Recursive prototyping for computational behavioral analysis from egocentric videos

Sam Perochon[1 [0000−0002−1759−4154]] and Laurent Oudre[1 [0000−0002−4750−2265]]

Université Paris Saclay, Université Paris Cité, ENS Paris Saclay, CNRS, SSA, INSERM, Centre Borelli, F-91190, Gif-sur-Yvette, France

Abstract. Large-scale egocentric video datasets captured during standardized procedural tasks offer unique opportunities for fine-grained behavioral analysis, yet extracting structured representations from these recordings without exhaustive annotations remains challenging. We propose a data-driven framework that leverages the temporal redundancy of repeated task executions to derive compact and interpretable symbolic representations of participants' visual experience. Actions are discovered by iteratively refining latent subspaces using minimal human feedback to ascertain their quality and recover their interpretability. Egocentric vision prototypes are then consolidated via hierarchical clustering, effectively reducing redundancy and improving robustness. The symbolic representations combine prototype vocabulary with an exact temporal segmentation estimate at a fine, globally defined granularity. Without relying on costly ground-truth annotations, we demonstrate that the symbolization procedure generalizes across participants and remains robust to annotator subjectivity. The resulting representations enable cross-subject behavioral comparison and multimodal integration.

Keywords: Computational behavior analysis · Egocentric vision · Procedural activity analysis · Unsupervised action segmentation · Prototype learning · Computational ethology

1 Introduction

Systematic recording of egocentric vision during procedural tasks offers an unprecedented opportunity to characterize behavior more accurately and objectively at fine granularity, holding promise for clinical, pedagogical, or industrial applications [8]. However, despite recent advances in machine learning (ML) and computer vision (CV), the automatic and accurate quantification of complex human behavior during naturalistic tasks remains an open challenge [8].

Computational ethology aims to fill this gap by developing automatic and scalable methods to quantify behavior in naturalistic settings [23]. A central hypothesis formulates that behavior comprises identifiable sequences of stereotyped action modules, often standardized using ethograms [31].

While valuable, traditional ethograms and supervised human action recognition (HAR) benchmarks [7,18] face several limitations: (i) subjectively pre-

defining action categories; (ii) time-consuming manual annotations; (iii) imperfect inter-rater reliability; and (iv) the number, granularity, and temporal resolution of categories is constrained [26,8]. These issues render ethograms difficult to scale to complex procedural tasks spanning multiple interconnected steps. In contrast, unsupervised methods aim to automatically identify and extract stereotyped behavioral atoms from large-scale recordings, thereby alleviating the need for manual annotations and promoting more objective characterization of behavior [8].

Egocentric sensing technologies (eye-trackers, head-mounted cameras) provide fine-grained, participant-centered information—including occlusion-free hand-object interactions and first-person perspectives that disambiguate actions [24,19]. In this work, we propose a robust and generalizable methodology to encode participants' visual experience during a procedural task into symbolic representations. Specifically, we model visual experiences as sequences of fine-grained visual atoms, where each atom denotes an interpretable and unequivocal unique action. Our hypothesis is that the redundancy and temporal regularity of short segments of egocentric vision during a standardized task can be exploited to build a common vocabulary of vision prototypes. We develop and evaluate this framework on a clinical dataset of 162 participants performing a standardized cooking task, demonstrating both methodological validity and practical utility for behavioral analysis. Code is available at `https://github.com/samperochon/recursive-prototyping`.

2 Related Work

Unsupervised temporal action segmentation. Unsupervised temporal action segmentation (uTAS) aims to partition untrimmed videos into coherent action segments without labels. Kukleva et al. learned continuous temporal embeddings using self-supervised consistency before clustering [16]. More recent methods integrate segmentation into the optimization, such as ASOT [35] which formulates segmentation as an optimal transport problem between frames and action classes, while HVQ [29] introduces discrete hierarchical codebooks learned end-to-end. Unlike fully unsupervised approaches, we incorporate minimal human feedback during prototype accumulation to recover the semantics of the action classes and ascertain their quality.

Understanding of egocentric recording during procedural tasks. Large-scale egocentric benchmarks such as Ego4D Goal-Step [28] and EgoExoLearn [13] have been proposed to drive progress through hierarchical procedural annotations spanning diverse activities. Procedure learning from egocentric video has been addressed via self-supervised temporal consistency [2], optimal transport formulations [21], and progress-aware online segmentation [27]. These approaches require dense supervision or target multi-environment settings with substantial visual variability. In contrast, we focus on single-environment and prescribed tasks where visual redundancy across participants can be exploited without ground-truth labels.

Symbolic representations of behavior. Computational ethology seeks to extract discrete, interpretable behavioral primitives from continuous observations [10,8]. In animal studies, pose trajectories obtained via deep learning–based tracking [25] are typically segmented using generative models such as Keypoint-MoSeq [34] or variational autoencoders with recurrent architectures [20], followed by clustering to identify stereotyped behavioral motifs. Applications of these frameworks to characterize behavioral alterations in psychiatric and neurological conditions remain at an early stage. In contrast with these approaches primarily relying on body kinematics, therefore omitting the rich contextual and multimodal cues (objects, gaze, environment) that encompass behavior, we seek to extend this paradigm to egocentric vision.

