BEHAVE: Behavioral Ethology for Human Assessment via Variational Encoding

Zhanqi Zhang
University of California San Diego
San Diego, CA 92122
zhz091@ucsd.edu

Chi K. Chou
University of California San Diego
San Diego, CA 92122
ckchou@ucsd.edu

Holden Rosberg
University of California San Diego
San Diego, CA 92122
hrosberg@health.ucsd.edu

William Perry
University of California San Diego
San Diego, CA 92122
wperry@health.ucsd.edu

Jared W Young
University of California San Diego
San Diego, CA 92122
j9young@health.ucsd.edu

Arpi Minassian
University of California San Diego
San Diego, CA 92122
aminassian@health.ucsd.edu

Gal Mishne
University of California San Diego
San Diego, CA 92122
gmishne@ucsd.edu

Mikio Aoi
University of California San Diego
San Diego, CA 92122
maoi@ucsd.edu

Abstract

Quantifying spontaneous human behavior remains a major challenge in psychiatry and neuroscience. We present BEHAVE (Behavioral Ethology for Human Assessment via Variational Encoding), a framework that combines computer vision and unsupervised latent-variable models to capture fine-scale, naturalistic behaviors. BEHAVE segments continuous motion into interpretable motifs and introduces novel metrics of temporal structure, repertoire diversity, and stereotypy. In a naturalistic open-field assay of individuals with euthymic bipolar disorder (BD) and healthy controls, these metrics revealed subtle yet robust BD-associated differences, including reduced exploratory transitions and repertoire narrowing. Compared to clinical scales and standard action-recognition models, BEHAVE achieved superior classification of BD. This approach offers a scalable, bias-resistant path to decoding neuropsychiatric states from natural behavior and lays the foundation for translational biomarkers.

1 Introduction

Machine learning (ML) has advanced the study of neural activity and transformed many domains in neuroscience and medicine, yet its impact on psychiatry remains limited. Free-moving human behavior, an essential manifestation of brain and mind, is especially difficult to quantify due to its multidimensional nature and variability across individuals, contexts, and environments. Unlike neural data, behavioral assessment lacks standardized metrics and consistency, and is further complicated

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

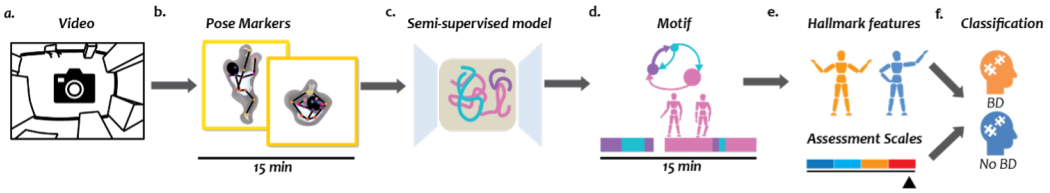

Figure 1: BEHAVE analysis pipeline. (a) Video recordings of free-moving human participants. (b) markers were placed on key points of participants (e.g., elbows) via [8]. (c) Pose markers were fed into a latent-variable model, and (d) the latent representations were used to segment the videos into motifs. (e) Hallmark behavioral features that characterized BD in different time scales were computed via metrics, and (f) were used to classify whether a participant is from the BD or no-BD groups.

by inconsistent camera angles, occlusions, and uncontrolled settings. As a result, clinical practice continues to rely on rating scales and self-reports to quantify human behavior, which reduces these rich dynamics into coarse categories and is vulnerable to bias and subjectivity [1–4].

This under-utilization of ML is particularly striking in bipolar disorder (BD), characterized by recurring episodes that alternate between mania and depression[5], yet the quantification still depends on tools such as the Hamilton Depression Rating Scale (HAM-D) and Young Mania Rating Scale (YMRS) [6, 3]. These measures aggregate behavior across contexts and may not capture micro-behaviors—such as subtle changes in fidgeting, movement variability, or the temporal structure of exploration—that may differentiate BD patients even during euthymia, the state between mania and depression. Such overlooked features could carry vital information about underlying pathophysiology but remain invisible to traditional methods.

To address these challenges, we introduce BEHAVE (Behavioral Ethology for Human Assessment via Variational Encoding), a computational ethology framework designed to move beyond subjective ratings and coarse categories, toward scalable, objective, and fine-grained behavioral assessment in naturalistic human settings.

