UCLA
UCLA Electronic Theses and Dissertations

Title

Representation Learning from Human Neural Activity: From Clinical Events to Memories Across Wake and Sleep

Permalink

https://escholarship.org/uc/item/16t0w1nv

ISBN

9798189269684

Author

Ding, Yuanyi

Publication Date

2026-08-28

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Representation Learning from Human Neural Activity:

From Clinical Events to Memories Across Wake and Sleep

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Electrical and Computer Engineering

by

Yuanyi Ding

2026

© Copyright by

Yuanyi Ding

2026

ABSTRACT OF THE DISSERTATION

Representation Learning from Human Neural Activity:

From Clinical Events to Memories Across Wake and Sleep

by

Yuanyi Ding

Doctor of Philosophy in Electrical and Computer Engineering

University of California, Los Angeles, 2026

Professor Vwani Roychowdhury, Chair

Learning representations from human neural activity is constrained by limited supervision, participant-specific sampling, heterogeneous recordings, and distribution shifts across institutions, tasks, and brain states. This dissertation develops representation-learning strategies under these constraints as targets progress from clinical events to episodic content during recall and sleep. Meaningfulness is operational: a representation must relate to its target, remain informative under variation, and be evaluated with evidence distinct from training. The first setting begins with high-frequency oscillation candidates proposed by legacy detectors but lacks exhaustive and consistently defined pathological labels. A self-supervised variational model learns event morphology, clustering discovers weak labels, and a classifier refines the decision boundary using real and generated latent

examples. This approach improves patient-level evaluation across multi-institutional datasets while preserving candidate generation and expert review as separate parts of the workflow. Omni-iEEG provides a harmonized benchmark spanning 302 patients, 178 hours, and more than 36,000 expert-validated annotations, enabling representation choices to be compared under common targets, splits, metadata, and clinical evidence. The second setting asks whether supervision available during one naturalistic experience can support inference during later internally generated cognition. Participant-specific multi-region attention models trained only on densely annotated movie viewing produce concept activations that preferentially precede matching verbal reports without recall labels for training. Models trained using only medial temporal lobe (MTL) spikes significantly decode concepts before but not after sleep, while models trained using only frontal cortex (FC) spikes decode concepts after but not before sleep. These combined issues create a single-exposure transfer-and-evaluation problem, where representations learned from sparse neuronal recordings during initial experience poorly generalize for evaluation in states with substantially different distributions. The final study addresses this problem using a dual-head objective that combines multi-label classification with contrastive concept alignment, followed by Recall-Adapted Sleep Alignment, which couples unlabeled sleep-domain adaptation with recall-based semantic regularization. Across 62 held-out concept-selective neurons, the adapted representation shows stronger concept-specific temporal coupling than unadapted and comparison models. Together, these studies show that progress in neural representation learning depends not only on model architecture, but also on how candidate spaces, supervision, distribution shifts, and independent evaluation evidence are jointly designed.

The dissertation of Yuanyi Ding is approved.

Hiroki Nariai

Jonathan Kao

Itzhak Fried

Vwani Roychowdhury, Committee Chair

University of California, Los Angeles

2026

DEDICATION

To everyone who guided, encouraged, and supported me along this journey. To anyone reading these words while pursuing a dream, may you be as fortunate as I have been to find kindness, wisdom, and companionship along the way.

愿君千万岁，无岁不逢春

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC	anterior cingulate cortex
ACF	activation-conditioned firing
AMY	amygdala
AUC	area under the receiver operating characteristic curve
BCE	binary cross-entropy
BECA	Bidirectional Event-Conditioned Alignment
CCA	cross-channel attention
CEBRA	Consistent Embeddings of high-dimensional Recordings using Auxiliary variables
CLAP	Contrastive Language–Audio Pretraining
CNN	convolutional neural network
CORAL	correlation alignment
CRA	cross-region attention
CSA	channel-specific attention
CT	computed tomography
EEG	electroencephalography
eHFO	epileptogenic high-frequency oscillation
ENT	entorhinal cortex
FC	frontal cortex
FCA	firing-conditioned activation
FCS	firing-conditioned similarity
HFO	high-frequency oscillation
HPC	hippocampus
HUP	Hospital of the University of Pennsylvania
iEEG	intracranial electroencephalography
KL	Kullback–Leibler
MCS	Memory Confidence Score
MNI	Montreal Neurological Institute
MRI	magnetic resonance imaging
MTL	medial temporal lobe
NREM	non-rapid-eye-movement
PCA	principal component analysis
PFC	prefrontal cortex
PHC	parahippocampal cortex
PSTH	peri-stimulus time histogram
RASA	Recall-Adapted Sleep Alignment
RSA	region-wise self-attention
SCF	similarity-conditioned firing
SD	standard deviation
SEM	standard error of the mean
SOZ	seizure-onset zone
spkHFO	spike-associated high-frequency oscillation
SS2LD	Self-Supervised to Label Discovery
STE	short-term energy
VAE	variational autoencoder

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Professor Vwani Roychowdhury, and my co-advisor, Professor Itzhak Fried, for shaping the direction of this research and for their sustained mentorship, scientific training, and the resources that made this work possible throughout my doctoral studies.

I thank my committee members, Professor Hiroki Nariai and Professor Jonathan Kao, for their thoughtful advice, constructive feedback, and careful review of this dissertation.

I am especially grateful to Dr. John Sakon and Dr. Soraya Dunn for their collaboration and for their contributions to the experimental and scientific work presented in this dissertation.

I am grateful to my lab members and peers for the daily scientific discussions, code feedback, experimental assistance, and companionship that supported this work and made the doctoral journey a shared experience.

I extend my deepest gratitude to the patients who participated in these studies and to their families. Their generosity and willingness to contribute during demanding periods of clinical care made this research on human intracranial neural activity possible.

I am profoundly grateful to my parents for their unconditional love, patience, and sacrifices. Their steadfast belief in me gave me the courage to pursue this path and the strength to continue through its most difficult moments. Whatever I have accomplished rests on the foundation of their support.

The Movie–Recall–Sleep study incorporated into Chapters 3 and 4 was supported by the National Institute of Neurological Disorders and Stroke under grants R01NS084017 and U01NS123128.

For the SS2LD work incorporated into Chapter 2, Hiroki Nariai was supported by the National Institute of Neurological Disorders and Stroke under grant K23NS128318, the Sudha Neelakantan Venky Harinarayan Charitable Fund, the Elsie and Isaac Fogelman Endowment, and the UCLA Children's Discovery and Innovation Institute Junior Faculty Career Development Grant (CDI-TTCF-07012021). Atsuro Daida was supported for research abroad by the Uehara Memorial Foundation and the SENSHIN Medical Research Foundation.

The PyHFO 2.0 publication incorporated into Chapter 2 reports that the research received no external funding.

Chapter 2 incorporates material from Y. Ding, Y. Zhang, C. Duan, A. Daida, Y. Zhang, S. Kanai, M. Lu, S. Hussain, R. J. Staba, H. Nariai, and V. Roychowdhury, "PyHFO 2.0: An open-source platform for deep learning-based clinical high-frequency oscillations analysis," *Journal of Neural Engineering*, vol. 22, no. 5, art. no. 056040, 2025, doi: 10.1088/1741-2552/ae10e0. Yuanyi Ding led the software, validation, original drafting, and manuscript revision and contributed equally to methodology, project administration, and visualization. Yipeng Zhang contributed equally to methodology, project administration, software, validation, and writing. Chenda Duan contributed software, validation, and writing; Atsuro Daida and Sotaro Kanai contributed data curation, validation, and manuscript review; Yun Zhang and Mingjian Lu contributed software and validation; Shaun Hussain contributed data curation and validation; Richard

J. Staba contributed methodology; Hiroki Nariai contributed data curation, methodology, project administration, supervision, and manuscript review; and Vwani Roychowdhury led project administration and supervision and contributed methodology and manuscript review.

Chapter 2 also incorporates material from Y. Zhang, Y. Ding, C. Duan, A. Daida, H. Nariai, and V. Roychowdhury, “Self-supervised distillation of legacy rule-based methods for enhanced EEG-based decision-making,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2025*, pp. 477–486, 2025, doi: 10.1007/978-3-032-04984-1_46. Yipeng Zhang led the development and evaluation of SS2LD and preparation of the manuscript. Yuanyi Ding contributed to method development, experimental analysis, interpretation, and manuscript preparation. Chenda Duan contributed to implementation and evaluation; Atsuro Daida contributed clinical data curation and validation; Hiroki Nariai contributed clinical interpretation and supervision; and Vwani Roychowdhury directed and supervised the project.

