DETACH : Decomposed Spatio-Temporal Alignment for Exocentric Video and Ambient Sensors with Staged Learning

Supplementary Material

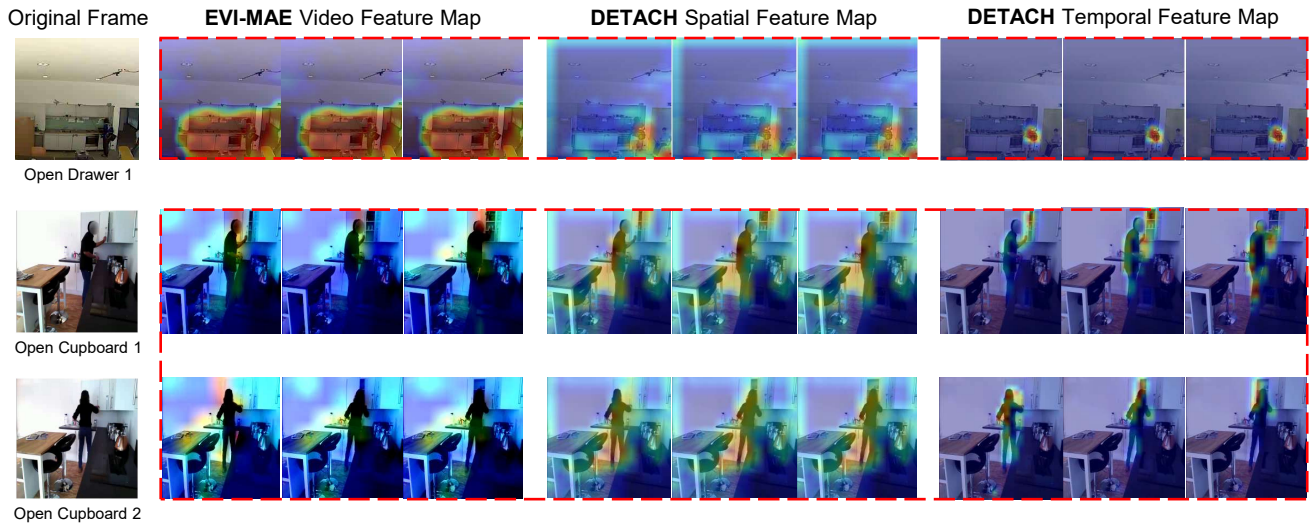


Figure 7. **Qualitative visualization of decomposed spatial-temporal features.** We visualize the sequence of frames to illustrate the temporal flow. This visualization validates DETACH's effectiveness in two aspects: (1) capturing subtle motion patterns that are diluted in the baseline (Row 1), and (2) distinguishing similar actions by explicitly decomposing specific spatial contexts from shared temporal patterns (Row 2 and Row 3).

A. Additional Qualitative Analysis

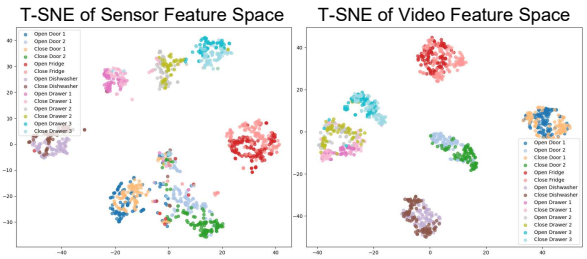
Analysis on Feature Decomposition. Fig. 7 visualizes the feature maps over the temporal frame sequence to qualitatively validate the effectiveness of DETACH in addressing the two fundamental limitations of Global Alignment discussed in the main text: (P1) the inability to capture local details and (P2) the over-reliance on modality-invariant temporal patterns.

Row 1 demonstrates the capability to capture fine-grained local details, specifically subtle motion patterns. As clearly observed in the EVI-MAE [16] column, the feature maps are broadly and diffusely activated, extending well beyond the subject's body into surrounding regions such as the ceiling and background structures, without precisely localizing the interaction area. This visual evidence reflects the limitation of Global Alignment where compressing the video into a single vector causes minimal visual changes to be diluted against the dominant static scenes. In contrast, DETACH decomposes spatial context from temporal motion and thereby preserves these subtle cues. As a result, it precisely activates the specific regions corresponding to the interaction.

