Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities

Supplementary Material

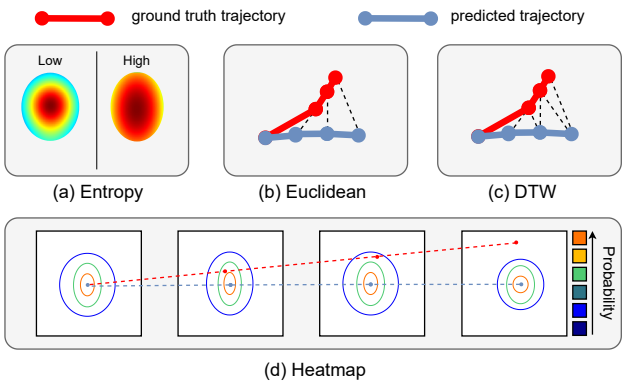


Figure 1. We consider three approaches to compare the ground truth with respect to the predicted trajectories in order to determine a mistake. (a) Entropy. (b) Euclidean distance between the two trajectories. (c) Dynamic Time Warping (DTW). (d) Average value of the ground truth trajectory at the predicted heatmaps.

A. Implementation details

Our model implementation is primarily based on the approach outlined in [3], with hyperparameters adjusted accordingly. Below, we elaborate on the key aspects of our implementation, highlighting the differences and specific choices made to enhance the model's performance.

Code Availability: To ensure reproducibility and provide further implementation details, we will share the complete codebase upon publication.

Stride Adjustment: In contrast to the stride of 8 used in [3], we opted for a stride of 1 during training. This modification allows the model to process consecutive frames without skipping, leading to finer granularity in temporal feature extraction. Our experiments indicated that this adjustment results in marginal improvements in both gaze estimation and mistake prediction accuracy.

Overfitting Prevention: To mitigate the risk of overfitting, we incorporated a weight decay parameter set to 0.07. This regularization technique helps in controlling the complexity of the model by penalizing large weights, thereby promoting generalization to unseen data.

Batch Size and Frame Processing: We configured the batch size to process 4 clips, each containing 8 frames. Specifically, our approach involves processing each video using non-overlapping windows of 8 consecutive frames. Consequently, each batch comprises 4 such windows, totaling 32 frames per batch (i.e., 8 frames/clip $\times$ 4 clips/batch $=$ 32 frames/batch). This setting ensures that the model captures sufficient temporal context while maintaining manageable memory usage.

Training Loss Following [3], we consider gaze prediction as defining a probability distribution over the 2D image plane of each input frame. Our proposed method leverages an architecture modified for gaze completion to predict missing segments of gaze trajectories. Ablations on single frames showed that looking at sequences of frames, which create a trajectory, is more effective for detecting anomalies due to the importance of changes over time. We train the model by minimizing the sum of the Kullback–Leibler divergence between the predicted gaze maps $\hat{P}(i)$ and the ground truth ones $Q(i)$ at each frame i:

$$L_{\mathrm{KL}}(\hat{P} \parallel Q) = \sum_i \hat{P}(i) \log \left(\frac{\hat{P}(i)}{Q(i)} \right) \tag{1}$$

Graphical Illustration of scoring functions Figure 1 illustrates the scoring function considered in this study. Entropy is the only scoring function which does not require any ground truth gaze as input, but only evaluates the level of uncertainty of the predicted heatmaps. The Euclidean and DTW scoring functions compute two forms of distances between predicted and ground truth trajectories. The heatmap scoring function evaluates the probability of predicted gaze under the points indicated by the ground truth trajectory. The heatmap scoring function achieves best results in our experiments. The main paper reports the formal definition of such scoring functions.

MoCoDAD Baseline Following methodologies from [2], we employ a sliding window approach to segment each agent's gaze/hands history. A window size of 8 frames is utilized, with the initial 4 frames dedicated to condition setting and the subsequent frames for the diffusion process. Hyperparameters are set as $\lambda_1 = \lambda_2 = 1$. Training proceeds end-to-end using the Adam optimizer with a learning rate of 1×10^{-4}, employing exponential decay over 25 epochs. The diffusion process utilizes $\beta_1 = 1 \times 10^{-4}$, $\beta_T = 2 \times 10^{-2}$ for $T = 10$, and incorporates the cosine variance scheduler.