Chapter 3 incorporates material from C. Duan, Y. Zhang, S. Kanai, Y. Ding, A. Daida, P. Yu, T. Zheng, N. Kuroda, S. A. Hussain, E. Asano, H. Nariai, and V. Roychowdhury, “Omni-iEEG: A large-scale, comprehensive iEEG dataset and benchmark for epilepsy research,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. Chenda Duan and Yipeng Zhang contributed equally and led dataset harmonization, benchmark development, model evaluation, and manuscript preparation. Sotaro Kanai contributed data harmonization and clinical-model evaluation. Yuanyi Ding contributed model development, benchmark analysis, interpretation, and manuscript preparation. Atsuro Daida, Pengyue Yu, Tiancheng Zheng,

Naoto Kuroda, Shaun A. Hussain, and Eishi Asano contributed data curation, metadata verification, and clinical validation across the constituent datasets. Hiroki Nariai contributed clinical interpretation and supervision, and Vwani Roychowdhury directed and supervised the project.

Chapters 3 and 4 incorporate material from Y. Ding, S. L. S. Dunn, J. J. Sakon, Z. M. Aghajan, C. Duan, Y. Zhang, J. I. Berger, A. E. Rhone, K. V. Nourski, H. Kawasaki, M. A. Howard III, V. P. Roychowdhury, and I. Fried, “Reading specific memories from human neurons before and after sleep,” *bioRxiv*, 2025, doi: 10.1101/2025.07.01.662486. Yuanyi Ding, Soraya L. S. Dunn, and John J. Sakon contributed equally to the work. Zahra M. Aghajan, Vwani P. Roychowdhury, and Itzhak Fried conceptualized the study. John J. Sakon, Vwani P. Roychowdhury, and Itzhak Fried supervised it. Yuanyi Ding, Soraya L. S. Dunn, John J. Sakon, Chenda Duan, and Yipeng Zhang performed the formal analysis. Yuanyi Ding, Soraya L. S. Dunn, John J. Sakon, Zahra M. Aghajan, Vwani P. Roychowdhury, and Itzhak Fried developed the methodology. Soraya L. S. Dunn, John J. Sakon, Joel I. Berger, Ariane E. Rhone, and Kirill V. Nourski collected the data; Hiroto Kawasaki, Matthew A. Howard III, and Itzhak Fried performed the surgeries; Matthew A. Howard III, Vwani P. Roychowdhury, and Itzhak Fried acquired funding; and Yuanyi Ding, Soraya L. S. Dunn, John J. Sakon, Vwani P. Roychowdhury, and Itzhak Fried wrote the manuscript.

Chapters 3 and 5 incorporate material from Y. Ding, C. Duan, S. L. S. Dunn, J. J. Sakon, Y. Zhang, J. I. Berger, A. E. Rhone, C. M. Garcia, K. V. Nourski, H. Kawasaki, M. A. Howard III, I. Fried, and V. P. Roychowdhury, “Concept-aligned neuronal representation transfer across brain states,” preprint, 2026. Yuanyi Ding led the study’s

conceptual and methodological development, model implementation, formal analysis, visualization, and manuscript preparation. Chenda Duan and Yipeng Zhang contributed to method and software development, evaluation, and manuscript review. Soraya L. S. Dunn and John J. Sakon contributed experimental design, neuronal data processing, analysis, scientific interpretation, and manuscript review. Joel I. Berger, Ariane E. Rhone, Christopher M. Garcia, and Kirill V. Nourski contributed participant data collection and curation. Hiroto Kawasaki, Matthew A. Howard III, and Itzhak Fried contributed the clinical and surgical work; Itzhak Fried also supervised the neuroscience component. Vwani P. Roychowdhury directed and supervised the project and contributed to its conceptual development and manuscript preparation.

VITA

2012–2016 Bachelor's degree in Electrical and Computer Engineering

University of Wisconsin–Madison

2016–2019 Master's degree in Electrical and Computer Engineering

University of California, Los Angeles

2021–2026 Doctoral candidate in Electrical and Computer Engineering; Doctor of Philosophy expected 2026

University of California, Los Angeles

PUBLICATIONS

Peer-Reviewed Publications

A. Daida, S. Kanai, Y. Zhang, C. Duan, N. Kuroda, T. Monsoor, **Y. Ding**, K. Kaneko, J. Qiao, S. A. Hussain, A. Fallah, N. Salamon, R. Sankar, R. J. Staba, J. Engel Jr., W. Speier, E. Asano, V. Roychowdhury, and H. Nariai, "Developmental profile of physiological high-frequency oscillations in the human brain," *NeuroImage*, vol. 336, art. no. 122017, 2026. doi: 10.1016/j.neuroimage.2026.122017.

C. Duan, Y. Zhang, S. Kanai, **Y. Ding**, A. Daida, P. Yu, T. Zheng, N. Kuroda, S. A. Hussain, E. Asano, H. Nariai, and V. Roychowdhury, "Omni-iEEG: A large-scale, comprehensive iEEG dataset and benchmark for epilepsy research," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

S. Kanai, A. Daida, Y. Zhang, **Y. Ding**, T. Monsoor, J. X. Qiao, N. Salamon, E. Ronne, S. S. Ahn, S. A. Hussain, R. Sankar, A. Fallah, W. Speier, V. Roychowdhury, J. Engel Jr., R. J. Staba, and H. Nariai, "Optimizing responsive neurostimulation targeting based on interictal high-frequency oscillations and phase-amplitude coupling," *Epilepsia*, vol. 67, no. 2, pp. 862–877, 2026. doi: 10.1111/epi.18675.

Y. Zhang, A. Daida, L. Liu, N. Kuroda, **Y. Ding**, S. Oana, S. Kanai, T. Monsoor, C. Duan, S. A. Hussain, J. X. Qiao, N. Salamon, A. Fallah, M. S. Sim, R. Sankar, R. J. Staba, J. Engel Jr., E. Asano, V. Roychowdhury, and H. Nariai, "Self-supervised data-driven approach defines pathological high-frequency oscillations in epilepsy," *Epilepsia*, vol. 66, no. 11, pp. 4434–4450, 2025. doi: 10.1111/epi.18545.

Y. Zhang, **Y. Ding**, C. Duan, A. Daida, H. Nariai, and V. Roychowdhury, “Self-supervised distillation of legacy rule-based methods for enhanced EEG-based decision-making,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2025*, pp. 477–486, 2025. doi: 10.1007/978-3-032-04984-1_46.

A. Daida, S. Panchavati, S. Oana, S. Kanai, Y. Zhang, **Y. Ding**, R. R. Rajsekar, N. Salamon, R. Sankar, A. Fallah, S. A. Hussain, V. Roychowdhury, W. Speier, and H. Nariai, “Evidence of thalamocortical network activation during epileptic spasms: A thalamic stereotactic EEG study,” *Epilepsia*, vol. 66, no. 7, pp. 2407–2420, 2025. doi: 10.1111/epi.18349.

Y. Ding, Y. Zhang, C. Duan, A. Daida, Y. Zhang, S. Kanai, M. Lu, S. Hussain, R. J. Staba, H. Nariai, and V. Roychowdhury, “PyHFO 2.0: An open-source platform for deep learning-based clinical high-frequency oscillations analysis,” *Journal of Neural Engineering*, vol. 22, no. 5, art. no. 056040, 2025. doi: 10.1088/1741-2552/ae10e0.

A. Daida, **Y. Ding**, Y. Zhang, S. Oana, S. Panchavati, B. D. Edmonds, S. S. Ahn, N. Salamon, R. Sankar, A. Fallah, R. J. Staba, J. Engel Jr., W. Speier, V. Roychowdhury, and H. Nariai, “Fast ripple band high-frequency activity associated with thalamic sleep spindles in pediatric epilepsy,” *Clinical Neurophysiology*, vol. 173, pp. 241–251, 2025. doi: 10.1016/j.clinph.2025.01.011.

Y. Zhang, L. Liu, **Y. Ding**, X. Chen, T. Monsoor, A. Daida, S. Oana, S. Hussain, R. Sankar, A. Fallah, C. Santana-Gomez, J. Engel, R. J. Staba, W. Speier, J. Zhang, H. Nariai, and V. Roychowdhury, “PyHFO: Lightweight deep learning-powered end-to-end high-frequency oscillations analysis application,” *Journal of Neural Engineering*, vol. 21, no. 3, art. no. 036023, 2024. doi: 10.1088/1741-2552/ad4916.

Preprints

Y. Ding, S. L. S. Dunn, J. J. Sakon, Z. M. Aghajan, C. Duan, Y. Zhang, J. I. Berger, A. E. Rhone, K. V. Nourski, H. Kawasaki, M. A. Howard III, V. P. Roychowdhury, and I. Fried, “Reading specific memories from human neurons before and after sleep,” *bioRxiv*, 2025. doi: 10.1101/2025.07.01.662486.

Y. Ding, C. Duan, S. L. S. Dunn, J. J. Sakon, Y. Zhang, J. I. Berger, A. E. Rhone, C. M. Garcia, K. V. Nourski, H. Kawasaki, M. A. Howard III, I. Fried, and V. P. Roychowdhury, “Concept-aligned neuronal representation transfer across brain states,” preprint, 2026.

L. Qiu, **Y. Ding**, and L. He, “Recurrent neural networks with pre-trained language model embedding for slot filling task,” *arXiv preprint arXiv:1812.05199*, 2018.