3 Symbolization procedure

Given an egocentric video $V = (v_1, \ldots, v_N)$ of N frames, our goal is to produce a *symbolic representation* $\mathbf{s} = (s_1, \ldots, s_n)$, where each symbol $s_i \in \mathcal{A} \cup \{-1\}$ belongs to a finite alphabet $\mathcal{A}$ of action prototypes (with -1 denoting a background category). A *prototype* is defined as an interpretable and unequivocal visual pattern representing a class of semantically consistent visual experiences.

The procedure consists of two independent modules: (i) egocentric vision prototyping to discover action classes, and (ii) unsupervised temporal segmentation to estimate robust action boundaries at fine granularity.

Egocentric videos are first encoded using a frozen VideoMAE-v2 model [33]. We apply it to overlapping 16-frame windows with stride 8, producing a sequence of D-dimensional latent embeddings $\mathbf{x} = \{x_i\}_{i=1}^{n} \in \mathbb{R}^{D \times n}$, where each embedding summarizes approximately 640 ms of video, and $D = 1408$. The goal of the prototyping step is to construct a compact and interpretable vocabulary $\mathcal{V}$ summarizing the egocentric vision data by clustering latent video representations into semantically consistent regions. Ideally, we aim for a sparse and exhaustive vocabulary, generalizable across participants.

Direct clustering of the full latent space ($N \approx 10^6$, $D = 1408$, $K \approx 10^2$) proved unreliable due to the emergence of clusters at heterogeneous semantic scales, reflecting a known limitation in extreme clustering regimes [15]. While many clusters correspond to well-defined unique actions, others systematically mix multiple visual primitives, regardless of clustering method, dimensionality reduction and normalization strategies. Additional connections between this observation and the empirical pairwise similarity matrices of latent embedding sequences are derived in Supplementary Material (SM) §1. To address this issue, we propose a recursive prototyping strategy that progressively refines under-segmented regions of the encoder latent space. The method alternates between (i) accumulating semantically unequivocal clusters, and (ii) refining under-segmented regions associated with ambiguous clusters. This recursive latent subspace exploration is guided by minimal human feedback to ascertain cluster quality and recover their interpretability.

3.1 Prototyping procedure using recursive latent subspace exploration guided by human feedbacks

The procedure operates in two stages: the centroids accumulation incorporating the human feedbacks, and the centroids' consolidation phase to improve the vocabulary sparsity and reduce its redundancy.

Centroids accumulation. Given training embeddings $X_{\text{train}} = \{x_i\}_{i=1}^{N}$, we iteratively estimate candidate centroids using cosine k-means clustering applied to a residual set $R_p \subseteq X_{\text{train}}$, initialized as $R_1 = X_{\text{train}}$. At iteration p, C centroids are obtained by minimizing the total within-cluster cosine dissimilarity:

$$\mathcal{L}_{\cos}(\mu, R_p) = \sum_{x_i \in R_p} \left(1 - \frac{x_i^\top \mu_{c_i}}{\|x_i\| \, \|\mu_{c_i}\|}\right),$$

where $c_i = \arg\min_c d(x_i, \mu_c)$ denotes the nearest centroid assignment of x_i under cosine distance. Each cluster is then reviewed by annotators and categorized as either *unequivocal* or *ambiguous*, using the annotation protocol described below. Centroids associated with unequivocal clusters are accumulated into an accumulator set, while samples (i) assigned to ambiguous clusters or (ii) weakly assigned to accepted centroids define the residual set for the next iteration. This mechanism, illustrated in Figure 1, focuses subsequent clustering steps on poorly represented regions of the latent space. Weakly associated samples are defined by a cosine distance to their closest centroid above a fixed threshold $d_{\min} = 0.3$, corresponding approximately to the first quartile of pairwise cosine distances in the latent space. The process is repeated until a target number of candidate prototypes is reached, to form the centroids vocabulary $\mu^{[K]} = \{\mu_i\}_{i=1}^{K}$. Additional details are provided in SM §2, and an analysis of residual set thresholding strategies is proposed in SM §3.

Annotation protocol for egocentric vision prototypes. To assess the semantic interpretability of the learned clusters, we introduce a *Visual Probing Mechanism* (VPM) to identify semantically consistent clusters and filter out ambiguous ones. At each iteration p, the C candidate centroids $\{\mu_c^p\}_{c=1}^{C}$ are obtained by k-means, minimizing within-cluster cosine distance. For each sample x_i, we compute its assignment distance to the centroid of its assigned cluster, $d(x_i, \mu_{c_i}^p)$, where $c_i = \arg\min_{c \in [C]} d(x_i, \mu_c^p)$. These distances are used to rank samples within each cluster.