2 Related Works

Several approaches have sought to bridge the gap between behavior and psychiatric assessment. In humans, wearable sensors have been used to capture gross activity patterns, and the human Behavioral Pattern Monitor (hBPM) [7] has adapted the rodent open-field assay to quantify exploratory behavior in BD. While promising, these methods still rely on manual annotation or predefined metrics, making them labor-intensive, susceptible to bias, and limited in their ability to capture subtle, concurrent micro-behaviors.

In animal behavior research, the field has advanced rapidly with pose estimation methods [8–10], paired with unsupervised segmentation frameworks including [11–16]. These methods have revealed stereotyped behavioral motifs linked to neural and pharmacological manipulations, underscoring the value of data-driven decomposition of continuous behavior. Inspired by this success, there is growing interest in applying similar approaches to humans. However, existing human pose-tracking systems (e.g., [17–19]) and action-recognition models (e.g., [20–22]) remain constrained by reliance on predefined action labels, controlled camera setups, and narrow behavioral taxonomies. These constraints limit their suitability for capturing the nuanced and variable nature of human behavior critical to psychiatry.

3 Method

We developed BEHAVE, an end-to-end framework for quantifying naturalistic human behavior in the hBPM open-field assay [7]. The pipeline combines four components: 1) data processing and pose extraction, 2) latent encoding, 3) motif segmentation and metrics, and 4) BD classification benchmarking (Fig.1).

Specifically, we used DeepLabCut [8] trained on 20–50 labeled top-view frames of each video, across 20 body landmarks. After three iterations with outlier correction, errors were 2.0 px (train) / 3.7 px (test). Landmarks were aligned to egocentric skeletons; estimates < 0.9 confidence were dropped.

Pose sequences were then embedded via a latent variable model (VAME [13]) into a 10-D latent space. Behavioral motifs, representing recurrent, identifiable actions of human behavior, were obtained by k-means clustering ($k = 10$ for comparability with human annotations). We divided the recording into three 5-min epochs.

To obtain the hallmark features of the motifs, we proposed several measures to quantify the behavior of the subjects, including two new measures to quantify the variability of the behavior.

Motif dwell time: the mean dwell time (number of frames) occupied by each motif in each of the different epochs in the BD and healthy comparison (HC) population.

Motif transitions: motifs were modeled as states of a Markov chain with transition matrix P (self-transitions excluded). We computed (i) transition frequency (from weighted adjacency), (ii) sparsity (count of non-transitioning rows with nonzero dwell time). We further proposed *Effective Number of Accessible States (ENAS)* per motif state E_{S_i}, and reported the average ENAS over motifs:

$$E_{S_i} = \begin{cases} 0, & \text{if } \sum_{j \neq i} P_{ij} = 0 \\ \left(\sum_{j \neq i} P_{ij}^2 \right)^{-1}, & \text{otherwise} \end{cases}, \qquad \text{ENAS} = \frac{1}{n} \sum_{i=1}^{n} E_{S_i}, \tag{1}$$

ENAS is intended to measure the number of accessible motifs (states) in each period by weighting motif counts by their transition probabilities, capturing how freely participants move between behaviors, from flexible exploration $n - 1$ to highly stereotyped patterns (0 or 1).

Motif-volume and population distance: Motif-volume captures variability in how a motif is expressed. For each subject, we calculated the trace of the covariance of latent vectors within a motif (per epoch and population), providing an estimate of the spread of that motif's expression. We also assessed interpopulation distance between BD and HC to test whether motifs were performed differently across groups. We fit Gaussian distributions to the latent samples for each motif and epoch, then computed the 2-Wasserstein distance between them. As a control, we also computed intrapopulation distances (within BD and within HC). Larger interpopulation distances indicate that BD and HC participants expressed the same motif with more distinct latent representations.

4 Experiments

Dataset Adults aged 18–55 were recruited into two groups: bipolar disorder (BD; n = 24, 10 men; one was cyclothymic, the rest BD Type I/II) and healthy comparisons (HC; n = 24, 12 men). Diagnoses were based on the Structured Clinical Interview for DSM-IV [23]. All BD participants were euthymic (YMRS < 12, HAM-D < 10), and HC were free of psychiatric or neurological conditions. Sessions took place in a 2.7 × 4.3 m furnished room with 11 objects, where participants remained alone for 15 min while a ceiling fisheye camera (640 × 480, 30 fps) recorded behavior. Clinicians annotated 10 exploratory behaviors, and the Spatial-D metric [7] was computed. Further methodological details are provided in [24].