Chapter 1 Introduction

1.1 The Scientific Problem: Meaningful Representations from Human Neural Activity

Human neural recordings offer a distinctive view of cognition and disease because they measure biological activity as it is generated rather than only its behavioral consequences. Neural decoding uses such recordings to predict external variables including stimuli, behavior, clinical state, or semantic content (Quian Quiroga and Panzeri, 2009; Glaser et al., 2020). Modern decoding systems have recovered information about visual objects, speech, music, and continuous language from multiple recording modalities (Horikawa and Kamitani, 2017; Metzger et al., 2023; Bellier et al., 2023; Tang et al., 2023). These results establish that neural activity contains information accessible to computational models. Predictive success alone, however, does not determine what structure the model has learned or how that structure relates to the neural phenomenon under study. A decoder may capture the intended phenomenon, exploit a correlate specific to the experiment, or depend on regularities that disappear when the participant, institution, task, or brain state changes.

This dissertation is driven by one central question:

How can machine learning recover and evaluate meaningful representations from human neural activity as the scientific target progresses from clinical neural events to episodic content across experience, recall, and sleep?

Here, meaningfulness is treated as a matter of empirical support rather than an intrinsic property of a learned latent space. A representation is considered more

meaningful when multiple forms of evidence support its relationship to the scientific target and its relevance under the conditions being studied. Operationally, three criteria organize that evidence: the representation relates to the intended neural or behavioral target; it remains informative under the changes relevant to the scientific question; and its interpretation is supported by evidence that is not identical to the supervision used to learn it. These criteria do not require one model or metric to apply to every target. They provide a common basis for asking what a learned representation captures, where it remains useful, and how strongly its interpretation is supported.

1.2 Why Human Neural Data Make This Problem Difficult

Human electrophysiology provides complementary views at different recording scales. Scalp electroencephalography measures population-level electrical activity noninvasively but offers limited spatial specificity. Intracranial electroencephalography records local field potentials directly from implanted electrodes, and microwires extending from clinical depth electrodes can additionally capture action potentials from individual neurons and local neuronal populations (Fried et al., 1999; Nagahama et al., 2023). These measurements support questions ranging from pathological events in field potentials to content-selective neuronal responses and population activity during perception and memory. They nevertheless sample only a small and clinically determined part of the brain. Neural signals are noisy and nonstationary; electrode coverage differs across participants; recording hardware, sampling, referencing, and preprocessing vary across institutions; and conventional single-unit analysis can require substantial post hoc curation (Buccino et al., 2022). Every model therefore learns from an incomplete and participant-specific observation of a larger biological system.

Reliable supervision is also scarce. Clinical biomarkers require specialized annotation, and experts may disagree about the presence or pathological significance of individual events. Rule-based detectors can provide broad coverage but also produce false positives. As the target shifts from pathology to cognition, the form of supervision changes rather than simply becoming smaller. Movie viewing supplies external events that can be annotated densely; free recall supplies sparse and temporally indirect verbal reports; sleep supplies no continuous behavioral account of which content, if any, is active. The neural distribution also changes across these conditions because sensory input, behavior, and sleep physiology differ.

These constraints limit what predictive performance can mean. A model evaluated within one recording process, annotation convention, or brain state may rely on structure that is unavailable elsewhere. Claims about a neural representation must therefore be interpreted in relation to what was recorded, which evidence supervised learning, and which observations were reserved to test the resulting interpretation. The methodological task is not merely to obtain more neural data, but to learn from the evidence that is available without claiming more than that evidence can support.

1.3 Representation Learning as the Dissertation's Approach

This dissertation treats the learned representation—not only the final prediction—as an object of scientific study. Representation learning transforms raw observations so that selected structure becomes accessible to downstream analysis (Bengio et al., 2013). In neural data, a representation may range from a hand-engineered spectral description to a latent variable learned by a deep network. The studies here focus primarily on learned representations while retaining clinical rules and features when they encode useful prior

knowledge. Legacy detectors can define a sensitive candidate space and preserve interpretability, while learned models can discover recurring structure that fixed thresholds do not capture. The central issue is therefore not whether human-defined and learned representations compete, but how their different sources of information are assigned appropriate roles.

The dissertation uses several learning settings because the available evidence changes with the scientific target. Supervised learning directly optimizes a defined target when suitable labels exist. Weak supervision uses labels or proxies that are incomplete, noisy, or defined at a different level from individual examples. Self-supervised learning derives structure from the observations themselves, and contrastive learning organizes a representation through relations among matched and unmatched observations. Domain adaptation further asks whether learned structure can remain informative when source and target recordings follow different distributions (Ben-David et al., 2010; Chen et al., 2020). These approaches are complementary tools rather than successive claims that one form of learning universally replaces another.

Evaluation must likewise match the scientific question. A clinical event representation may be tested through morphology, expert review, anatomical localization, surgical outcome, and performance across heterogeneous cohorts. A memory representation may instead be tested through matching behavioral reports or content-selective neuronal activity outside the condition used for supervised learning. Cross-state analysis does not seek to erase all differences among movie viewing, recall, and sleep, because those differences are biologically meaningful. It asks whether target-related structure remains informative while state-specific neural dynamics change. This

relationship between the target and the evidence available to learn and evaluate it determines the progression of the dissertation.

1.4 From Clinical Events to Episodic Content

The apparent distance between epilepsy biomarkers and episodic memory can obscure their shared computational problem. In both settings, a high-dimensional neural signal is only indirectly related to the scientific target, reliable supervision is unevenly available, and a model may exploit structure specific to the dataset rather than the intended phenomenon. High-frequency oscillations (HFOs) provide a comparatively constrained starting point. Established detectors can propose candidate events with operational start times, channels, and time-frequency morphology, yet artifacts, physiological oscillations, disagreement over pathological definitions, and variability in expert annotation make their clinical relevance difficult to determine at scale (Gotman, 2010; Zijlmans et al., 2012; Frauscher et al., 2018). This setting isolates the problem of learning clinically relevant structure within an observation space that existing signal-processing knowledge has already partially defined.

Episodic content presents a more abstract target. An episodic memory concerns an event situated in a particular experiential context rather than a waveform defined by a local threshold (Tulving, 1972). Retrieval can reinstate content associated with that experience, while consolidation engages distributed interactions between the hippocampal formation and neocortex (Wang and Morris, 2010; Brodt et al., 2023). Concept-selective neurons provide direct evidence that human medial temporal lobe activity can represent particular people, places, or objects across sensory forms and can reactivate before matching free recall (Quian Quiroga et al., 2005; Gelbard-Sagiv et al.,

2008). Such neurons are sparse, however, and cannot provide dense labels for the many elements of a naturalistic experience. Population-level models are therefore needed to examine content beyond the concepts represented by individually identified cells.

Sleep creates the most weakly observed condition in this progression. Memory traces are thought to reactivate and reorganize during systems consolidation while neural dynamics differ substantially from waking perception and recall (Rasch and Born, 2013; Brodt et al., 2023). A model signal aligned with content-specific evidence during recall or sleep would support a stronger claim than movie decoding alone, but the absence of continuous behavioral labels makes that claim harder to test. The dissertation thus follows one question across targets of increasing abstraction: first learning structure among clinically proposed events, then decoding content tied to a single experience, and finally testing whether concept-related structure transfers to internally generated states. This progression also supports a longer-term translational vision: closed-loop systems that detect memory-related activity during sleep and use model-guided stimulation to influence consolidation. Human studies have already shown that stimulation timed to sleep-related hippocampal–prefrontal activity can alter neural synchrony and improve later memory, establishing causal intervention during sleep as a plausible, though still early, translational direction (Geva-Sagiv et al., 2023). A model-guided system targeting specific memory content would additionally require reliable access to human neural signals, a population-level estimator of content-related activity, verification during behaviorally observable recall, and an independent strategy for evaluating model signals during sleep. The present dissertation develops and tests these prerequisites at the levels of representation

learning, decoding, and cross-state verification; the design and causal evaluation of memory-enhancing stimulation remain future directions.

1.5 Data Design and Empirical Identifiability

The progression toward more abstract targets changes not only the learning method but also the observations required to make a claim testable. In human neuroscience, data are not a neutral container presented to an algorithm. The recording paradigm determines which aspects of a phenomenon are observed; annotations and auxiliary variables determine which relations can supervise learning; held-out measurements determine which interpretations can be evaluated; and the sampling design determines which forms of variation the resulting claim can address. This dissertation uses *empirical identifiability* to describe the practical condition in which a target has been operationalized through sufficient observations to test a model's relationship to it. The term does not imply formal identifiability of statistical parameters or unique recovery of an underlying biological mechanism.

Four dimensions organize this data-design problem: how the target is operationalized, what evidence supervises learning, how evaluation evidence is separated from that supervision, and across which participants, institutions, tasks, or brain states the representation is expected to remain informative. Clinical studies may begin with recognizable event candidates but still require harmonized metadata and common benchmarks before performance can be compared across sources. Episodic-memory studies must instead anchor neural activity to a specific experience, use behavior to observe later retrieval, and introduce an auxiliary neural reference when behavioral reports disappear. These considerations motivate the central argument developed in

Chapter 3: claims about meaningful neural representations cannot be separated from the data paradigm that makes their target observable and their interpretation testable. The four research questions below apply that argument to a progression from imperfect clinical supervision to cross-state memory representations.