Our VPM relies on the following principles: (i) cluster semantics can be inferred from samples closest to the centroid; (ii) within-cluster semantic consistency can be assessed from the most distant assigned samples; and (iii) the central frame of each input sequence ($\approx 640ms$) is sufficient for visual inspection. In practice, inspecting the nearest and furthest samples from 30 participants per cluster was sufficient for reliable annotation. Representative examples of annotation probes are shown in Figure 2-a, and detailed validation of these hypothesis are reported in SM §4.

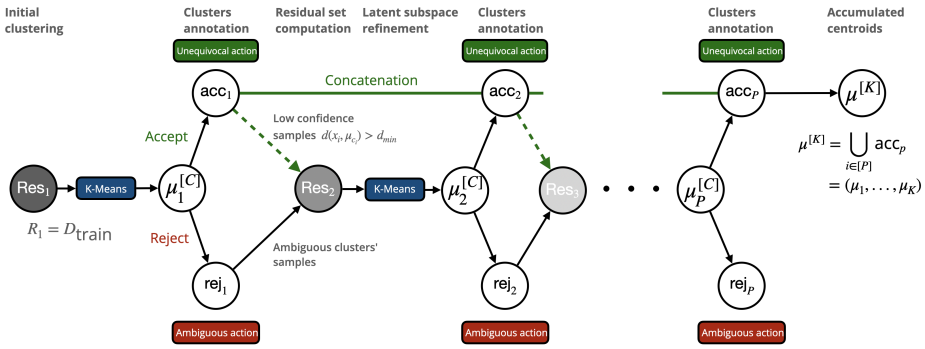

Fig. 1. Illustration of the iterative centroids accumulation step guided by human feedbacks. Under-segmented regions of the VideoMAE-v2 latent space are iteratively refined to extract fine-grained cluster centroids validated by human annotation. acc_p: accepted set of centroids; rej_p: rejected set of centroids; Res_p: Residual sets.

During annotation, clusters whose nearest neighbors exhibit consistent visual patterns corresponding to a single action are accepted, while clusters with heterogeneous semantics are rejected for further refinement. Furthest-neighbor probes typically exhibit weaker semantic alignment, and serve as a quality control by providing a qualitative estimate of assignment errors. As the procedure progresses, furthest neighbors exhibit increased homogeneity and stronger semantic alignment with nearest samples, indicating improved cluster-level semantic consistency, as illustrated in Figure 2-b. Additional examples of accepted and rejected clusters are presented in SM §4.

Lastly, the annotation step enables practitioners to assign application-relevant labels to the clusters. In our clinical setting, each accepted centroid is classified as *task-related* or *exogenous* to support downstream behavioral analyses, as described in Section 6.

Vocabulary consolidation via hierarchical clustering. The accumulation of centroids introduces semantic redundancy, where multiple clusters correspond to the same underlying concept under varying visual conditions. To promote sparsity of the final vocabulary, the accumulated centroids are subsequently consolidated via hierarchical agglomerative clustering (HAC), which defines a reduction mapping Φ from centroids to prototypes. The prototypes vocabulary size G is selected by maximizing the average silhouette score over dendrogram cuts, computed using cosine distance between centroids. In particular, we show in Figure 3-a that HAC outperforms alternatives such as K-means, DBSCAN/HDBSCAN, GMMs, and spectral clustering. In practice, two consecutive consolidation stages are applied to address residual redundancy: the first stage groups similar centroids, while the second stage uses mean-aggregate representations within each prototype to further consolidate near-duplicate clusters (see Figure 3-(b,d)).

Fig. 2. (a) Two illustrative examples of accepted centroids annotation images, and (b) furthest assigned samples after model training. We add for each cluster their average assignment distance (d_avg); the normalized sum-of-squared cosine distance (SSDn); the prevalence across subjects and embeddings (N); the average (s.t.d) number of segments occurrences across participants ($N\_segm/session$); and the average (s.t.d.) segments duration across participants (T).

3.2 Unsupervised temporal action segmentation

To leverage temporal structure, we perform unsupervised temporal action segmentation independently for each participant. Given embeddings $X = (x_1, \ldots, x_n)$, the goal is to partition the sequence into contiguous segments corresponding to semantically consistent actions. We use kernel-based change-point detection (KCP) [1,11], which detects mean-shifts in a Reproducing Kernel Hilbert Space (RKHS), focusing on action boundaries rather than classes.