TrajREC Baseline We followed the implementation proposed in *TrajREC* [5] official code release[1], adapting it for

[1] https://github.com/alexandrosstergiou/TrajREC

Method	Fusion	F1	Recall	Precision
Gaze Prediction	//	0.37	0.65	0.26
Gaze Completion	CH	0.38	0.67	0.29
Gaze Completion	CH + CORR	**0.40**	**0.70**	**0.31**

Table 1. Comparison of GLC and the proposed Gaze Completion approach for Gaze Estimation on EPIC-Tent.

gaze/hands trajectory analysis. The approach encodes temporally occluded gaze/hands trajectories, jointly learns latent representations of occluded segments, and reconstructs trajectories based on expected motions across different temporal segments.

For both methods, if a frame does not contain gaze or hand keypoints, we exclude that frame from the score calculation for the segment.

Action Type Classification To assess whether the type of performed action affects the performance of our method, we grouped actions contained in all three datasets into four categories: *Hand-Eye Coordination*, *Object Manipulation*, *Task Preparation*, and *Inspection/Verification*. For categorization, we prompted GPT-4 with a full list of actions using the following prompt:

In the context of how gaze affects actions, organize the following actions into groups that align with gaze literature. Group the actions from those that involve the most fine-grained gaze coordination to those that involve less gaze precision.

The list was then manually revised. The full classification is shown in Table 2

B. Additional Ablations

This section reports additional ablations which could not be included in the submitted paper due to space limits.

B.1. Performance Comparison Across Action Types

Table 3 compares the performance of the considered baselines and our proposed method across different action types.

The fully supervised C2F method achieves overall F1 and AUC scores of 0.58 and 0.72, respectively, maintaining stable performance across the four action types. The best performance is observed in *Inspect/Verify* actions (F1: 0.74, AUC: 0.85), likely due to the strong visual cues inherent to these tasks (e.g., instruction sheets).

In contrast, under both *One-Class* and *Unsupervised* scenarios, the proposed gaze-based approaches show varying performance depending on the action type. Stronger results are observed in tasks requiring *Hand-Eye Coordination* and *Object Manipulation* skills. Under the *One-Class* supervision level, our method achieves an F1 score of 0.741 and an AUC of 0.839 for hand-eye coordination tasks (+21% vs Overall AUC). This indicates its effectiveness in learning "normal" attention patterns and detecting mistakes during complex actions where gaze and motor coordination are crucial.

Conversely, for simpler actions, such as *Task Preparation* and *Inspect/Verify*, the proposed approaches are less effective. This is likely due to the high gaze variability inherent in less skill-intensive tasks. For instance, under the *One-Class* supervision level, our method achieves an F1 score of 0.257 and an AUC of 0.543 (-24% vs Overall AUC) for task preparation actions.

B.1.1. Performance of the proposed gaze completion approach vs standard gaze prediction

Results in main paper Table 1 compared the performance of the proposed mistake detection method based on gaze completion versus different methods, including a baseline method based on the standard gaze prediction task implemented with the method of [3]. In Table 1, we instead compare the performance of the proposed gaze completion approach with standard gaze prediction based on [3] on the EPIC-Tent dataset. Just using channel fusion brings a performance boost, achieving an F1-score of 0.38, recall of 0.67, and precision of 0.29, while combining channel and correlation fusion brings best results with an F1-score of 0.40, recall of 0.70, and precision of 0.31, suggesting that conditioning on partial trajectories makes gaze prediction less uncertain and the proposed approach can leverage the informative content provided by the input trajectory surpassing the performance of standard gaze prediction. Moreover, the performance in Table 1 correlates with the results in Table 1 of the main paper, suggesting that accurate gaze prediction enhances mistake detection performance. Specifically, the proposed approach excels in gaze prediction for "Correct execution" frames, although it loses accuracy for "Mistake" frames. Given that "Correct execution" frames are generally more frequent, the F1 score improves overall, but the gap in prediction accuracy between "Correct execution" and "Mistake" frames widens. This discrepancy, however, benefits trajectory-based comparisons in mistake detection, as the increased accuracy in "Correct execution" frames helps to better identify errors in subsequent frames.