1.6 Specific Research Questions and Dissertation Strategy

The central question introduced in Section 1.1 is divided into four research questions, each defined by a distinct scientific target, source of supervision, and form of evaluation:

1. **Research Question 1:** How can clinically meaningful neural-event representations be learned when expert labels are scarce, pathological definitions remain uncertain, and legacy detectors provide useful but noisy prior knowledge? Chapter 2 addresses this question through HFO candidate events, interactive clinical tools, weak supervision, and self-supervised morphology learning.
2. **Research Question 2:** What observations, annotations, and evaluation evidence are required to make neural representation targets comparable, observable, and learnable as those targets become more abstract? Chapter 3 addresses this question across clinical and memory settings, examining how the required data and evaluation designs change with the scientific target.
3. **Research Question 3:** Can specific content from a single naturalistic episode be decoded from neuronal population activity during subsequent recall? Chapter 4 formulates this problem and evaluates a participant-specific population decoder before and after sleep.

4. **Research Question 4:** Can concept-level neural representations transfer across movie viewing, recall, and sleep when supervision becomes sparse or absent? Chapter 5 addresses this question through concept-aligned learning, cross-state adaptation, and evaluation using independent concept-selective neuronal activity.

The questions are ordered by dependency rather than chronology. Chapter 2 asks how to learn when candidate instances exist but supervision is imperfect. Chapter 3 examines the observations that make such learning and evaluation possible, first through multi-center clinical comparability and then through a purpose-designed memory paradigm. In Research Question 2, “required” refers to the evidential roles needed to support each claim as operationalized in these studies, not to a universal or minimal set of observations for all neural representation problems. Chapter 4 uses the resulting memory paradigm to test recall decoding, and the state dependence and absence of sleep labels left unresolved by that analysis motivate Chapter 5.

1.7 Contributions and Organization

1. **Learning from imperfect clinical supervision:** Computational methods and open tools that use legacy detectors and expert knowledge to define candidate events, learn morphological structure, support annotation, and refine the identification of pathological neural events without requiring exhaustive expert labels.
2. **Data design and empirical identifiability:** Harmonized clinical resources, benchmark tasks, and standardized evaluation for multi-center intracranial EEG, together with experimental paradigms that connect naturalistic experience, recall, sleep, semantic annotations, and single-neuron activity.

3. **Episodic-content decoding:** A population-level framework that learns from a single naturalistic viewing experience and evaluates internally generated recall, extending analysis beyond individually identified concept cells or paradigms requiring repeated presentations.
4. **Cross-state concept alignment:** Concept-aligned representation learning that separates direct source-state decoding from the geometry of a reusable concept space and addresses the mismatch between densely labeled perception and sparsely labeled or unlabeled internal states.
5. **Evaluation without behavioral labels:** The use of held-out concept-selective neurons as independent, content-specific references for testing whether model-derived concept signals align with concept-selective neuronal activity during sleep.

Together, these contributions connect clinical signal analysis, data design, neural decoding, and memory neuroscience through a single progression of representation-learning problems. Chapter 2 develops learning under imperfect clinical supervision. Chapter 3 establishes the data and evaluation conditions required for comparable clinical targets and observable episodic content. Chapter 4 tests whether specific content from one experience can be decoded during recall, and Chapter 5 tests whether concept-aligned structure transfers across movie viewing, recall, and sleep. Chapter 6 synthesizes the scientific conclusions and unresolved generalization problems, and Chapter 7 returns to the central question.

Figure 1.1 summarizes this progression while emphasizing that the commonality lies in the representation-learning problem and its evidential structure, not in one latent space shared across all studies.

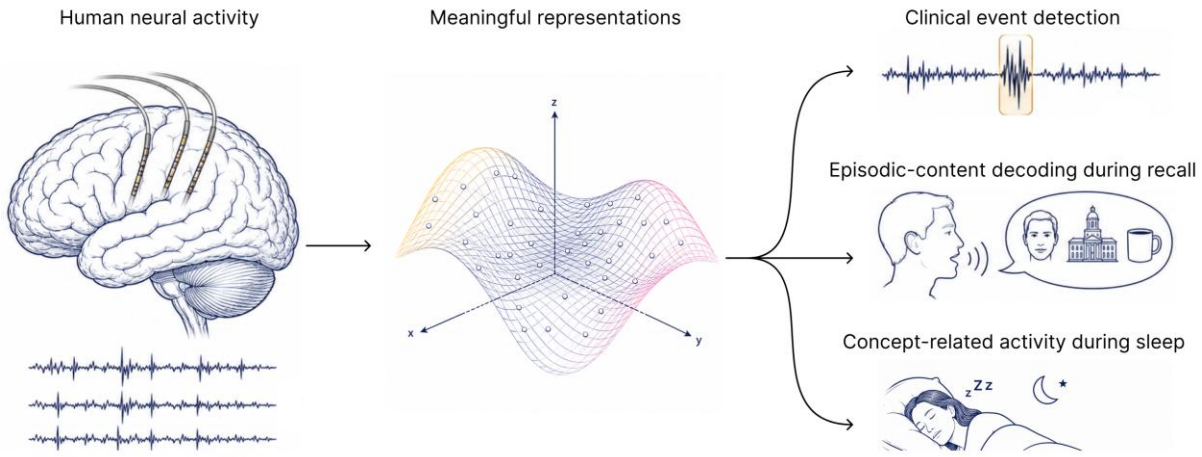

Figure 1.1. Meaningful representations from human neural activity. This dissertation studies how representations learned from intracranial human neural activity can be evaluated and used across distinct scientific targets. The central surface symbolizes the common representation-learning problem, while each application uses a representation learned for its own target and evidence. Across the studies, these representations support identification and refinement of clinical events, decoding of episodic content during recall, and testing of concept-related activity during sleep. The figure highlights representation learning as the framework connecting clinical neural-signal analysis, recall decoding, and cross-state sleep analysis.

Chapter 2 Learning Neural Representations from Imperfect Clinical Supervision

2.1 Clinical Setting and Scientific Question

This chapter begins with a comparatively constrained representation problem: learning the structure of clinically relevant neural events in intracranial electroencephalography (iEEG). High-frequency oscillations (HFOs) are brief, high-frequency events that have been investigated as biomarkers of epileptogenic tissue and as potential aids to surgical planning (Jacobs et al., 2010; Zijlmans et al., 2012). Their apparent concreteness makes them a useful starting point for this dissertation. An HFO candidate can be assigned an operational start time, end time, recording channel, and time-frequency morphology. Yet operational detectability does not make its clinical meaning self-evident. A candidate event may reflect pathological activity, physiological

activity, or background and artifact, and there remains no universally accepted event-level definition of a pathological HFO.

This distinction between detecting an event and interpreting it creates the central learning problem. Established signal-processing methods can scan long, multichannel recordings with high sensitivity and return candidate HFOs, but their outputs include false positives that require manual review. Expert annotation is costly, does not scale to hundreds of thousands of events, and is itself an imperfect reference because visual HFO assessment exhibits inter-rater variability (Nariai et al., 2018; Spring et al., 2017). Fully supervised learning does not remove this ambiguity; it transfers the ambiguity into a labeled training set and can make model performance dependent on a small cohort, a particular detector, or one institution's annotation practice.

The scientific question is therefore: **can clinically meaningful neural-event representations be learned when expert labels are scarce, pathological definitions remain uncertain, and legacy detectors provide useful but noisy prior knowledge?** This chapter tests the claim that clinically relevant event representations can be learned without exhaustive event-level labels by separating candidate generation, self-supervised morphology learning, weak clinical orientation, and evaluation through converging morphological, anatomical, and patient-level evidence. Legacy algorithms contribute accumulated clinical and signal-processing knowledge without being treated as sources of ground-truth labels; self-supervised learning models structure within their candidate space; and clinical variables orient and evaluate that structure without being interpreted as exact event-level annotations. HFOs thus allow the central representation-learning

problem to be studied in a setting where candidate instances can already be proposed by signal-level detectors.

2.2 Why Existing Detection and Supervision Are Insufficient

Computational HFO analysis commonly begins with rule-based detectors such as short-term energy (STE), the Montreal Neurological Institute (MNI) detector, and the Hilbert detector. Although their definitions differ, these methods encode explicit assumptions about frequency, amplitude, duration, and local signal statistics. Their chief value for the present problem is recall: they reduce a continuous recording to a tractable collection of candidate intervals while retaining a large share of events that may be clinically relevant (Zelmann et al., 2012). Their chief limitation is semantic specificity. Crossing a detector threshold does not establish that an event is a genuine HFO, much less that it is pathological.

