Disentangling Intent: Sparse Autoencoders for Interpretable Action Transition Prediction in Egocentric Video

Kosta Gjorgjievski
ghm25@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Eduardus Tjitrahardja
xhp25@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Raphaël El Haddad
tai-s25@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

José Manuel Davila
dhz25@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Abstract

As wearable AI moves from science fiction to everyday devices, building proactive systems that can anticipate a user's next action while remaining transparent and trustworthy has become essential. This project addresses the *interpretability gap* in egocentric intent prediction. We propose training Sparse Autoencoders (SAEs) on latent representations produced by a frozen V-JEPA video foundation model, decomposing the dense 768-dimensional embedding space into a sparse, high-dimensional feature space where each active dimension corresponds to a single, interpretable concept. A linear probe trained on these disentangled features predicts action transitions from the Ego4D dataset. We compare this interpretable pipeline against black-box baselines—a supervised MLP, a linear probe on raw embeddings, and a PCA-reduced linear probe—quantifying the accuracy/interpretability trade-off and producing the first catalog of interpretable features in a video foundation model applied to egocentric intent prediction.

Keywords

egocentric video, sparse autoencoders, mechanistic interpretability, V-JEPA, action transition prediction, intent forecasting

1 Introduction

As wearable computing devices—smart glasses, AR headsets—become ubiquitous, there is growing demand for proactive AI systems capable of anticipating a user's next action before it occurs. The foundation for such systems lies in egocentric video: the first-person visual stream captured by body-mounted cameras, rich in hand–object interactions, gaze patterns, and environmental context. Recent self-supervised models, most notably V-JEPA [2], can produce dense 768-dimensional latent representations that encode powerful spatiotemporal features without requiring human labels. When paired with supervised classifiers such as MLPs trained on Ego4D annotations [8], these representations enable promising action-transition prediction. However, such pipelines remain fundamentally *opaque*: the classifier maps the dense embedding to a label without revealing *which aspects* of the representation drive the prediction.

This project addresses that opacity. We propose applying Sparse Autoencoders (SAEs) to V-JEPA's latent space to reveal the fine-grained features that constitute "intent" in egocentric video, then using those features for transparent, auditable downstream prediction.

1.1 Gaps Identified

Three critical shortcomings motivate this work. First, supervised classifiers on video embeddings (MLPs on V-JEPA or VideoMAE [16] representations) function as black boxes: they confirm that predictive information exists in the latent space but provide no insight into what the model is actually tracking, which is unacceptable for safety-critical wearable applications. Second, neural networks suffer from *superposition* [6], packing multiple distinct concepts into the same dimensions; individual dimensions of dense embeddings are therefore polysemantic and uninterpretable. Third, while SAEs have proven powerful for disentangling language model representations [3, 5, 14] and are beginning to be applied to static vision transformers [7, 11], they have not yet been applied to *video* foundation models. The unique temporal and egocentric character of V-JEPA representations presents an entirely unexplored setting.

1.2 Novelty and Contributions

This project makes three primary contributions:

(1) **First SAE applied to a video foundation model.** To our knowledge, this is the first application of SAEs to V-JEPA for mechanistic interpretability in the video domain.
(2) **Interpretable intent-prediction pipeline.** SAE features replace the opaque MLP with a transparent linear probe, enabling human auditing of which disentangled features drive each action-transition prediction.
(3) **Systematic accuracy–interpretability trade-off study.** We compare against three baselines (MLP, raw-embedding linear probe, PCA linear probe) to quantify the cost of interpretability and isolate the contribution of sparse disentanglement.

2 Background and Related Work

2.1 Egocentric Video and Action Forecasting

Egocentric video, captured from head- or chest-mounted cameras, has emerged as a critical modality for understanding daily human activity. The Ego4D dataset [8]—over 3,700 hours from 931 participants across 74 locations—catalyzed research in action recognition, action forecasting, and intent prediction. EgoVLP [12] introduced

egocentric video–language pre-training, achieving strong performance on anticipation benchmarks, while direct supervised training on Ego4D labels has shown moderate-to-high accuracy for action-transition prediction. All such approaches produce black-box predictions that reveal nothing about the model's internal reasoning.

2.2 Self-Supervised Video Representation Learning

The field progressed from handcrafted features through 3D convolutional architectures [4] and two-stream networks [15] to masked autoencoder methods such as VideoMAE [16]. V-JEPA [2] advances this paradigm by performing prediction entirely in latent space—without pixel reconstruction—yielding dense 768-dimensional embeddings that capture high-level semantic and motion features efficiently.