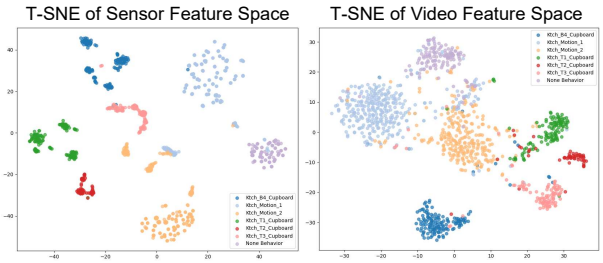
Row 2 and Row 3 compare two actions that share similar temporal patterns but involve different spatial contexts. The EVI-MAE [16]'s feature maps for both rows focus primarily on dominant temporal patterns such as the human body rather than distinct spatial contexts. This observation visualizes the over-reliance on modality-invariant features which prevents the baseline from distinguishing actions with such similar dynamics. Conversely, DETACH successfully decomposes the representation. Its spatial feature map explicitly captures distinct spatial interaction location, while the temporal feature map independently handles the temporal dynamics. This decomposition effectively resolves the ambiguity that confounds the baseline model.

Visualization of Video and Sensor Feature Spaces. We visualize the feature spaces of both modalities using t-SNE. Fig. 8(a) presents the results on the Opportunity++ dataset [6], where the visualizations consistently reveal a context-dependent structure across both modalities. At a global level, samples are primarily grouped by spatial context. Within these spatial clusters, fine-grained differentiation emerges based on temporal context, where distinct actions form well-separated sub-clusters. This confirms our model successfully encodes discriminative temporal features within the relevant spatial context for both modalities.

To complement this analysis, we further present the t-SNE visualization of the spatial features for the HWU-USP



(a) **t-SNE visualization of video and sensor feature spaces on Opportunity++**. Samples are primarily grouped by spatial context at the global level, with fine-grained action-based sub-clusters emerging within each spatial group.



(b) **t-SNE visualization of video and sensor feature spaces on HWU-USP**. Samples are grouped by spatial context for both modalities.

Figure 8. **Visualization of Video and Sensor Feature Spaces.**

dataset [12] in Fig. 8(b). It is important to note the structural difference between the two datasets regarding label granularity. While Opportunity++ provides mid-level labels that allow for analyzing atomic actions within a specific context such as distinguishing opening a door from closing, the HWU-USP contains only high-level activity labels. Consequently, demonstrating the separation of atomic actions is not feasible for this dataset. Instead, Fig. 8(b) visualizes the feature space at the spatial context level for both video and sensor modalities. The formation of cohesive clusters corresponding to underlying spatial contexts confirms that DETACH effectively learns discriminative spatial representations.

B. Additional Ablation Studies

Ablation on Sensor Encoder Capacity. To investigate whether the performance improvements of DETACH stem from the proposed alignment framework or the specific encoder architecture, we evaluated our method using different sensor backbones:

- **Lightweight:** A simple 1D-CNN followed by a GRU, representing previous egocentric multimodal studies [7, 11] with limited capacity.
- **Heavyweight:** A Transformer encoder with increased depth and parameters, designed to maximize representational capacity for multivariate sensor data, following recent approaches in time-series modeling [14, 15].

Table 4. **Robustness analysis across different sensor encoder architectures.** We report the performance using three different levels of sensor encoders on both datasets.

Sensor Encoder	Opportunity++		HWU-USP	
	F1 (Weighted)	mAP	F1 (Weighted)	mAP
A. Lightweight (1D CNN + GRU)	0.62	0.73	0.71	**0.67**
B. Heavyweight (Transformer)	0.71	0.85	0.71	0.66
C. DETACH (1D CNN + GRU + Attn.)	**0.73**	**0.87**	**0.73**	**0.67**

Table 5. **Performance comparison across different video encoder architectures.** We report the performance using four different video encoders on both datasets.

Video Encoder	Opportunity++		HWU-USP	
	F1 (Weighted)	mAP	F1 (Weighted)	mAP
A. ViViT	0.58	0.65	0.68	0.61
B. VideoMAE	0.64	0.73	**0.71**	0.60
C. MViT	0.62	0.72	0.70	**0.62**
D. 3D CNN (Ours)	**0.70**	**0.82**	**0.71**	**0.62**

Table 6. **Ablation on the momentum encoder used for calculating W_{ij}^{temporal}.**

Method	Opportunity++		HWU-USP	
	F1 (Weighted)	mAP	F1 (Weighted)	mAP
w/o Mom. Enc.	0.59	0.72	0.65	0.61
w Mom. Enc. (Ours)	**0.73**	**0.87**	**0.73**	**0.67**

- **DETACH:** Our default architecture, incorporating self-attention mechanism to capture spatio-temporal dependencies.