Classification Results To assess whether BEHAVE-derived features could distinguish BD from HC, we compared against labels from human annotation and from two pretrained computer vision (CV) action-recognition models [19, 25] as baselines. Features from each video were used for classifying BD. Backward sequential feature selection (4-fold cross-validation, target = 30 features) preceded logistic regression. We used stratified splits (75/25 train/test) and leave-5-subjects-out cross-validation for subject-independent evaluation, reporting mean ± SD accuracy across iterations. Pairwise method comparisons used Tukey's HSD with multiple-comparison control. Table 1 shows that the motif features achieved significantly higher accuracy than models trained on rating scales (HAM-D, YMRS), manual annotations, or CV-based features. Among motif-derived metrics, our proposed ENAS and motif-volume were the strongest predictors, underscoring the importance of transition structure and expression variability.

Analysis Results In Figure 2, we present examples of specific features and embeddings which demonstrate that BEHAVE captures meaningful behavior characteristics in the latent space. We summarize some of the more profound findings below.

Differences in motif usage between BD and HC. We found BD participants spent more time in stretching (motif 1) and fidgeting (motif 4), and less time in object-directed exploration (motif 9).

Table 1: Classification accuracy of BD vs HC across approaches.

Model	Pretrained	Accuracy (%)
Assessment Scales [6, 3]	-	37.41
Spatial-D [7]	-	33.33
K-means on Deeplabcut	-	71.48
hBPM video ratings [7]	-	71.11
S3D [25]	Kinetics-400	63.70
MMAction [19]	Kinetics-400	72.96
Ours	-	**77.04***

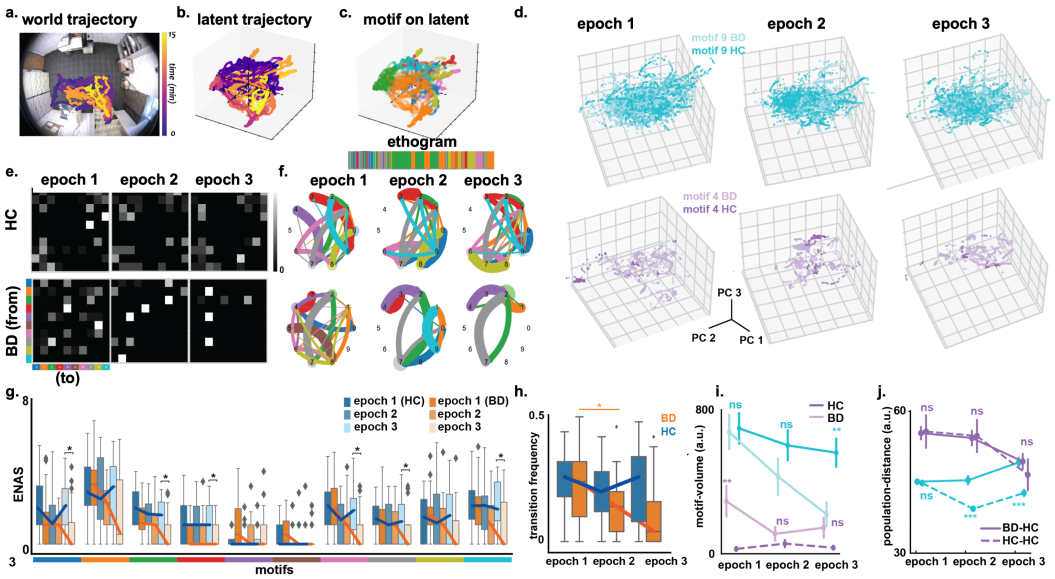

Figure 2: (a-c) World trajectory, latent trajectory, and motif segmentation on the latent trajectory and ethogram corresponding to the motifs. (d) The latent representation of motif 9 “inspect and get distracted” and 4 "fidget" for BD and HC over 3 epochs. (e) Transition matrices of an HC video and a BD video, and (f) probability transition graphs corresponding to the matrices. (g) ENAS, and (h) transition frequency in BD and HC in 3 epochs. (i-j) Motif volume and population distance in 3 epochs for the motifs in (d).