Machine-learning classifiers can refine this candidate set by removing artifacts or separating HFO subtypes. In a conventional supervised pipeline, however, each refinement stage requires a target label. This produces two related bottlenecks. First, generating enough annotations requires specialized clinical labor. Second, more annotation does not necessarily resolve uncertainty in the construct being labeled. Experts may disagree about whether an event is present, and the distinction between physiological and pathological HFOs is not reducible to a universally agreed visual rule. A model trained on such labels can reproduce an annotation convention without necessarily learning a representation that is clinically informative outside the training cohort.

Weak supervision offers a possible alternative, but it still requires a defensible source of evidence. Treating detector outputs as pseudo-labels would merely reproduce the original detector. Treating surgical resection as an event-level label would also be too strong: resected tissue may contain both pathological and non-pathological activity, while preserved tissue may include events whose relevance is uncertain. The methodological gap is therefore more specific than a shortage of labels. What is needed is a way to preserve the broad candidate space contributed by legacy detectors, discover recurring event morphology without labels, and use patient-level clinical evidence to orient that structure without claiming that the evidence provides exact event-level ground truth.

2.3 Clinical Knowledge as a Candidate Space

The candidate-space interpretation is instantiated in the PyHFO 2.0 workflow (Ding et al., 2025a). The workflow keeps three stages conceptually distinct: rule-based detectors propose candidate events; learned models assign provisional categories; and users inspect, accept, or revise those predictions. PyHFO 2.0 integrates STE, MNI, and Hilbert detection with classifiers for artifact rejection, spike-associated HFOs (spkHFOs), and epileptogenic HFOs (eHFOs). Its scientific role in this chapter is narrower than that of a representation-learning method: it defines a reproducible process for generating the candidate space, preserving the origin of model outputs, and returning uncertain cases to expert review. This provenance is important because a final event designation can be traced to the recording, the detector that proposed the interval, the classifier that refined it, and any subsequent human revision. Candidate definition is thereby separated from learned refinement; changing a classifier need not silently change the rules that generated the event pool. The interface supports this separation by presenting predictions

as suggestions alongside the signal views needed for expert assessment rather than as irreversible labels.

Evaluation across human iEEG from UCLA and Zurich and a smaller rodent dataset showed that the implemented detectors closely reproduced their reference counterparts when preprocessing was held constant. The addition of the Hilbert detector also tested whether existing artifact and spkHFO classifiers remained useful on candidates generated by a different detection rule. Against independent annotations from five UCLA patients, both classifiers retained high F1 scores. Runtime benchmarking further established that a broad candidate space could be processed at practical scale, but computational efficiency is supporting evidence for the workflow rather than the scientific contribution pursued in the remainder of this chapter.

PyHFO 2.0 therefore establishes the candidate-generation and expert-review context for the representation-learning problem; it does not solve that problem. Its pretrained classifiers still depend on an existing source of supervision, while the candidate space remains heterogeneous when reliable event-level labels are unavailable. The next section addresses this unresolved issue by treating detector outputs not as ground truth, but as an unlabeled collection from which event morphology and provisional supervision can be learned.

Figure 2.1 summarizes the staged PyHFO 2.0 workflow and the provenance preserved between candidate generation, model refinement, and expert review.

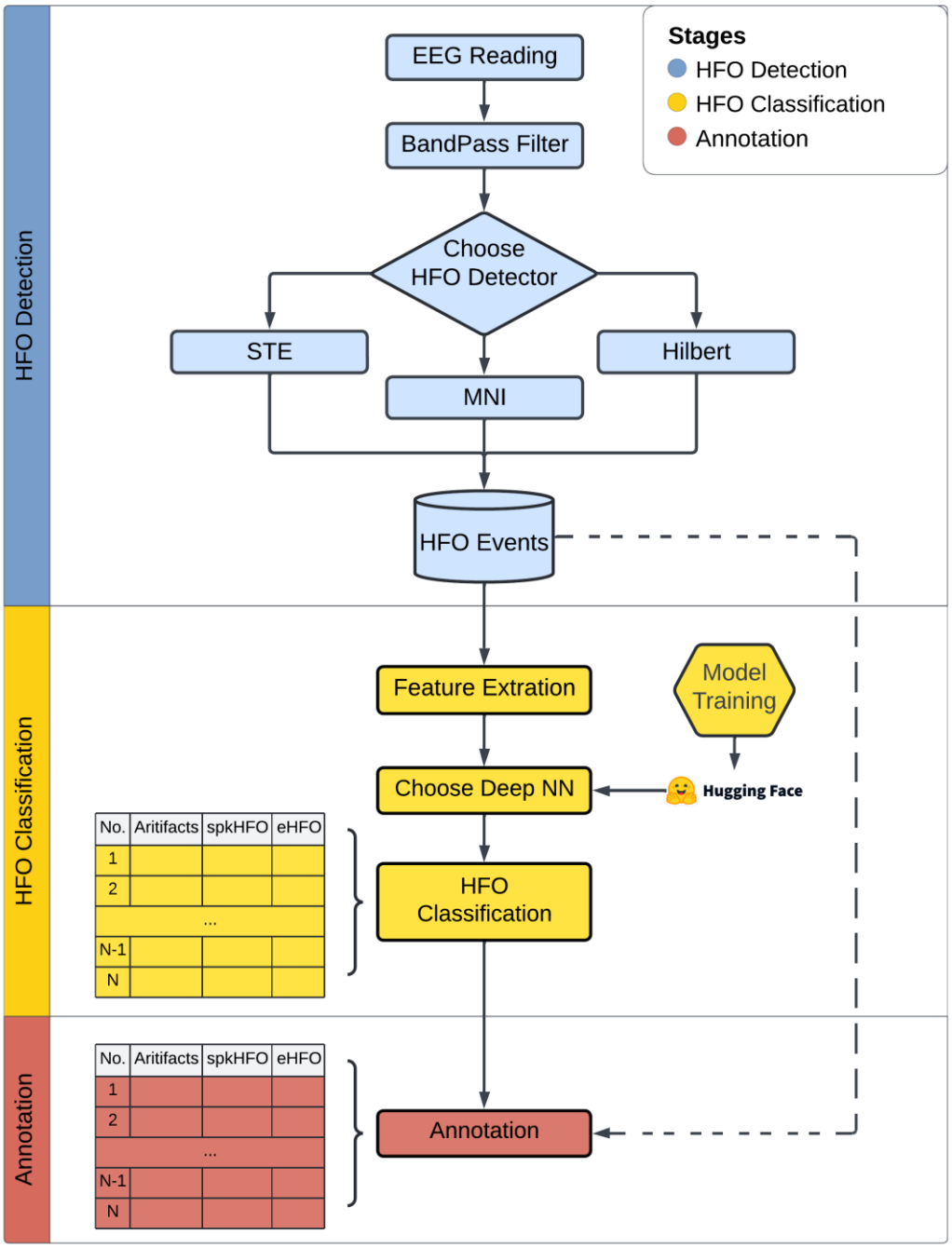

Figure 2.1. PyHFO 2.0 workflow for HFO detection, classification, and annotation. After EEG loading and band-pass filtering, STE, MNI, or Hilbert detection defines a candidate-event pool. Feature extraction and pretrained or custom deep neural networks then provide provisional artifact, spkHFO, and eHFO classifications, whose outputs remain available for interactive expert review and revision. The staged workflow separates candidate definition, model-based refinement, and final annotation. Reproduced from Ding et al. (2025a), Figure 1.

2.4 Self-Supervised Morphological Representation Learning

Starting from this unlabeled candidate space, the core methodological answer is Self-Supervised to Label Discovery (SS2LD) (Zhang et al., 2025). SS2LD does not assume that a detector or an existing classifier has already solved the interpretation problem. Instead, STE and MNI detections supply a large and heterogeneous pool containing pathological HFOs, physiological HFOs, and background events. The method asks whether recurring morphology within this pool can first reveal a useful representation and then support the discovery of provisional labels without exhaustive event-level annotation.