Kernel-based change-point detection. Let $k(\cdot, \cdot)$ be a positive semidefinite kernel applied to latent embeddings. KCP reformulates change-point detection as a segmentation problem in the associated RKHS, where each segment is approximated by its empirical mean [1,11]. For a candidate segmentation $\tau = \{0 = t_0 < \cdots < t_M = n\}$ with M segments, we minimize the penalized empirical risk

$$\hat{\mathcal{R}}_{n,\lambda}(\tau) = \frac{1}{n} \sum_{m=0}^{M-1} \sum_{i=t_m}^{t_{m+1}-1} \left\| \phi(x_i) - \bar{\mu}_{t_m:t_{m+1}} \right\|_{\mathcal{H}}^2 + \frac{\lambda}{n} M, \tag{1}$$

where $\phi : \mathbb{R}^D \to \mathcal{H}$ is the implicit feature map induced by k, $\bar{\mu}_{t_m:t_{m+1}}$ denotes the empirical RKHS mean of segment m, and the regularization hyperparameter λ controls the number of change points. The optimization is solved exactly using the pruned exact dynamic-programming (PELT) algorithm [14], yielding an efficient nonparametric procedure that does not require specifying the number of segments in advance. The detailed optimization procedure to solve Equation (1)

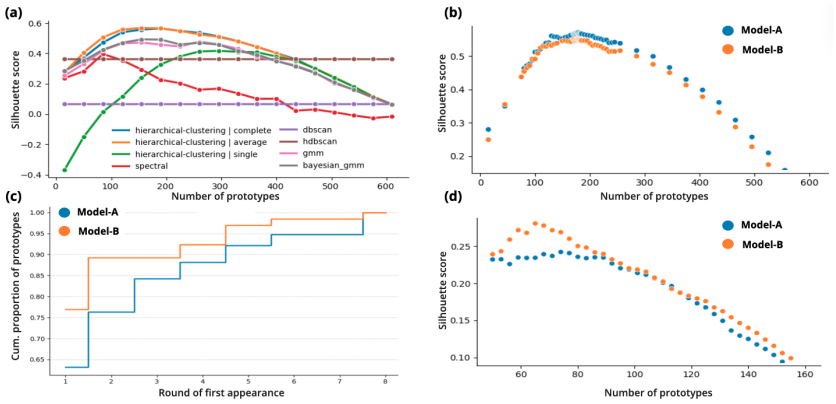

Fig. 3. Evolution of the silhouette score across number of estimated prototypes (a) across clustering methods, (b) during the 1^{st} and 2^{nd} (d) consolidation stages. (c) Cumulative proportion of final prototypes per round of appearance.

using dynamic programming can be found in [1], and is implemented efficiently in the `ruptures` package [32].

Penalty calibration. Following [1,4], we select the penalty parameter using the slope heuristic, which identifies the optimal regularization by detecting a characteristic slope change in the model complexity versus penalty curve. We ensure comparability across participants by estimating a single shared penalty using a fixed-effect formulation rather than tuning λ independently for each sequence.

3.3 Symbolic representation

The final symbolic representation combines the consolidated prototype vocabulary $\mathcal{V}$ with participant-specific temporal segmentations. Given a latent embedding sequence $\mathbf{X} = (x_1, \ldots, x_n)$, each sample is assigned to its nearest centroid and mapped to a prototype via the reduction mapping Φ. Each embedding x_i is then assigned to its closest centroid μ_k under cosine distance, yielding a prototype label $y_i = \Phi(\arg\min_k d(x_i, \mu_k))$. To reduce assignment noise, we filter unreliable assignments using a prototype-specific threshold: for each prototype g, we compute the mean m_g and standard deviation σ_g of their empirical assignment distances, and reject assignments exceeding $m_g + \sigma_g$ by assigning them to a background label ($\tilde{y}_i = -1$).

Segment-level label aggregation. Given a temporal segmentation $\tau = (t_0, \ldots, t_M)$ and filtered labels $\tilde{\mathbf{y}}$, we assign a single label to each segment via majority voting over its central half (excluding boundary frames prone to unstable assignments):

$$s_j = \text{MajVote}\{\tilde{y}_i : i \in [t_{j-1} + \tfrac{\ell_j}{4},\ t_j - \tfrac{\ell_j}{4}]\}, \quad \ell_j = t_j - t_{j-1}.$$

The final symbolic sequence assigns s_j to all indices within segment j. Lastly, short segments (fewer than four embeddings, i.e., $\approx 1.2s$) flanked by identical labels are merged, yielding a piecewise-constant representation.

4 Experimental setting

4.1 Dataset

Data were collected from 162 administrations of a standardized cooking task in a controlled environment, consisting of following the instructions of a recipe to bake a chocolate cake [5]. These sessions correspond to $N = 122$ unique participants: 26 healthy controls and 96 patients with dysexecutive syndromes. Egocentric videos were recorded using an eye-tracker camera at 25 fps. Task durations ranged from 15 to 90 minutes (median: 32 min), yielding 905,903 embeddings in total.

4.2 Model estimation

Sessions are split at the subject level into training (80%, 129 sessions from 98 participants, 725,335 embeddings) and test (20%, 33 sessions from 24 participants, 180,568 embeddings) sets to assess generalization. To evaluate robustness to annotator subjectivity, two independent vocabularies (`Model-A` and `Model-B`) are estimated using feedback from different annotators. Centroids are collected over $P = 8$ rounds with $C = 100$ centroids per round—the maximum that did not degrade clustering quality given the training sample size—yielding $K_{\text{in}} = 800$ centroids prior to consolidation, chosen to exceed significantly the average number of detected action segments ($\bar{M} = 274$, STD$= 24$). The final vocabulary size is determined by HAC consolidation. Cosine k-means is implemented via the `Faiss` library [9], which outperformed alternatives in preliminary experiments.