While differences were statistically significant, the were modest, suggesting that duration-based metrics alone provide limited discriminative power.

Motif transitions reveal increased stereotypy in BD. Both groups showed a decrease in transitions over time, but this decline was more pronounced in BD. Transition matrices for BD participants became progressively sparser, with an increasing number of “non-transitioning” motifs where behavior remained confined to a single state. Larger ENAS indicates greater repertoire flexibility. BD participants exhibited a marked decline in ENAS across epochs, reflecting progressive behavioral narrowing and stereotypy. This result was robust across different model granularities (n = 10 vs. 30 motifs), indicating that reduced flexibility is a reliable feature of BD behavior (Figure 2 e-h).

Motif variability distinguishes BD expression. Using motif-volume, we found that BD participants showed reduced variability in exploratory behaviors (motifs 2, 6, 9) and increased variability in fidgeting and static postures (motifs 4, 5). These differences became most pronounced in the final epoch, suggesting progressive divergence between groups as the session unfolded. Interpopulation analyses confirmed that BD and HC expressed motifs in increasingly distinct ways over time, with inter-group distances exceeding intra-group variability for multiple motifs (Figure 2 d, i-j). This indicates that BD participants not only used motifs differently but also performed them with qualitatively distinct kinematic profiles.

Group-level differences were robust to potential confounds. Group comparisons identified 12 of 35 features that remained significant after controlling for age. Medication subgroup analyses suggested that group differences were robust, with no clear medication effects. However, the modest sample size limits generalizability and warrants follow-up in larger studies.

5 Conclusions

In summary, BEHAVE uncovered rich, fine-grained behavioral motifs that human raters or pre-trained CV models did not capture. The new measures we introduce, such as ENAS and motif-volume, provide sensitive, interpretable metrics of behavioral organization that reveal subtle BD-related signatures previously inaccessible. Using these metrics, we found participants showed decreased transition diversity and reduced within-motif variability, leading to progressive behavioral stereotypy. These features outperformed clinical scales and standard vision baselines in classifying BD versus controls, underscoring the promise of unsupervised behavioral motif analysis for computational psychiatry. By quantifying these dynamics directly from large-scale naturalistic behavior, BEHAVE shows how machine learning can extract unbiased, interpretable features that advance both computational methods and their application to core questions in psychiatry.

References

[1] R. Michael Bagby, Andrew G. Ryder, Deborah R. Schuller, and Margarita B. Marshall. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *The American Journal of Psychiatry*, 161(12):2163–2177, 12 2004. PMID: 15569884.

[2] Lynn P. Rehm and Michael W. O'Hara. Item characteristics of the hamilton rating scale for depression. *Journal of Psychiatric Research*, 19(1):31–41, jan 1 1985.

[3] A rating scale for mania: reliability, validity and sensitivity - PubMed. https://pubmed.ncbi.nlm.nih.gov/728692/. [Online; accessed 2023-05-09].

[4] Constantina Demetriou, Bilge Uzun Ozer, and Cecilia A. Essau. *Self-Report Questionnaires*, pages 1–6. John Wiley & Sons, Ltd, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118625392.wbecp507 DOI: 10.1002/9781118625392.wbecp507.

[5] Frederick K. Goodwin and Kay Redfield Jamison. *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression*. Oxford University Press, Oxford, New York, second edition, second edition edition, mar 22 2007.

[6] M. Hamilton. Development of a rating scale for primary depressive illness. *The British Journal of Social and Clinical Psychology*, 6(4):278–296, 12 1967. PMID: 6080235.

[7] Jared W. Young, Arpi Minassian, Martin P. Paulus, Mark A. Geyer, and William Perry. A Reverse-Translational Approach to Bipolar Disorder: Rodent and human studies in the Behavioral Pattern Monitor. *Neuroscience and biobehavioral reviews*, 31(6):882–896, 2007. PMID: 17706782 PMCID: PMC2025688.

[8] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 9 2018. number: 9 publisher: Nature Publishing Group.

[9] Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning | eLife. https://elifesciences.org/articles/47994. [Online; accessed 2023-04-25].