2.3 Superposition and Sparse Autoencoders

Neural networks exhibiting superposition [6] compress more features than available dimensions, creating polysemantic representations. SAEs address this by decomposing a dense input $\mathbf{x} \in \mathbb{R}^d$ into a sparse feature vector $\mathbf{f} \in \mathbb{R}^{d'}$ ($d' \gg d$) via an encoder–decoder pair trained with reconstruction and sparsity losses. Bricken et al. [3] demonstrated that SAE features trained on transformer activations correspond to recognizable concepts; Cunningham et al. [5] confirmed their high human interpretability. SAEs have since been scaled [14] and applied to CLIP [9, 11], InceptionV1 [7], and medical imaging [1], with the Prisma toolkit [10] emerging as an open-source framework for vision and video. None of these works target video foundation models or use SAE features for downstream prediction in egocentric settings.

2.4 Interpretability for Action Recognition

Traditional interpretability for video action recognition relies on post-hoc methods such as Grad-CAM [13] or attention visualization. These methods identify *where* the model looks but not *what concepts* it tracks. Our approach is complementary: rather than highlighting spatial regions, we decompose the model's internal representation into interpretable features, identifying the specific latent concepts that constitute an intent transition.

3 Challenges

Data and representation. Processing hundreds of hours of Ego4D with V-JEPA requires substantial compute. Annotation quality for action transitions varies across the dataset, and the boundary between an "intent transition" and a natural action continuation can be ambiguous. Determining optimal clip length and temporal stride for embedding extraction requires careful experimentation.

SAE training. Video embeddings encode spatiotemporal information simultaneously, so features may correspond to spatial patterns, temporal dynamics, or combinations of both. The optimal expansion factor, sparsity coefficient λ, and training duration are not established for video representations. There is an inherent tension between reconstruction fidelity (preserving information) and sparsity (forcing disentanglement).

Evaluation. Interpretability is subjective and requires human inspection or automated labeling protocols not designed for video features. Establishing a fair comparison between the MLP and the SAE pipeline demands careful control of training splits and hyperparameters.

4 Objectives

The project aims to accomplish three measurable goals. First, train a Sparse Autoencoder on V-JEPA embeddings from Ego4D and evaluate quality via reconstruction loss, L0 sparsity, and explained variance, targeting an L0 below 100 active features per input. Second, train a linear probe on SAE activations for action-transition prediction and compare against three baselines (MLP, raw-embedding linear probe, PCA linear probe), with success defined as the SAE probe matching or approaching MLP accuracy while providing full feature-level interpretability. Third, identify and catalog the top SAE features driving predictions, mapping them to interpretable spatiotemporal concepts through manual inspection and activation statistics, with success defined as at least 50% of top-predictive features being mappable to human-understandable concepts.

5 Proposed Methodology

5.1 Pipeline Overview

Figure 1 illustrates the end-to-end pipeline, which follows a four-stage modular architecture: (1) encode Ego4D clips with a frozen V-JEPA to obtain 768-dim dense embeddings; (2) train an SAE unsupervised on these embeddings to produce sparse feature activations; (3) train a linear probe on the SAE features using Ego4D action-transition labels; (4) evaluate against baselines and analyze the most predictive features for interpretability.

5.2 Step 1: Embedding Extraction

We use the pre-trained V-JEPA (ViT-L) to extract 768-dim embeddings from Ego4D clips sampled at 16 frames per clip. We experiment with embeddings from early, middle, and late transformer layers to determine which level of abstraction best supports action-transition prediction, as different layers capture different balances of low-level motion and high-level semantics.

5.3 Step 2: Sparse Autoencoder Training

The SAE maps $\mathbf{x} \in \mathbb{R}^{768}$ through a single encoder to $\mathbf{f} \in \mathbb{R}^{d'}$, with $d' = \alpha \times 768$ for expansion factors $\alpha \in \{16, 32, 64\}$, producing feature spaces of 12K, 24K, and 49K dimensions. The training objective is:

$$\mathcal{L}(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \lambda \|\mathbf{f}\|_1$$

where ReLU activations ensure non-negative features. We tune λ to achieve L0 in $[20, 100]$, following practices from language model SAE literature [3, 14]. Training uses Adam with learning-rate scheduling on unlabeled Ego4D embeddings to maximize coverage of the representation space.