For patient k, let $X_c^{(k)}$ denote the iEEG signal on channel c. A detector proposes event $h_i = X_c^{(k)}[s_i : e_i]$, with start and end times s_i and e_i. Because event durations vary, a fixed window w_i is centered on the event midpoint:

$$w_i = X_c^{(k)}[\frac{s_i + e_i}{2} - L : \frac{s_i + e_i}{2} + L].$$

A Morlet transform converts this signal window into a time-frequency representation W_i. In the reported implementation, each input spans 570 ms and 10–290 Hz and is resized to 64×64. The SS2LD convolutional variational autoencoder (VAE) with residual layers maps W_i to a 16-dimensional latent distribution with mean $\mu_\theta(w_i)$ and covariance $\Sigma_\theta(w_i)$ (Zhang et al., 2025). A latent sample

$$z_i \sim \mathrm{N}(\mu_\theta(w_i), \Sigma_\theta(w_i))$$

is decoded to reconstruct the event's time-frequency morphology. Rather than measuring reconstruction only at the pixel level, SS2LD uses a perceptual loss computed from feature maps of a pretrained VGG16 network (Johnson et al., 2016):

$$\mathrm{L}_{\mathrm{perceptual}} = \sum_{l} \quad \|\phi_l(W_i) - \phi_l(\hat{W}_i)\|_2^2.$$

This objective emphasizes structured morphological similarity across multiple feature scales. The learnable β-weighted VAE objective reported in SS2LD balances reconstruction with regularization of the posterior toward a standard normal prior (Zhang et al., 2025):

$$\mathrm{L}_{\mathrm{pretrain}} = (1-\beta)\mathrm{L}_{\mathrm{perceptual}} + \beta \mathrm{L}_{\mathrm{KL}},$$

$$\beta_{\mathrm{new}} = \beta + \eta_\beta(\mathrm{E}[\mathrm{L}_{\mathrm{KL}}] - \mathrm{E}[\mathrm{L}_{\mathrm{perceptual}}]).$$

The learned space is converted into weak supervision through two-stage hierarchical clustering. First, k-means with $k = 2$ separates candidate HFO morphology from heterogeneous background events; the cluster with greater VAE reconstruction error is assigned as background. Second, the remaining events are clustered again with $k = 2$. The cluster occurring predominantly on channels later included in the surgical resection is oriented as pathological, while the other is treated as physiological. This procedure uses resection information as group-level evidence to interpret an emergent morphological partition, not as an exact label for every event. Events in the pathological cluster receive weak label $l_i = 1$; physiological and background events receive $l_i = 0$.

Clustering alone imposes a simple boundary on a latent distribution that may have more complex geometry. SS2LD therefore freezes the pretrained VAE and trains a classification head F_ψ on the latent mean. The VAE also generates a surrogate reconstruction $\hat{W}_i$, which is encoded as $\hat{\mu}_i$. Assuming that the reconstructed morphology retains the event category, the classifier is optimized on both observed and generated examples:

$$L_{classifier} = BCE\ (F_\psi(\mu_i), l_i) + BCE\ (F_\psi(\hat{\mu_i}), l_i).$$

The three stages play different roles. Self-supervised reconstruction learns recurring morphology from all candidates; clustering discovers a provisional organization and connects it to clinical evidence; discriminative refinement and VAE augmentation produce a smoother decision boundary. The result is not label-free in the absolute sense. Rather, it relocates supervision from exhaustive expert event labeling to detector-defined candidates, recording morphology, and weaker patient-level clinical evidence.

Figure 2.2 shows how these stages connect the detector-defined candidate pool to the learned morphological space, weak-label discovery, and final classifier.

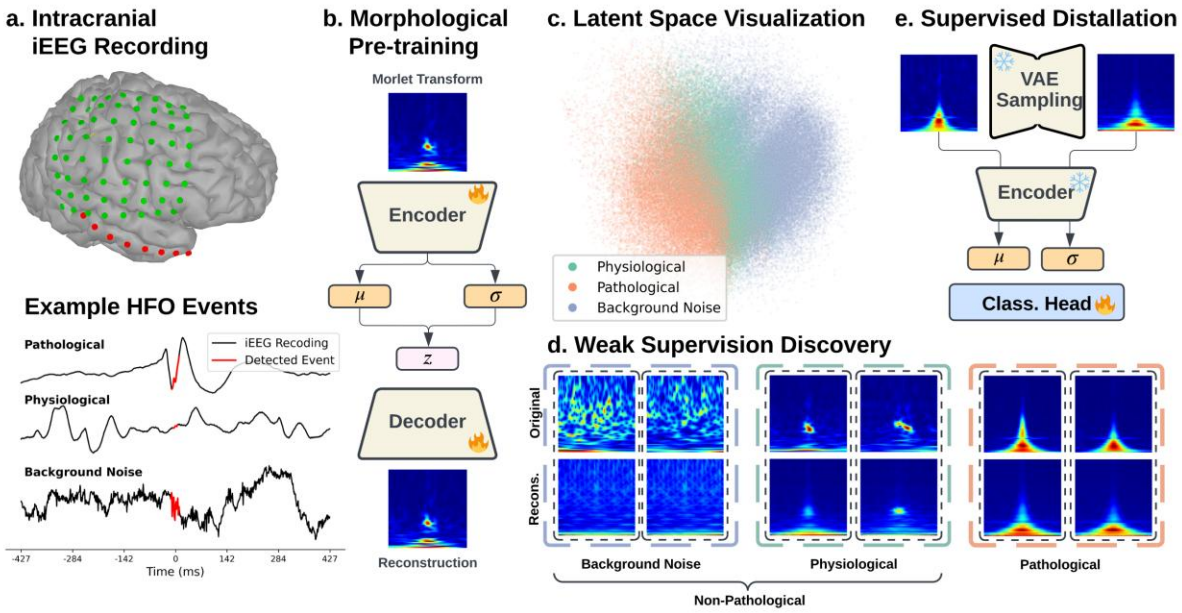

Figure 2.2. Overview of the proposed SS2LD framework. (a) HFO events are detected in interictal intracranial EEG recordings using rule-based detectors and include pathological, physiological, and background-noise events. (b) A VAE is pretrained to reconstruct HFO morphological features derived using the Morlet transform, with the encoder mapping inputs to latent distributions and the decoder reconstructing sampled latent codes. (c) Latent representations are projected using principal component analysis to visualize clustered event types. (d) Hierarchical clustering identifies distinct morphological groups, providing weak supervision for pathological-event classification. (e) A classifier is trained on both original latent codes and VAE-generated samples to refine the decision boundary. Reproduced from Zhang et al. (2025), Figure 1.

2.5 Does the Learned Representation Capture Clinically Relevant Structure?

The evaluation cannot rely on authoritative pathological labels because their absence motivates the method. It therefore tests the representation through converging evidence: patient-level surgical outcomes, the anatomical distribution of predicted events, comparison with existing pathological-HFO classifiers, ablation of the proposed learning stages, cross-institutional evaluation, and inspection of the latent dimensions. This design evaluates whether the learned event partition is clinically useful without presenting any single proxy as definitive ground truth.

2.5.1 Cohorts and Evaluation Protocol

The principal Open iEEG cohort contains recordings from 185 patients with epilepsy at UCLA and Detroit and 686,410 candidate HFOs generated by STE and MNI. Surgical outcomes are available for 162 patients, including 113 who became seizure-free, and each channel is annotated as resected or preserved. Evaluation uses subject-wise five-fold cross-validation so that events from the same patient do not appear in both training and test sets. Within each fold, the reported implementation assigns 119 patients to training, 30 to validation, and 36 to testing while balancing institutional representation. This chapter introduces the resource only to the extent required to evaluate SS2LD; Chapter 3 examines the construction, harmonization, and benchmarking of multi-center clinical data as a scientific problem in its own right.

External evidence is provided by the Zurich iEEG cohort. Bipolar rereferencing and quality control yield 15 patients with valid resection annotations, including 10 who became seizure-free, and 97,511 STE- and MNI-detected HFOs. The model is evaluated on all valid events across recording runs. The difference in institution, acquisition, referencing,

and patient composition makes Zurich a test of transport beyond the development cohort, although the modest number of patients limits the precision of conclusions about generalization.

The VAE is pretrained for 100 epochs with Adam, a learning rate of 10^{-3}, weight decay of 10^{-5}, and a batch size of 512. To prevent patients with many detections from dominating representation learning, sampling is capped at 2,500 events per patient per epoch. The refinement classifier is trained for nine epochs with a learning rate of 3×10^{-4}, weight decay of 10^{-5}, and a batch size of 4,096.

2.5.2 Clinical Evidence Without Event-Level Ground Truth

For patient k, the resection ratio measures the proportion of events predicted to be pathological that occur on resected channels:

$$RR_k = \frac{\text{predicted pathological HFOs on resected channels}}{\text{all predicted pathological HFOs}}.$$

A value near one indicates that the predicted pathological-event burden was largely removed, whereas a value near zero indicates little overlap between the prediction and resection. A class-balanced logistic regression uses RR_k to predict postoperative seizure freedom. Accuracy and F1 score are reported because approximately two thirds of the patients with known outcome were seizure-free.

A complementary measure examines preserved tissue in patients who became seizure-free. If surgery succeeded without removing a region, events remaining in that region provide evidence against labeling most of its activity as pathological. Patient-level specificity is therefore defined as

$$Specificity_k = \frac{\text{predicted non-pathological HFOs on preserved channels}}{\text{all detected HFOs on preserved channels}},$$

and averaged across seizure-free patients. Neither measure proves the status of an individual event. Together, however, they test two clinically relevant expectations: predicted pathological events should be preferentially removed in successful surgery, and preserved tissue in seizure-free patients should not be saturated with pathological predictions.

2.5.3 Main Comparison

SS2LD is compared with two existing refinement strategies: eHFO, a weakly supervised classifier based on clinical evidence, and spkHFO, a classifier trained from expert annotations of HFOs associated with spike discharges. Across both cohorts, SS2LD produces the highest outcome-prediction accuracy, F1 score, and specificity, as summarized in Table 2.1.

Table 2.1. Clinical evaluation of pathological-HFO refinement. Values reproduce the SS2LD paper's five-fold evaluation. The zero standard deviations reported for the two Zurich baselines arise because their fixed pretrained decision boundaries produce the same external-cohort predictions across folds; they should not be interpreted as uncertainty-free estimates.