5 Evaluation protocol and results

In the absence of ground-truth action labels, we adopt a multi-level evaluation protocol designed to assess: (i) the semantic consistency of the learned prototypes, (ii) the robustness of the symbolization procedure to subjective human feedback, and (iii) the fidelity and generalizability of prototypes assignments to empirical samples, in particular across unseen particants.

5.1 Prototyping outcomes and semantic validity

Since consolidation may merge visually distinct actions, we evaluate prototype validity through systematic visual inspection. For each prototype, a visual probe is constructed by aggregating the probes of its constituent centroids (SM §4.4). Prototypes are retained if all constituent centroids encode a single, unambiguous action.

Results. After eight centroid-collection rounds ($K_{\text{in}} = 800$), annotators retained $\approx 80\%$ of centroids (642 and 613 for `Model-A` and `Model-B`). While the majority of final prototypes emerge from the first iteration, subsequent rounds contribute additional semantically distinct actions, demonstrating that the iterative procedure effectively discovers novel prototypes (Figure 3-(c)). Two successive HAC consolidation stages reduced these to $G = 75$ and $G = 65$ final prototypes (detailed progression in SM §5). Of the consolidated prototypes, 55 (73.3%) and 45 (69.2%) were validated as capturing unique actions—rejections mainly resulted from merges of visually similar but semantically distinct actions. Qualitatively, the discovered prototypes span the full range of task-relevant actions (e.g., recipe reading, ingredient manipulation, oven operation) as well as exogenous behaviors (e.g., looking around the room, examiner interactions), without requiring predefined action categories. In the remainder of the evaluation, only validated prototypes are considered.

5.2 Sensitivity to human feedback subjectivity

We compare `Model-A` and `Model-B` at three levels: (i) inter-rater reliability on a shared annotation set measured using Cohen's κ coefficient [6], (ii) vocabulary-level semantic comparison using embedding co-occurrence Hungarian matching, and (iii) agreement between downstream symbolic representations using Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI).

Results. On $n = 400$ shared images, annotators exhibited moderate agreement [17] on ambiguity judgments (Cohen's $\kappa = 0.44$, 95% CI: $[0.35, 0.54]$) but substantial agreement on the task/exogenous distinction ($\kappa = 0.78$, 95% CI: $[0.57, 0.93]$), indicating variability stems from subjective ambiguity rather than semantic disagreement. Importantly, both vocabularies capture the same range of actions: prototype matching reveals one-to-one correspondences with no action missing from either model. Downstream symbolic representations show strong agreement (NMI= 0.791, 95% CI: $[0.790, 0.792]$; ARI= 0.682, 95% CI: $[0.680, 0.685]$) and nearly identical temporal granularity: median durations 6.4s vs. 6.6s, segment counts $1{,}247 \pm 290$ vs. $1{,}298 \pm 310$. Additional details and visualizations are provided in SM §7–8.

5.3 Fidelity and generalizability of symbolic representations

In the absence of ground-truth labels, we evaluate fidelity and generalizability through visual probing: (i) generalizability is assessed by comparing nearest-neighbor probes between training and held-out participants, (ii) fidelity is probed by inspecting furthest-neighbor samples of each prototype to test if distant assignments preserve semantic consistency, and (iii) coverage is assessed qualitatively via both UMAP projection of prototypes' centroids against empirical samples and visual inspections of the furthest-neighbors.

Results. The estimated prototypes transfer well to held-out participants: nearest-neighbor visual probes revealed highly stable visual semantics between training and test subjects. This is corroborated by the overlap between UMAP projections of both training and testing samples, presented in SM §6. We note that this strong generalizability can be theoretically attributable to the *prescribed* nature of the task, performed in a reproducible setting, and the consistent use of the same head-mounted camera and video encoder. Furthest-neighbor samples preserved prototype semantics, validating the distance-based filtering heuristic and ensuring the accuracy of the symbolic representations. However, visually similar actions (e.g., pouring different ingredients) were not consistently disentangled at the centroid level, reflecting both the limited discriminative power of the video encoder and the need for refined residual space definitions in the iterative procedure.

Taken together, these results demonstrate that the proposed symbolization procedure yields fine-grained and temporally accurate proxy representations of participants' behavior, capturing a broad range of actions with strong generalizability to unseen participants. Independent annotators produced vocabularies that, despite minor subjective differences in centroid selection, exhibited strong semantic alignment and temporal consistency—validating their use for downstream behavioral analysis.

6 Application of the symbolic representations to the characterization of behavioral manifestations associated with dysexecutive syndromes

We demonstrate the clinical utility of the symbolic representations by analyzing behavioral differences across diagnosis groups: 26 healthy controls, 59 patients with traumatic brain injury (TBI), and 37 with radiation-induced leukoencephalopathy (RIL). All analyses use the `Model-A` vocabulary, combining validated prototypes with original centroids when consolidation introduced ambiguity, yielding $G = 150$ prototypes manually grouped into 28 action categories (Figure 4-a).