B.2. Length of prediction and performance

Table 4 ablates performance for different prediction lengths. Smaller windows lead to higher precision due to short future trajectories being more predictable, but also lower recall, with the best F1 score when predicting 4 frames into the future.

As the prediction window extends from 1 to 4 frames, the model's recall improves, indicating more mistakes detected.

Category	Dataset	Actions
Hand-Eye Coordination	EpicTent	assemble, insert stake, insert support, insert support tab, tie top
	HoloAssist	touch, place, lift, press, flip, unscrew, rotate, slide, insert, close, turn, screw, disassemble
	IndustReal	fit, plug, tighten, loosen
Object Manipulation	EpicTent	spread tent, place guyline
	HoloAssist	adjust, empty, drop, clean, make, pour, split, mix-stir, stack-pile, load, mount, lock, unlock, shift, grab, pull
	IndustReal	put, take, pull
Task Preparation	EpicTent	pickup/open stakebag, pickup/open supportbag, pickup/open tentbag
	HoloAssist	withdraw, exchange, hold, break, approach, stand, align
	IndustReal	align
Inspection/Verification	EpicTent	instruction, place ventcover
	HoloAssist	inspect, validate, point, tap, click, push
	IndustReal	check, browse

Table 2. Classification of actions by category across datasets based on gaze involvement.

Method	Sup. Level	Overall F1	Overall AUC	Hand-Eye Coord.		Object Manip.		Task Prep.		Inspect/Verif.	
				F1	AUC	F1	AUC	F1	AUC	F1	AUC
Random	//	0.36	0.51	–	–	–	–	–	–	–	–
TimeSformer [1]	Fully Supervised	<u>0.49</u>	<u>0.67</u>	0.452	0.615	0.474	0.636	0.551	0.691	0.532	0.678
C2F [4]	Fully Supervised	**0.58**	**0.72**	0.506	0.600	0.5622	0.771	0.5138	0.686	0.741	0.857
GLC [3]	One-Class	0.46	0.66	0.524	0.704	0.495	0.665	0.425	0.579	0.396	0.556
Ours	One-Class	<u>0.52</u>	<u>0.69</u>	0.741	0.839	0.612	0.734	0.489	0.643	0.244	0.543
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.72**	0.753	0.872	0.631	0.764	0.498	0.657	0.257	0.543
GLC [3]	Unsupervised	0.44	0.61	0.542	0.694	0.474	0.657	0.406	0.563	0.338	0.526
Ours	Unsupervised	<u>0.51</u>	<u>0.69</u>	0.711	0.839	0.603	0.723	0.483	0.637	0.240	0.531
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.70**	0.714	0.862	0.602	0.754	0.489	0.646	0.253	0.535

* Late fusion

Table 3. Mistake detection results on EPIC-Tent by category. Best results are **in bold**, second best results are <u>underlined</u>.

Baseline	Future frames	F1	Precision	Recall
Gaze Completion	1	0.46	0.39	0.59
Gaze Completion	2	0.47	0.38	0.62
Gaze Completion	3	0.47	0.37	0.63
Gaze Completion	4	0.49	0.34	0.88

Table 4. Performance ablation for different prediction lengths. Smaller windows yield higher precision but lower recall. The best F1 score is achieved when predicting 4 frames into the future.

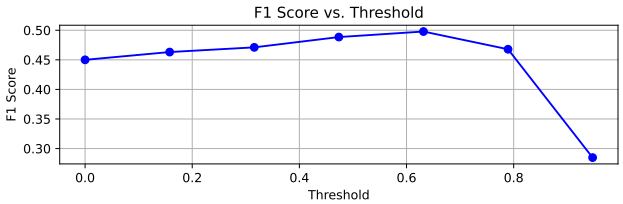


Figure 2. Length of prediction.

However, this is offset by a reduction in precision, leading to more false positives.