[10] Talmo D. Pereira, Nathaniel Tabris, Arie Matsliah, David M. Turner, Junyu Li, Shruthi Ravindranath, Eleni S. Papadoyannis, Edna Normand, David S. Deutsch, Z. Yan Wang, Grace C. McKenzie-Smith, Catalin C. Mitelut, Marielisa Diez Castro, John D'Uva, Mikhail Kislin, Dan H. Sanes, Sarah D. Kocher, Samuel S.-H. Wang, Annegret L. Falkner, Joshua W. Shaevitz, and Mala Murthy. Sleap: A deep learning system for multi-animal pose tracking. *Nature Methods*, 19(4):486–495, 4 2022. number: 4 publisher: Nature Publishing Group.

[11] Caleb Weinreb, Mohammed Abdal Monium Osman, Libby Zhang, Sherry Lin, Jonah Pearl, Sidharth Annapragada, Eli Conlin, Winthrop F. Gillis, Maya Jay, Shaokai Ye, Alexander Mathis, Mackenzie Weygandt Mathis, Talmo Pereira, Scott W. Linderman, and Sandeep Robert Datta. Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. mar 17 2023. page: 2023.03.16.532307 section: New Results.

[12] Alexander B. Wiltschko, Matthew J. Johnson, Giuliano Iurilli, Ralph E. Peterson, Jesse M. Katon, Stan L. Pashkovski, Victoria E. Abraira, Ryan P. Adams, and Sandeep Robert Datta. Mapping Sub-Second Structure in Mouse Behavior. *Neuron*, 88(6):1121–1135, dec 16 2015. PMID: 26687221 PMCID: PMC4708087.

[13] Kevin Luxem, Petra Mocellin, Falko Fuhrmann, Johannes Kürsch, Stefan Remy, and Pavol Bauer. Identifying Behavioral Structure from Deep Variational Embeddings of Animal Motion. jan 14 2022. page: 2020.05.14.095430 section: New Results.

[14] Alexander I. Hsu and Eric A. Yttri. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature Communications*, 12(1):5188, aug 31 2021. number: 1 publisher: Nature Publishing Group.

[15] Changhao Shi, Sivan Schwartz, Shahar Levy, Shay Achvat, Maisan Abboud, Amir Ghanayim, Jackie Schiller, and Gal Mishne. Learning Disentangled Behavior Embeddings. nov 9 2021. [Online; accessed 2025-03-03].

[16] Nastacia L. Goodwin, Jia J. Choong, Sophia Hwang, Kayla Pitts, Liana Bloom, Aasiya Islam, Yizhe Y. Zhang, Eric R. Szelenyi, Xiaoyu Tong, Emily L. Newman, Klaus Miczek, Hayden R. Wright, Ryan J. McLaughlin, Zane C. Norville, Neir Eshel, Mitra Heshmati, Simon R. O. Nilsson, and Sam A. Golden. Simple Behavioral Analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience. *Nature Neuroscience*, 27(7):1411–1424, 7 2024. publisher: Nature Publishing Group.

[17] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime Multi-Person 2d Pose Estimation using Part Affinity Fields. may 30 2019. arXiv:1812.08008 [cs].

[18] Movenet: Ultra fast and accurate pose detection model. | TensorFlow Hub. https://www.tensorflow.org/hub/tutorials/movenet. [Online; accessed 2023-04-25].

[19] MMAction2 Contributors. Openmmlab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2, 7 2020. original-date: 2020-07-11T07:19:10Z.

[20] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. aug 2 2016. arXiv:1608.00859 [cs].

[21] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. feb 12 2018. arXiv:1705.07750 [cs].

[22] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 318–335, Cham, 2018. Springer International Publishing.

[23] Carl C. Bell. Dsm-IV: Diagnostic and Statistical Manual of Mental Disorders. *JAMA*, 272(10):828–829, sep 14 1994.

[24] Brook L. Henry, Arpi Minassian, Virginie Patt, Jessica Hua, Jared W. Young, Mark A. Geyer, and William Perry. Inhibitory deficits in euthymic bipolar disorder patients assessed in the Human Behavioral Pattern Monitor. *Journal of affective disorders*, 150(3):948–954, sep 25 2013. PMID: 23759280 PMCID: PMC3759601.

[25] Xin Xiong, Weidong Min, Wei-Shi Zheng, Pin Liao, Hao Yang, and Shuai Wang. S3d-CNN: skeleton-based 3d consecutive-low-pooling neural network for fall detection. *Applied Intelligence*, 50(10):3521–3534, oct 1 2020.