6.1 Temporal organization

Task fragmentation, quantified as symbolic segments per unit time, differed significantly across groups (likelihood-ratio test $\chi^2(2) = 11.53$, $p = 3.31 \times 10^{-3}$). Participants with RIL exhibited 22.9% longer segment durations than controls, while those with TBI showed 8.1% longer durations. These findings align with recent reports linking reduced activity fragmentation to cognitive impairment in aging cohorts [30]. Statistical details are provided in SM §9.

6.2 Action composition across diagnosis groups

We analyzed group differences in action composition by computing, for each of the 28 categories: number of occurrences, mean duration, and proportion of

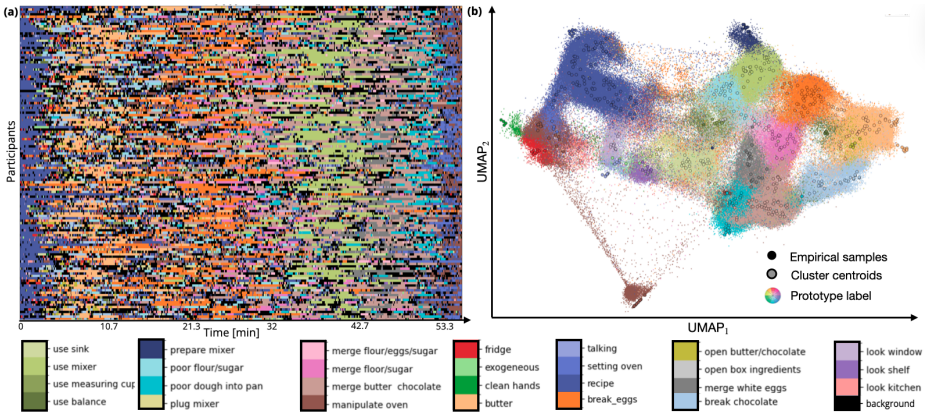

Fig. 4. (a) Symbolic representations of participants' egocentric vision during the procedural task, and (b) two-dimensional UMAP projection of training samples and prototype centroids. Each row in (a) corresponds to a participant, with colors indicating prototype assignments (sequences upsampled for visualization).

task time. Pairwise group comparisons were performed using Mann-Whitney U tests [22] with rank-biserial correlation (r_{rb}) as effect size, and the Benjamini-Hochberg procedure [3] to control the false discovery rate (FDR) across multiple comparisons. Two key patterns emerged (SM §9).

Task-related vs. exogenous actions. Participants with RIL spent a significantly higher proportion of time on exogenous actions (e.g., looking around the room, talking to the examiner) compared to both controls ($p = 8.15 \times 10^{-3}$, median 17% vs. 8.3%) and TBI patients ($p = 2.18 \times 10^{-3}$, median 17% vs. 9.1%). No significant difference was found between TBI and controls. This pattern is consistent with known deficits in sustained attention and goal-directed behavior in severe dysexecutive syndromes [12].

Category-specific findings. Severity of executive impairment was associated with increased recipe consultations (RIL vs. Control: $p_{\text{corr}} = 3.8 \times 10^{-3}$, $r_{rb} = 0.65$; TBI vs. Control: $p_{\text{corr}} = 0.015$) and reduced time spent on core task actions such as merging the dough (RIL vs. Control: $p_{\text{corr}} = 2.3 \times 10^{-3}$, $r_{rb} = -0.68$). These findings corroborate impairments in prospective and episodic memory leading to failures to remember instructions [12,5]. Participants with RIL also exhibited longer oven manipulation durations ($p_{\text{corr}} = 0.014$, $r_{rb} = 0.61$) and more frequent examiner interactions compared to TBI patients ($p_{\text{corr}} = 8.2 \times 10^{-3}$, $r_{rb} = -0.46$). Notably, examiner interactions are part of the standardized error categories counted by neurologists during this ecological assessment [5], suggesting that the symbolic representations recover clinically relevant behavioral markers. The complete per-category analysis is provided in SM §10.

7 Discussion

We presented a framework for constructing interpretable symbolic representations of egocentric video by combining a pretrained video encoder with iterative human-in-the-loop prototype refinement. Unlike pose-based behavioral representations that operate on body coordinates, the prototype explicitly encode rich egocentric visual experience that is crucial to understanding goal-directed behavior. While human feedback guides the rejection of ambiguous clusters, the definition of action categories remains fully unsupervised, and independent annotators produced semantically aligned vocabularies with consistent downstream representations. Importantly, this approach drastically reduces annotation effort compared to exhaustive frame-level or segment-level labeling: annotators provide simple binary accept/reject decisions on candidate cluster probes ($\approx$1-15s per cluster and $\approx$1-60s per prototype), replacing hours of manual video annotation with minutes of visual inspection. The recursive procedure including clustering, visual probes generation, and annotations can be applied entirely (P=8) within a single day on a standard server[1], making the approach computationally efficient and practical for large-scale behavioral studies.