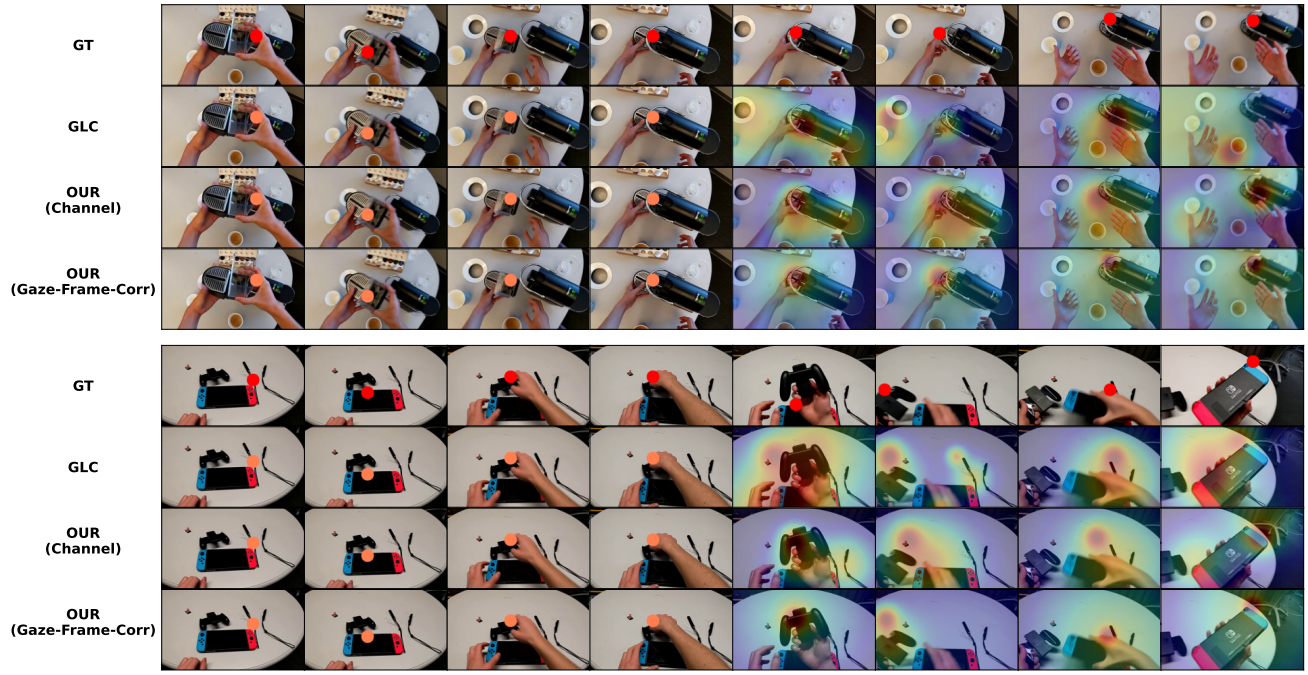
B.3. Chosen thresholds and sensitivity

We report F1 scores obtained at each method's optimal thresholds, which we'll report in the paper. Figure 2 shows how the F1 score of our best method (*Unsupervised - Ours*, Table 2 of main paper) changes when varying the threshold. Performance is stable for a range of threshold values.