5.4 Step 3: Linear Probe and Baseline Comparisons

With the SAE encoder frozen, we train a logistic regression classifier on sparse feature activations using Ego4D action-transition labels. We compare against: (a) **MLP baseline**: multi-layer perceptron on raw 768-dim embeddings; (b) **Linear probe on raw embeddings**: controls for the effect of non-linearity; (c) **PCA + linear probe**: raw

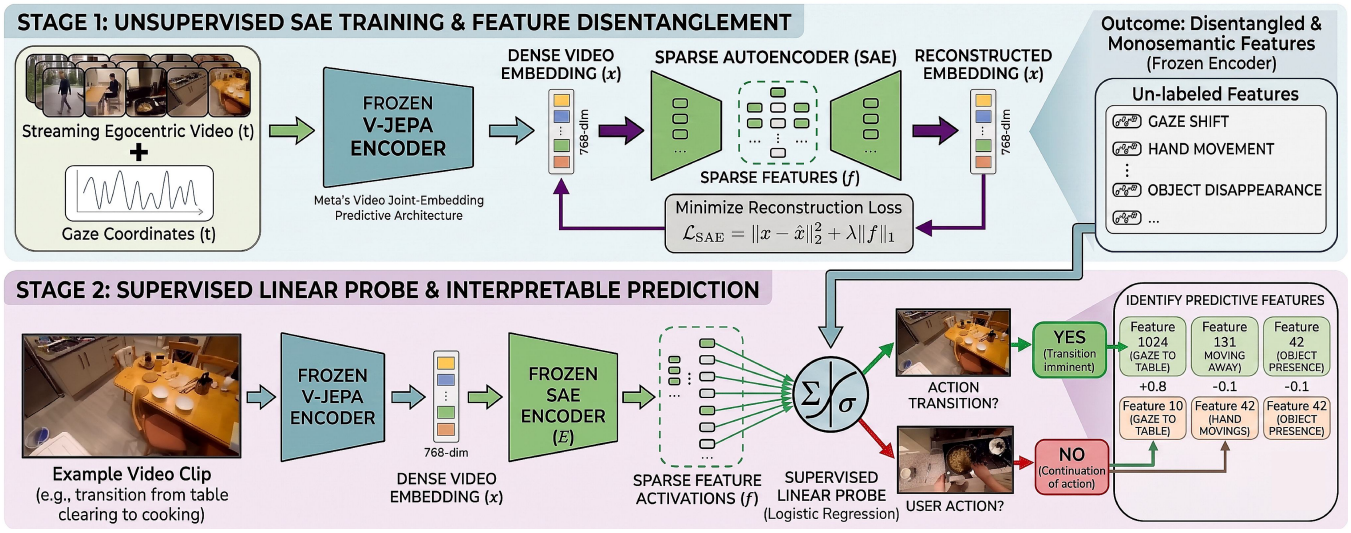

Figure 1: End-to-end pipeline. Ego4D clips are encoded by a frozen V-JEPA to produce dense 768-dim embeddings. A Sparse Autoencoder decomposes these into sparse feature activations (unsupervised). A linear probe trained on those activations predicts action transitions (supervised), and the top-weighted features are analyzed for interpretability.

embeddings reduced to the SAE's typical L0 dimensionality, isolating the contribution of sparse disentanglement over dimensionality reduction alone. All models share the same train/validation/test splits and are evaluated with accuracy, F1-score, and AUC-ROC.

5.5 Step 4: Feature Interpretability Analysis

For each action-transition class, we identify the top-weighted SAE features in the linear probe and analyze clips that maximally activate each feature through: (i) **manual inspection** to label observed spatiotemporal patterns (e.g., "hand retracting from object," "rapid gaze shift"); (ii) **activation statistics** to identify features consistently active for specific transition types; (iii) **feature co-occurrence analysis** to test whether intent is represented as the simultaneous activation of specific feature combinations. This analysis is the key deliverable that distinguishes our approach from the black-box baseline.

6 Dataset

The primary dataset is **Ego4D** [8], released by Meta AI in collaboration with 13 universities. It contains over 3,700 hours of egocentric video from 931 participants across 74 locations in 9 countries, capturing cooking, cleaning, social interactions, and occupational tasks. We use the Action Recognition and Future Forecasting benchmarks, which provide verb–noun action labels and temporal boundaries. The preprocessing pipeline extracts clips around action-transition boundaries (5 s before and after the transition) at a consistent frame rate, encodes them with frozen V-JEPA to produce 768-dim embeddings, and pairs them with binary labels (transition vs. continuation), creating a classification dataset for both unsupervised SAE training (all embeddings) and supervised linear-probe training (labeled subset).

7 Experimental Plan and Evaluation

We compare four approaches: the SAE linear probe, an MLP baseline, a raw-embedding linear probe, and a PCA linear probe. Within the SAE branch we ablate the expansion factor (16×/32×/64×), the L1 coefficient ($\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$), the V-JEPA layer (early/middle/late), and clip length (8/16/32 frames). All models are evaluated on the same held-out test split using accuracy, F1-score, and AUC-ROC. Interpretability is assessed through the feature catalog and the fraction of top-predictive features mappable to human-understandable concepts (target: $\geq 50\%$). Success for the full project is defined as the SAE probe achieving within 5–10% of MLP accuracy while providing complete feature-level interpretability.