Cohort	Method	Outcome accuracy	Outcome F1	Specificity
Open iEEG	eHFO	0.538 ± 0.119	0.397 ± 0.069	0.543
Open iEEG	spkHFO	0.594 ± 0.129	0.434 ± 0.070	0.703
Open iEEG	**SS2LD**	**0.612±0.131**	**0.464±0.069**	**0.749**
Zurich iEEG	eHFO	0.533 ± 0.000	0.364 ± 0.000	0.745
Zurich iEEG	spkHFO	0.600 ± 0.000	0.500 ± 0.000	0.819
Zurich iEEG	**SS2LD**	**0.640±0.037**	**0.583±0.050**	**0.909**

These improvements are consistent across the internal multi-institutional cohort and the external Zurich cohort. The absolute outcome-prediction values remain moderate, which is important to retain in interpretation: a model that improves a clinically relevant proxy has not thereby solved seizure-outcome prediction. The result instead supports the

narrower claim that self-supervised morphology and discovered weak supervision identify a more clinically aligned subset of detector candidates than the evaluated alternatives.

2.5.4 Ablation and Representation Analysis

Ablation separates the effects of label discovery, discriminative refinement, VAE augmentation, and latent dimensionality. Weak labels produced directly by clustering achieve an accuracy of 0.557 and an F1 score of 0.423 on Open iEEG. Training the classifier without generated augmentation increases these values to 0.568 and 0.433. Adding augmentation in the full 16-dimensional SS2LD model increases them to 0.612 and 0.464, respectively. Table 2.2 shows that the discovered clusters contain useful information, but that refining their boundary and augmenting the latent examples are both needed for the strongest patient-level result.

Table 2.2. SS2LD ablation on Open iEEG. The source paper reports slightly different fold variability for the 16-dimensional full model in its main-comparison and ablation tables. The values are retained here as reported in the source table rather than silently reconciled across analyses.

Model variant	Latent dimensions	Outcome accuracy	Outcome F1	Specificity
Weak supervision only	16	0.557 ± 0.101	0.423 ± 0.066	0.739
Classifier, no VAE augmentation	16	0.568 ± 0.116	0.433 ± 0.058	0.715
Full SS2LD	8	0.562 ± 0.134	0.426 ± 0.065	0.736
Full SS2LD	**16**	**0.612±0.108**	**0.464±0.063**	**0.749**
Full SS2LD	32	0.593 ± 0.128	0.442 ± 0.065	0.761
Full SS2LD	64	0.574 ± 0.126	0.431 ± 0.061	0.752

Eight latent dimensions appear too restrictive to retain the morphology needed for outcome-aligned discrimination. Increasing the dimensionality beyond 16 raises specificity slightly but lowers accuracy and F1 score, consistent with a more conservative classifier that predicts fewer events as pathological. This tradeoff is scientifically relevant:

maximizing specificity alone could produce a clinically unhelpful representation if it removes too many potentially informative events.

The decoder permits a qualitative test of what individual latent dimensions encode. Starting from the mean latent code of predicted pathological events, each dimension is varied between its 0.001 and 0.999 empirical percentiles and decoded into a time-frequency representation. The resulting interpolations modulate spectral intensity, peak frequency, and the presence of additional event-like components. In a complementary knockout analysis, individual dimensions are set to zero before recalculating the projected distribution and classifier predictions. The first dimension produces the largest disruption to the decision boundary and varies between background-like and pathological-event morphology. These analyses do not assign a unique biological mechanism to each coordinate, but they show that the decision is linked to structured, inspectable time-frequency variation rather than only to an opaque classification score.

Figure 2.3 visualizes this interpretation through latent-dimension interpolation and knockout analysis.

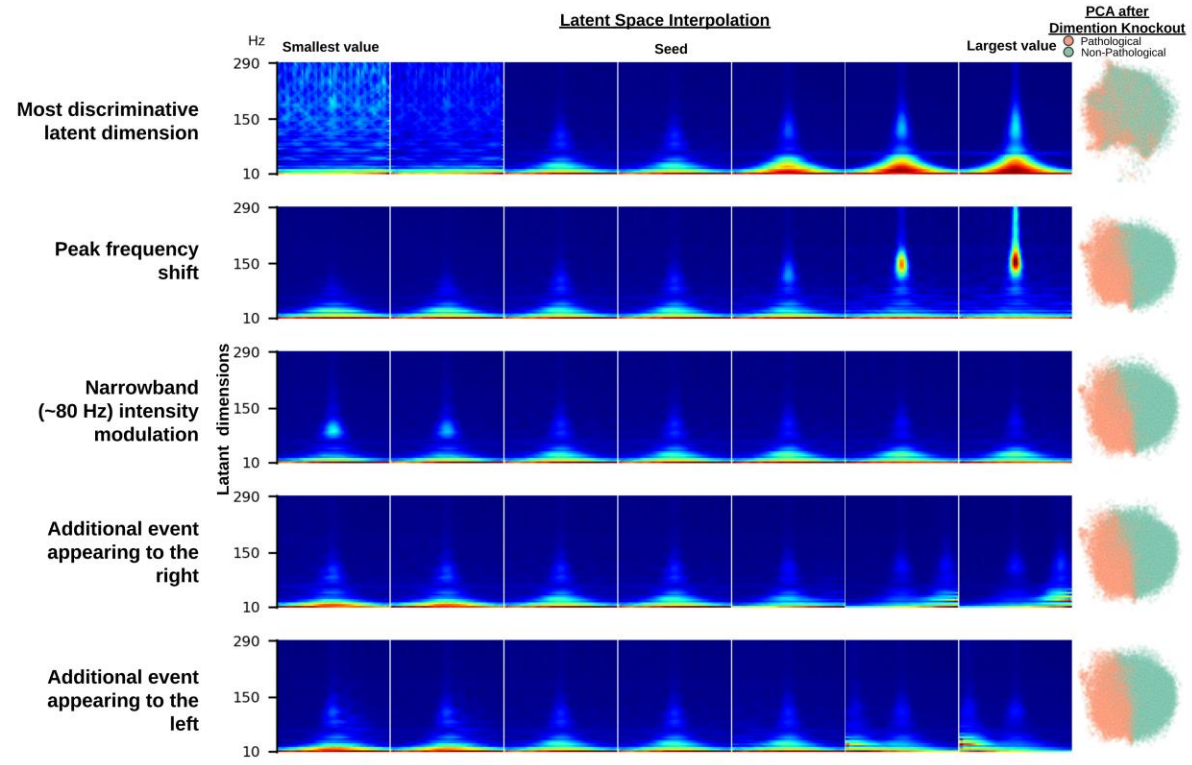

Figure 2.3. Latent-dimension interpolation and dimension knockout. The visualization focuses on the five most expressive latent dimensions in one representative fold. Each row on the left corresponds to a latent dimension and is annotated with its inferred role in encoding HFO morphology, including spectral-intensity variation, peak-frequency shifts, and additional events. The central spectrogram in each row is the unperturbed seed derived from the mean latent code of predicted pathological events. The first dimension spans a prominent transition between background-noise-like and pathological-event morphology. On the right, principal component analysis is recomputed after each latent dimension is set to zero, with points colored by classifier-predicted pathological or non-pathological status. Greater mixing after knockout indicates a stronger contribution to the learned decision boundary; the first dimension produces the largest effect. Reproduced from Zhang et al. (2025), Figure 2.

2.6 Scientific Interpretation

Together, these studies support the chapter's central claim: clinically relevant event representations can be learned without exhaustive event-level labels by separating candidate generation, self-supervised morphology learning, weak clinical orientation, and evaluation through converging morphological, anatomical, and patient-level evidence. Legacy detectors constrain the search space using accumulated domain and signal-processing knowledge. Self-supervised learning identifies recurring morphology within

that space, while resection location and postoperative outcome provide indirect evidence about its clinical relevance. PyHFO 2.0 demonstrates how model predictions can remain inspectable and revisable within an expert-facing workflow; SS2LD was evaluated as an offline representation-learning pipeline rather than as a deployed PyHFO 2.0 component. No component is asked to provide evidence it cannot support: detector thresholds establish candidacy, morphology establishes similarity, clinical variables establish indirect relevance, and experts retain interpretive authority. The principal contribution is therefore not simply a new HFO detector, classifier, or software platform, but a redistribution of supervision. Expert knowledge is concentrated in defining candidate-generation procedures, interpreting the discovered structure, and testing clinical consequences rather than in exhaustively labeling every event. This distinction also clarifies the meaning of “self-supervised” in this setting: the representation is learned from the signal itself, but the overall scientific inference remains grounded in clinical design and evidence.

The results further suggest that representation quality cannot be judged by event-classification performance alone. Because authoritative event labels are unavailable, the strongest evaluation comes from alignment among multiple forms of evidence: separable and decodable morphology, improvement over supervised and weakly supervised alternatives, stability across institutions, anatomical consistency with successful resection, and compatibility with expert review. Each is incomplete on its own, but their convergence supports the conclusion that useful clinical structure can be recovered from imperfectly supervised neural events.