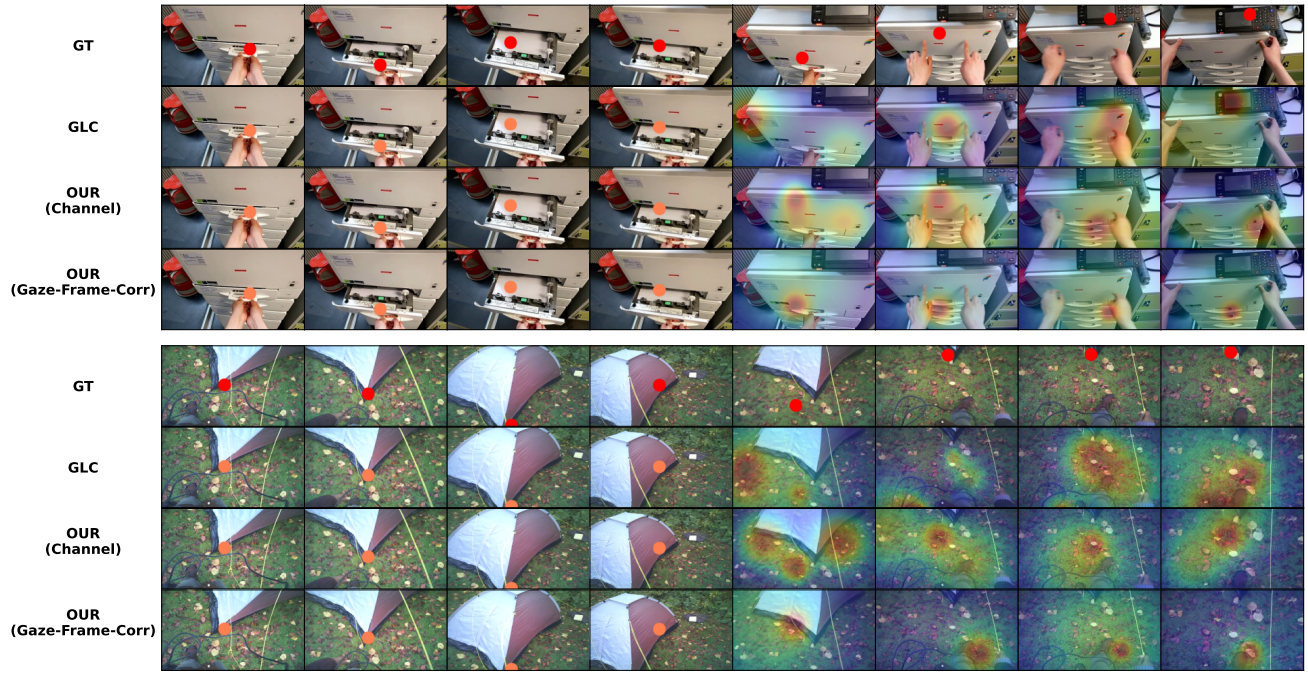
B.4. Qualitative Results and Failure Cases

Figure 3 illustrates the performance of various baselines and the proposed approach for both correct predictions in *Correct Execution* cases (a) and in *Incorrect Execution* cases (b). The top row displays the ground truth, followed by predictions from the GLC method, our proposed "Channel" approach, and, at the bottom, the "Gaze Frame Correlation" approach. The last four columns display the predicted heatmap, where red peaks symbolize the 2D gaze predicted points.

Figure 3a focuses on correct predictions related to *Correct Execution*. The first row shows the actual gaze coordinates. Notably, in the second row corresponding to GLC, the predicted heatmaps exhibit inconsistencies, with



(a) Correct prediction of Correct action.



(b) Correct prediction of Mistake Action.

Figure 3. Qualitative examples. The first four columns represent the inputs (with the input gaze 2D points highlighted in orange). The latter four columns show the predicted outputs in the form of heatmaps.

varying peaks across consecutive frames. In contrast, our proposed method leverages temporal information to produce temporally consistent predictions. The "Channel" approach demonstrates better consistency than GLC, while the "Gaze Frame Correlation" method generates more defined heatmaps with fewer, more localized peaks around the gaze region. In this case, a *Correct Execution* is identified based on the small gap between the ground truth and the predicted gaze trajectory.

Figure 3b highlights predictions related to *Incorrect Execution*. Here, our approach's gaze predictions diverge from the ground truth, effectively flagging mistakes in action execution.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. 3

[2] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10318–10329, 2023. 1

[3] Bolin Lai, Miao Liu, Fiona Ryan, and James Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *British Machine Vision Conference*, 2022. 1, 2, 3

[4] Dipika Singhania, Rahul Rahaman, and Angela Yao. C2f-tcn: A framework for semi-and fully-supervised temporal action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11484–11501, 2023. 3

[5] Alexandros Stergiou, Brent De Weerdt, and Nikos Deligiannis. Holistic representation learning for multitask trajectory anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 1