8 Project Schedule and Progress

Completed. The team has conducted a comprehensive literature review spanning SAEs for mechanistic interpretability [3, 5, 14], video foundation models [2, 16], egocentric video understanding [8, 12], and emerging SAE applications to vision transformers [7, 10, 11]. The overall pipeline architecture has been designed and responsibilities divided. The evaluation framework is established, including the three baseline comparisons and the interpretability analysis protocol.

Ongoing. Current work focuses on configuring the V-JEPA inference pipeline, establishing Ego4D access, and defining the clip-sampling strategy around action-transition boundaries. Initial SAE hyperparameter experiments on a small embedding subset are being planned to validate the training pipeline before full-scale execution.

Wk	Dates	Task	Details
1	Apr 27 – May 4	Setup & Extraction	V-JEPA inference pipeline; Ego4D access; extract embeddings; layer selection
2	May 5 – May 11	SAE Training & MLP	Train SAE (16×/32×/64×); tune λ; train MLP baseline; PCA baseline
3	May 12 – May 18	Probes & Comparison	Train linear probes on SAE/raw/PCA features; evaluate on held-out set
4	May 19 – May 25	Interpretability	Feature catalog; manual inspection; co-occurrence analysis; validation
5	May 26 – Jun 1	Ablations & Analysis	Ablation studies; robustness across activity types; refine interpretations
6	Jun 2 – Jun 8	Report & Presentation	Final report; visualizations; presentation slides

Table 1: Project timeline. Kosta leads SAE; Edu & Rafa lead MLP baseline; Manuel leads preprocessing and PCA.

9 Expected Outcomes

We expect to deliver: (i) a trained SAE on V-JEPA embeddings with documented reconstruction quality and sparsity metrics; (ii) a comparative evaluation showing the accuracy/interpretability tradeoff between the SAE linear probe, MLP, raw-embedding linear probe, and PCA baseline; (iii) a catalog of interpretable SAE features mapped to visual and temporal concepts in egocentric video; and (iv) a test of whether intent is better characterized as the coactivation of specific sparse features rather than a single monolithic prediction—the central hypothesis of the mechanistic interpretability program applied to video.

References

[1] Ahmed Abdulaal, Hannah Fry, Nina Montaña-Brown, Ayodeji Ijishakin, Junaid Gao, Stephanie Hyland, Daniel C. Alexander, and Daniel C. Castro. 2024. An X-Ray Is Worth 15 Features: Sparse Autoencoders for Interpretable Radiology Report Generation. *arXiv preprint arXiv:2410.03334* (2024).

[2] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, Nicolas Ballas, Amy Shuster, and Emmanuel Dupoux. 2024. Revisiting Feature Prediction for Learning Visual Representations from Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA.

[3] Trenton Bricken, Adly Templeton, Joshua Marcus, et al. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Anthropic Research* (2023). https://transformer-circuits.pub/2023/monosemantic-features.

[4] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, USA, 6299–6308.

[5] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse Autoencoders Find Highly Interpretable Features in Language Models. *arXiv preprint arXiv:2309.08600* (2023).

[6] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, et al. 2022. Toy Models of Superposition. *arXiv preprint arXiv:2209.10652* (2022).

[7] Lucy Gorton. 2024. The Missing Curve Detectors of InceptionV1: Applying Sparse Autoencoders to InceptionV1 Early Vision. *arXiv preprint arXiv:2406.03662* (2024).

[8] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, et al. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 18995–19012. doi:10.1109/CVPR52688.2022.01842

[9] Sonia Joseph, Pranav Suresh, Evan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Russ Graham, Danilo Bzdok, Wojciech Samek, and Blake A. Richards. 2025. Steering CLIP's Vision Transformer with Sparse Autoencoders. *arXiv preprint arXiv:2504.08729* (2025). CVPR 2025 Workshop on Mechanistic Interpretability for Vision.

[10] Sonia Joseph, Pranav Suresh, Lorenz Hufe, Emma Stevinson, Russ Graham, Yilun Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake A. Richards. 2025. Prisma: An Open Source Toolkit for Mechanistic Interpretability in Vision and Video. *arXiv preprint arXiv:2504.19475* (2025). CVPR 2025 Workshop on Mechanistic Interpretability for Vision.

[11] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. 2025. PatchSAE: Sparse Autoencoders Reveal Selective Remapping of Visual Concepts During Adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, Singapore. arXiv:2412.05276.

[12] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, et al. 2022. Egocentric Video-Language Pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, New Orleans, LA, USA. arXiv:2206.01670.

[13] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 618–626.

[14] Dong Shu, Xuzhong Wu, Hao Zhao, Devvrit Rai, Zijian Yao, Ninghao Liu, and Mengnan Du. 2025. A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. ACL, Miami, FL, USA. arXiv:2503.05613.

[15] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Montreal, QC, Canada, 568–576.

[16] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, New Orleans, LA, USA. arXiv:2203.12602.