As shown in Tab. 4, the Lightweight encoder outperforms all state-of-the-art baselines reported in the main text (Tab. 1), regardless of their backbone complexity. This result provides compelling evidence that the substantial performance gains are primarily driven by our novel decomposed spatio-temporal alignment strategy rather than model capacity. Furthermore, even when the heavyweight Transformer yields higher performance, the improvement over our default encoder is marginal. This suggests that our chosen architecture achieves an optimal balance between performance and efficiency, making it a highly effective and practical choice for ambient sensing environments without the computational cost of Transformers.

Impact of Video Backbone on Performance. To justify our choice of the video backbone, we evaluate the proposed DETACH, a 3D CNN-based architecture, against three representative Transformer-based encoders: ViViT [1], VideoMAE [13], and MViT [9]. For a fair comparison, all models were trained under identical experimental settings, including a unified batch size of 96. As shown in Tab. 5, 3D CNN significantly outperforms the Transformer-based encoders across both datasets. We attribute this to the inherent spatio-temporal inductive bias of 3D CNN, which captures

Table 7. **Ablation on the confidence threshold for selecting confident samples in the clustering stage.**

Threshold	Opportunity++		HWU-USP	
	F1 (Weighted)	mAP	F1 (Weighted)	mAP
50%	0.56	0.68	0.64	0.60
75% (Ours)	**0.73**	**0.87**	**0.73**	**0.67**
100%	0.53	0.71	0.60	0.63

Figure 9. **Stabilization of cluster assignments in Stage 1.** The pseudo-label change rate decreases and plateaus, indicating the convergence of the spatial encoder. The transient spike observed in the middle corresponds to the activation of the refinement loss (L_{refine}), which triggers a re-adjustment of cluster assignments. Subsequently, the rate stabilizes close to zero, serving as the stopping criterion for our unsupervised training protocol.

subtle motion cues more effectively than Transformers in data-limited regimes [8].

Necessity of the Momentum Encoder. Tab. 6 examines the necessity of the momentum encoder for computing W_{ij}^{temporal}. Directly using the training temporal encoder causes significant performance drops due to unstable similarity estimations. The momentum encoder provides stable feature representations, enabling reliable temporal similarity measurement and effective false negative mitigation through precise W_{ij}^{temporal} weighting.

Importance of Confidence-based Filtering. Tab. 7 highlights the critical role of confidence thresholding. Utilizing all samples (100%) causes drastic degradation on both datasets, indicating that indiscriminate usage of samples injects noisy supervision. Our strategy effectively filters these ambiguous predictions, ensuring that only high-confidence pseudo-labels drive the cross-modal alignment.

Sensitivity to λ_{hard} value. Tab. 8 highlights the effect of λ_{hard} on downstream performance. When λ_{hard} is too small, hard negatives do not contribute sufficiently to the contrastive loss, limiting representation learning. Conversely, when λ_{hard} is too large, the excessive weighting on hard negatives leads to unstable training and degrades performance. We find that $\lambda_{hard} = 3.0$ achieves the best balance, consistently outperforming other values across both datasets.

Table 8. **Ablation on the λ_{hard} value.**

λ_{hard}	Opportunity++		HWU-USP	
	F1 (Weighted)	mAP	F1 (Weighted)	mAP
2.0	0.67	0.74	**0.73**	0.65
3.0 (Ours)	**0.73**	**0.87**	**0.73**	**0.67**
4.0	0.68	0.73	0.72	0.65
5.0	0.64	0.66	0.65	0.62

C. Implementation Details

To ensure reproducibility, we provide a detailed breakdown of the training strategy for DETACH. Our framework is trained in two sequential stages, each employing a scheduling strategy suited to its specific objective.

Two-Stage Training Schedule. Since the convergence rate of online clustering varies with dataset distribution and complexity [2, 3], we adopted an adaptive training schedule for Stage 1 rather than a fixed epoch limit. We trained the spatial encoders until cluster assignments stabilized, which typically occurred around 25 epochs for Opportunity++ and 22 epochs for HWU-USP. These total epochs include the joint learning phase (10 epochs) described in the main text. This difference in convergence speed reflects the distinct complexities and distributions of the two datasets.

