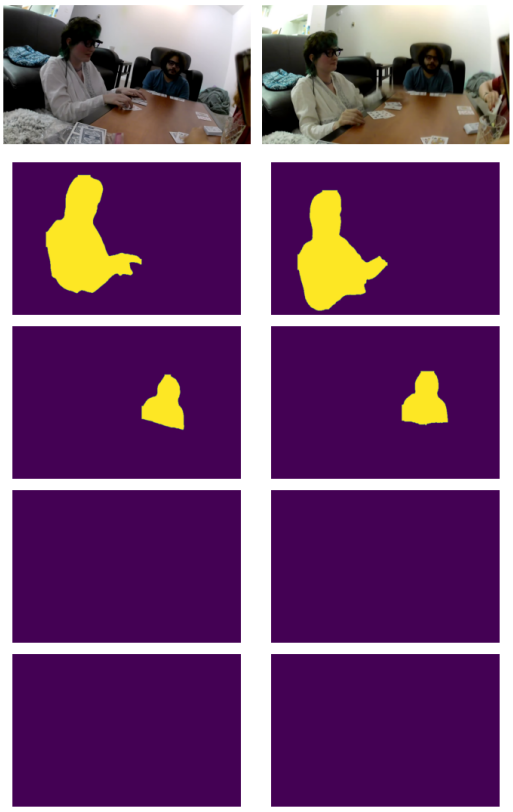
Candidates:	2	3	4	
Motion scores:	1.5	22.7	2.4	Conf.=2.4/1.5=1.6
Re-ID scores: (Frame 1)				
Re-ID scores: (Frame 16)	11.4 6.8	13.2 17.5	12.7 16.3	**Conf.=12.7/11.4=1.1** Conf.=16.3/6.8=2.4
Re-ID scores: (Frame 30)	12.7	7.3	15.0	Conf.=12.7/7.3=1.7

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

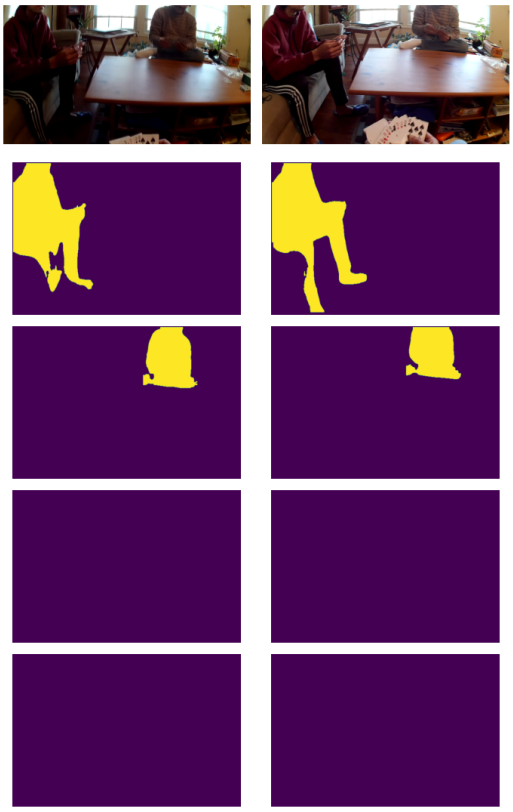
Candidates:	3	4	
Motion scores:	22.7	2.4	**Conf.=22.7/2.4=9.5**
Re-ID scores: (Frame 1)			
Re-ID scores: (Frame 16)	17.5	16.3	**Conf.=17.5/16.3=1.1**
Re-ID scores: (Frame 30)	7.3	15.0	**Conf.=15.0/7.3=2.1**

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

Figure 10. **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.



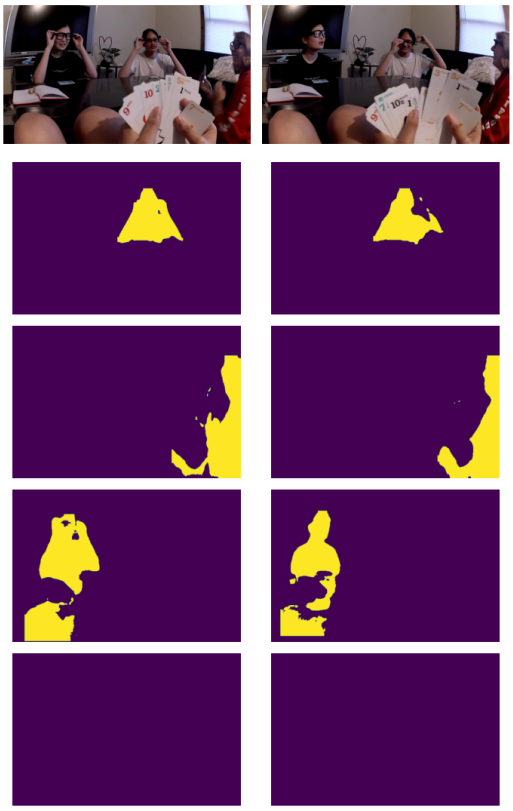
(a) Accepted annotation.



(b) Accepted annotation.)

Figure 11. 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 12. Examples of rejected annotations for **Ego4D-TF**.

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

Figure 13. **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.