The learned prototypes generalize well to held-out participants, a property we attribute to the prescribed nature of the task, reproducible acquisition conditions, and the use of a shared encoder. This suggests that the approach could extend to other standardized procedural tasks where visual regularities are expected; however, validation on multi-environment benchmarks remains an open direction, as existing large-scale egocentric datasets typically lack the single-environment characteristic that enables cross-participant prototype transfer. A current limitation lies in the consolidation stage: similarity-based hierarchical clustering may merge visually close but semantically distinct actions, requiring manual inspection to preserve vocabulary integrity. Beyond consolidation, fine-grained discrimination of visually similar substeps is bounded by the frozen encoder, and the iterative procedure can propagate under-segmentation errors across rounds despite per-iteration human review; integrating temporal dynamics, object-level cues, or domain-adapted encoders could mitigate both.

Beyond enabling scalable behavioral comparisons without dense action labeling, the symbolic representations provide a principled anchor for multimodal integration: complementary signals such as gaze patterns, pupillometry, or speech can be analyzed within specific action contexts, while temporally aligned modalities could enrich the characterization of action segments themselves.

Acknowledgements We wish to thank all the participants of the study, without whom this research would not have been possible. We gratefully acknowledge the collaboration of the neurologists (Pr. Flavie Bompaire, Pr. Damien Ricard), and the research assistants and nurses from HIA Percy, in particular Celine Mizoule, Elodie Busson, and Mona Michaud. This study is funded by the Direction

[1] We used for the experiments a dual-socket Intel Xeon Gold 5220R system (48 cores, 256 GB RAM)

Centrale du Service de Santé des Armées (DCSSA), and using computational resources from the "Mésocentre" computing center of Université Paris-Saclay, CentraleSupélec and École Normale Supérieure Paris-Saclay supported by CNRS and Région Île-de-France (https://mesocentre.universite-paris-saclay.fr/).

Ethics statement Participants provided written informed consent, and the study protocols were approved by the Comité de Protection des Personnes (CPP) Sud Mediteranee IV (CPP: 210504). Trial is registered at `ClinicalTrials.gov` with identifier: NCT05017051, IDRCB: 2021-A00087-34. For additional access or request, please contact the corresponding author at `sam.perochon@ens-paris-saclay.fr`.

References

1. Arlot, S., Celisse, A., Harchaoui, Z.: A kernel multiple change-point algorithm via model selection. Journal of Machine Learning Research **20**(162), 1–56 (2019), `http://jmlr.org/papers/v20/16-155.html`
2. Bansal, S., Arora, C., Jawahar, C.V.: My view is the best view: Procedure learning from egocentric videos. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII. p. 657–675. Springer-Verlag, Berlin, Heidelberg (2022). `https://doi.org/10.1007/978-3-031-19778-9_38`
3. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: Statistical Methodology **57**(1), 289–300 (Jan 1995). `https://doi.org/10.1111/j.2517-6161.1995.tb02031.x`
4. Celisse, A., Marot, G., Pierre-Jean, M., Rigaill, G.: New efficient algorithms for multiple change-point detection with reproducing kernels. Computational Statistics & Data Analysis **128**, 200–220 (2018). `https://doi.org/https://doi.org/10.1016/j.csda.2018.07.002`, `https://www.sciencedirect.com/science/article/pii/S0167947318301683`
5. Chevignard, M.P., Taillefer, C., Picq, C., Poncet, F., Noulhiane, M., Pradat-Diehl, P.: Ecological assessment of the dysexecutive syndrome using execution of a cooking task. Neuropsychological Rehabilitation **18**(4), 461–485 (Aug 2008). `https://doi.org/10.1080/09602010701643472`
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960). `https://doi.org/10.1177/001316446002000104`
7. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: The epic-kitchens dataset: Collection, challenges and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **43**(11), 4125–4141 (2021). `https://doi.org/10.1109/TPAMI.2020.2991965`
8. Datta, S.R., Anderson, D.J., Branson, K., Perona, P., Leifer, A.: Computational neuroethology: A call to action. Neuron **104**(1), 11–24 (Oct 2019). `https://doi.org/10.1016/j.neuron.2019.09.038`
9. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data **12**(2), 346–361 (Apr 2026). `https://doi.org/10.1109/tbdata.2025.3618474`