For Stage 2, we used a fixed schedule of 50 epochs. As this stage focuses on optimizing our proposed Spatial-Temporal Weighted Contrastive Loss, a fixed budget ensures consistent evaluation of the alignment process, following standard evaluation protocols in contrastive learning [5]. Thus, "Stage2 - Epoch 0" in Fig. 4 represents the state immediately after Stage 1 completion, where spatial features are already discriminative from Stage 1 convergence, while temporal features are initialized but not yet aligned.

Stopping Criterion for Stage 1. To maintain strict adherence to the unsupervised protocol, we did not use ground-truth labels to determine when to stop Stage 1. Instead, we monitored the stability of cluster assignments by tracking the rate of change in pseudo-labels, defined as the fraction of samples that switched their predicted cluster assignments between consecutive epochs [2]. As shown in Fig. 9, training was terminated when this rate reached a plateau close to zero, indicating that the model had learned discriminative spatial representations and the cluster assignments had converged.

Baseline Implementation Details. In the original CO-MODO [4] architecture, the video encoder remains frozen, and a FIFO queue stores video embeddings to compute the target video-video similarities. However, as we unfreeze the video encoder for end-to-end fine-tuning to ensure a fair comparison, directly updating the queue with embeddings from a rapidly updating encoder causes severe representation shift and training instability. To mitigate this, inspired

Table 9. **Detailed statistics and configurations of the datasets.**

Attribute	Opportunity++ [6]	HWU-USP [12]
Sensor Configuration		
Sensor Types	Logic Switches (×13) Accelerometers (×7)	Logic Switches (×4) Logic Motion Sensors(×2)
Sensor Loc.	Doors, Drawers, Fridge, Dishwasher	Cupboards, Wall
Annotation Details		
Label Level	Mid-level	High-level
Total Classes	14	5
Class Names	Open/Close Door 1,2 Open/Close Drawer 1-3 Open/Close Fridge Open/Close Dishwasher	Cereals Tea Sandwich Dishes Tidy
Experimental Setup		
Subjects	Single subject	16 subjects (one per video)
Total Samples	5324 windows	3485 windows

by the Momentum Contrast (MoCo) framework [10], we incorporate a momentum update strategy. Specifically, we introduce a momentum video encoder whose parameters are updated via an exponential moving average (EMA) of the trainable video encoder. The similarity queue is then strictly updated using the embeddings generated by this momentum encoder. This design ensures a consistent and stable target distribution for the contrastive loss while allowing the main video encoder to learn robust, task-specific features.

D. Dataset Details

Opportunity++. As detailed in Table 9, Opportunity++ focuses on mid-level actions with short durations, such as "Open Door 1" or "Close Fridge". The dataset employs a hybrid sensor configuration of logic switches and accelerometers, requiring the model to align video representations with both discrete state transitions and continuous temporal dynamics. Data collected from a single subject provides a controlled environment for evaluating fine-grained alignment capabilities.

HWU-USP. HWU-USP targets high-level daily activities with longer temporal contexts, such as "Making a Sandwich" or "Tidy Up". With 16 different subjects, this dataset introduces intra-class variance in action execution. The sensor combines logic switches and logic motion sensors to capture object interactions and spatial occupancy. This multi-user setting evaluates the model's generalization ability and robustness in complex action recognition scenarios.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[4] Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora Salim. Comodo: Cross-modal video-to-imu distillation for efficient egocentric human activity recognition. *arXiv preprint arXiv:2503.07259*, 2025.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[6] Mathias Ciliberto, Vitor Fortes Rey, Alberto Calatroni, Paul Lukowicz, and Daniel Roggen. Opportunity++: A multimodal dataset for video-and wearable, object and ambient sensors-based human activity recognition. *Frontiers in Computer Science*, 3:792065, 2021.

[7] Arnav M Das, Chi Ian Tang, Fahim Kawsar, and Mohammad Malekzadeh. Primus: Pretraining imu encoders with multimodal self-supervision. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[8] 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. In *International Conference on Learning Representations*.

[9] 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.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

[12] Caetano Mazzoni Ranieri, Scott MacLeod, Mauro Dragone, Patricia Amancio Vargas, and Roseli Aparecida Francelin Romero. Activity recognition for ambient assisted living with videos, inertial units and ambient sensors. *Sensors*, 21 (3):768, 2021.

[13] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[14] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 220–233, 2021.

[15] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124, 2021.

[16] Mingfang Zhang, Yifei Huang, Ruicong Liu, and Yoichi Sato. Masked video and body-worn imu autoencoder for egocentric action recognition. In *European Conference on Computer Vision*, pages 312–330. Springer, 2024.