10. Fazzari, E., Romano, D., Falchi, F., Stefanini, C.: Animal behavior analysis methods using deep learning: A survey. Expert Systems with Applications **289**, 128330 (2025). https://doi.org/https://doi.org/10.1016/j.eswa.2025.128330, https://www.sciencedirect.com/science/article/pii/S0957417425019499
11. Garreau, D., Arlot, S.: Consistent change-point detection with kernels. Electronic Journal of Statistics **12**(2) (Jan 2018). https://doi.org/10.1214/18-ejs1513
12. Hendry, K., Ownsworth, T., Beadle, E., Chevignard, M.P., Fleming, J., Griffin, J., Shum, D.H.K.: Cognitive deficits underlying error behavior on a naturalistic task after severe traumatic brain injury. Front. Behav. Neurosci. **10**, 190 (Oct 2016)
13. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Lu, D., Wang, Y., Wang, L., Qiao, Y.: 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 (2024)
14. Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association **107**(500), 1590–1598 (Nov 2012). https://doi.org/10.1080/01621459.2012.737745
15. Kobren, A., Monath, N., Krishnamurthy, A., McCallum, A.: A hierarchical algorithm for extreme clustering. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 255–264. KDD '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3097983.3098079
16. Kukleva, A., Kuehne, H., Sener, F., Gall, J.: Unsupervised learning of action classes with continuous temporal embedding. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12058–12066 (2019), https://api.semanticscholar.org/CorpusID:102351314
17. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159 (Mar 1977). https://doi.org/10.2307/2529310
18. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European Conference on Computer Vision (ECCV) (Sep 2018)
19. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 287–295 (2015). https://doi.org/10.1109/CVPR.2015.7298625
20. Luxem, K., Mocellin, P., Fuhrmann, F., Kürsch, J., Miller, S.R., Palop, J.J., Remy, S., Bauer, P.: Identifying behavioral structure from deep variational embeddings of animal motion. Communications Biology **5**(1) (Nov 2022). https://doi.org/10.1038/s42003-022-04080-7
21. Mahmood, S.A., Ali, A.S., Ahmed, U., Fateh, F.J., Zia, M.Z., Tran, Q.H.: Procedure learning via regularized gromov-wasserstein optimal transport. ArXiv **abs/2507.15540** (2025), https://api.semanticscholar.org/CorpusID:280271567
22. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics **18**(1), 50–60 (1947)
23. Mobbs, D., Wise, T., Suthana, N., Guzmán, N., Kriegeskorte, N., Leibo, J.Z.: Promises and challenges of human computational ethology. Neuron **109**(14), 2224–2238 (Jul 2021). https://doi.org/10.1016/j.neuron.2021.05.021

24. Núñez-Marcos, A., Azkune, G., Arganda-Carreras, I.: Egocentric vision-based action recognition: A survey. Neurocomputing **472**, 175–197 (2022). https://doi.org/https://doi.org/10.1016/j.neucom.2021.11.081, https://www.sciencedirect.com/science/article/pii/S0925231221017586
25. Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, J.W., Murthy, M.: Sleap: A deep learning system for multi-animal pose tracking. Nature Methods **19**(4), 486–495 (Apr 2022). https://doi.org/10.1038/s41592-022-01426-1
26. Perochon, S., Oudre, L.: Unsupervised action segmentation of untrimmed egocentric videos. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10097216
27. Shen, Y., Elhamifar, E.: Progress-aware online action segmentation for egocentric procedural task videos. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18186–18197 (2024). https://doi.org/10.1109/CVPR52733.2024.01722
28. Song, Y., Byrne, E., Nagarajan, T., Wang, H., Martin, M., Torresani, L.: Ego4d goal-step: Toward hierarchical understanding of procedural activities. In: Advances in Neural Information Processing Systems. vol. 36, pp. 38863–38886. Curran Associates, Inc. (2023)
29. Spurio, F., Bahrami, E., Francesca, G., Gall, J.: Hierarchical vector quantization for unsupervised action segmentation. In: AAAI Conference on Artificial Intelligence (AAAI) (2025)
30. Tian, Q., Studenski, S.A., An, Y., Kuo, P.L., Schrack, J.A., Wanigatunga, A.A., Simonsick, E.M., Resnick, S.M., Ferrucci, L.: Association of combined slow gait and low activity fragmentation with later onset of cognitive impairment. JAMA Network Open **4**(11), e2135168–e2135168 (11 2021). https://doi.org/10.1001/jamanetworkopen.2021.35168
31. Tinbergen, N.: The Study of Instinct. Clarendon Press (1951)
32. Truong, C., Oudre, L., Vayatis, N.: ruptures: change point detection in python (2018). https://doi.org/10.48550/ARXIV.1801.00826, https://arxiv.org/abs/1801.00826
33. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14549–14560 (Jun 2023)
34. Weinreb, C., Pearl, J.E., Lin, S., Osman, M.A.M., Zhang, L., Annapragada, S., Conlin, E., Hoffmann, R., Makowska, S., Gillis, W.F., Jay, M., Ye, S., Mathis, A., Mathis, M.W., Pereira, T., Linderman, S.W., Datta, S.R.: Keypoint-moseq: parsing behavior by linking point tracking to pose dynamics. Nature Methods **21**(7), 1329–1339 (Jul 2024). https://doi.org/10.1038/s41592-024-02318-2
35. Xu, M., Gould, S.: Temporally consistent unbalanced optimal transport for unsupervised action segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 14618–14627. IEEE (Jun 2024). https://doi.org/10.1109/cvpr52733.2024.01385