2.7 Remaining Limitations and the Need for Designed Observations

Several limitations bound this conclusion. SS2LD can only learn from events admitted by STE and MNI, so clinically relevant activity missed by both detectors is absent from the representation. Its time-frequency input is restricted to 10–290 Hz and 570 ms, potentially excluding informative slower context, higher-frequency activity, or longer temporal organization. The clustering procedure depends on the morphology captured by the VAE and on resection location as an orienting clinical prior. Resection is not a pure event-level label, surgical outcome is influenced by factors beyond HFO burden, and the Zurich evaluation contains only 15 valid patients. The latent interpolation provides morphological interpretability but does not establish a one-to-one mapping between latent dimensions and neurophysiological mechanisms. The implementation also remains centered on HFOs, image-like time-frequency inputs, supported clinical file formats, and offline processing. Underrepresented recording domains may require model adaptation, and real-time or closed-loop applications would require a different systems design. Future work should evaluate broader frequency ranges, longer temporal context, additional biomarkers, more diverse institutions, and alternative representations that do not require converting a neural signal into an image.

Most importantly for the dissertation's argument, the chapter's conclusion is conditional on a candidate space and sources of clinical evidence that already exist. Clinical practice specifies what is recorded; legacy detectors determine where candidate events are sought; surgery supplies anatomical and outcome evidence; and experts can inspect the resulting events. These conditions allow Chapter 2 to treat representation learning under imperfect supervision as its primary problem. Chapter 3 examines those

conditions rather than continuing to take them for granted. It first asks how an existing clinical target, its candidate events, and its evaluation evidence can be made comparable when recordings and metadata differ across institutions. It then considers the more fundamental case in which target instances cannot be enumerated by a signal-level detector and the observation paradigm itself must make them empirically identifiable. The unresolved question is therefore: **what observations, annotations, and evaluation evidence are required to make increasingly abstract neural representation targets comparable, observable, and learnable?**

Chapter 3 Designing Data for Representation Learning

Chapter 2 began with a target that clinical practice had already made partially observable. Intracranial recordings supplied the signal, legacy detectors proposed candidate intervals, surgical variables supplied indirect clinical evidence, and experts could inspect individual events. The central difficulty was learning within that imperfectly supervised candidate space. Chapter 3 asks what happens when these supporting conditions are unstable or absent. In the clinical setting, the candidate events and proxies exist, but their definitions and availability vary across institutions. In the memory setting, the target instances cannot be enumerated by a signal-level detector, so the experiment must determine what counts as an observation of the target, which annotations can supervise learning, and what evidence can evaluate the learned structure.

This chapter therefore asks: **what observations, annotations, and evaluation evidence are required to make increasingly abstract neural representation targets empirically identifiable and learnable?** Here, *empirical identifiability* means that the

recording and study design provide sufficient observations to operationalize a target and test a model's relationship to it. The term does not imply formal identifiability of statistical parameters, nor does it imply that one dataset can uniquely determine the underlying biological mechanism. It instead emphasizes a practical scientific requirement: performance can support a representational claim only when the target, the available supervision, and the evaluation evidence have been made explicit.

Two cases develop this argument. The first concerns a clinical target that is already recognized but is measured inconsistently across institutions. Omni-iEEG shows how harmonization, expert-validated metadata, shared annotations, and clinically grounded benchmarks can make such claims comparable. The second concerns episodic content, whose instances cannot be enumerated by a signal-level detector. The Movie–Recall–Sleep studies show how a naturalistic experience, time-resolved semantic annotations, behavioral memory tests, neural recordings, and limited independent neuronal references can make progressively less observable expressions of that content available for study. These cases are not linked because epilepsy data directly produce a memory experiment. They are linked because both demonstrate that a dataset is part of the scientific argument rather than a neutral container presented to a learning algorithm.

3.1 From Given Events to Designed Observations

To compare how different data paradigms support representation-learning claims, this chapter analyzes each setting along four dimensions. **Target operationalization** asks how a theoretical target is connected to observable events, variables, or reference measurements; this is a problem of scientific measurement rather than a purely technical choice. **Learning supervision** identifies which labels, proxies, or unlabeled relations

constrain the representation. **Evaluation evidence** asks what observations test the resulting interpretation and how far they are separated from the information used for training or selection. **Relevant variation** specifies the participants, institutions, tasks, or brain states across which the representation is expected to remain informative. This final dimension matters because the interpretation of a decoder depends on the level of generalization it actually achieves, not on predictive performance in the abstract (Koh et al., 2021). These dimensions form a comparative framework synthesized from the five studies in this dissertation rather than a universal checklist for all neural datasets. They make explicit which parts of a representational claim are supplied by data design, which are learned by the model, and which remain to be tested. In Chapter 2, for example, detector-defined morphology supported representation learning, while resection and outcome variables provided indirect clinical evidence. When the target changes, the same scientific roles may require different observations.

The two scientific cases introduced above give rise to three distinct observation regimes. Multi-center clinical analysis begins with candidate events and clinical variables that already exist but require harmonization. Episodic recall provides sparse behavioral reports of internally generated content, whereas sleep provides no concurrent behavioral report. Recall and sleep are therefore considered separately because the same semantic target becomes observable through fundamentally different forms of evidence across these states. Table 3.1 compares the roles played by target operationalization, learning supervision, evaluation evidence, and relevant variation in each regime.

Table 3.1. Comparative framework for the empirical identifiability of neural representation targets. The rows distinguish three observation regimes rather than three equivalent datasets or scientific domains.

Setting	Target operationalization	Learning supervision	Evaluation evidence and separation	Relevant variation
Multi-center clinical events	Detector-defined candidate events linked to harmonized clinical variables	Expert annotations or clinically constructed proxy labels	Held-out participants and clinical evidence not reducible to event-level training labels	Across participants and contributing institutions within the assembled resource
Behaviorally anchored episodic recall	Semantically annotated content from a naturalistic experience linked to later behavioral recall events	Dense movie-viewing annotations only; recall labels are not used for training	Held-out recall vocalizations together with content and temporal controls	From externally observed experience to internally generated recall
Sleep without concurrent behavioral reports	Experience-linked semantic content examined in a state without concurrent behavioral reports	Dense movie-viewing annotations plus unlabeled recall and NREM inputs used for consistency and adaptation	A held-out concept-selective neuronal reference excluded from model input and from the adaptation interval	From wake states to sleep within participant

The framework exposes two connected design problems. When a target is already operationally recognizable, its observations and evidence may still need to be made comparable across institutions. When the target cannot be enumerated by a detector, the experiment must create the observations that make its content learnable and its interpretation testable.

3.2 Making Existing Clinical Targets Comparable Across Institutions

3.2.1 Why More Data Alone Are Insufficient

Public iEEG datasets have made it possible to study epilepsy with machine learning, but their independent origins create a second layer of uncertainty beyond the signals themselves. Recordings differ in file format, sampling rate, montage, channel naming, duration, device, and available clinical metadata. Seizure-onset zones, resection status, surgical outcome, and anatomical labels may be represented differently or omitted. Studies also define tasks, splits, and metrics independently. Under these

conditions, a difference in reported model performance may reflect the model, the cohort, the preprocessing pipeline, the clinical label definition, or the evaluation rule.

This section deliberately returns to the same clinical target studied in Chapter 2 but asks a different scientific question. Chapter 2 examined how to learn meaningful structure within a given HFO candidate space under imperfect supervision; the present section asks how evidence about that target can be made comparable when recordings, metadata, and evaluation conventions differ across institutions. SS2LD could be evaluated across multiple cohorts because the required event, resection, and outcome variables were available, but a broader claim of clinical generalizability requires those variables to be aligned systematically. Adding patients without reconciling their metadata would increase sample size while leaving the scientific target unstable. The design objective is therefore not merely aggregation. It is to preserve source-level information while constructing a common representation of the recordings, clinical variables, and evaluation tasks.

3.2.2 Omni-iEEG Construction and Harmonization

Omni-iEEG was constructed to address this problem (Duan et al., 2026). The resource brings together pre-surgical iEEG from independently released sources, including Open iEEG, Zurich iEEG HFO, HUP, and Epilepsy Interictal/SourceSink collections. Across the contributing resources, it spans 302 patients, 178 hours of high-resolution recordings, and data originating from eight epilepsy centers. After selection and quality control, the released benchmark split contains 255 patients, including 215 with resective surgery and 155 who became seizure-free according to the dataset summary. These nested counts should remain explicit: the full assembled resource and the records

eligible for a particular benchmark are not interchangeable populations. Table 3.2 preserves this distinction while summarizing the contribution of each source.

Table 3.2. Composition of the Omni-iEEG cohort by contributing source. "Total subjects" denotes all participants in each source, whereas "Selected" denotes those included in the released benchmark split. Training and testing rows from the source table are aggregated here; the split itself remains patient-disjoint. Adapted from Duan et al. (2026), Table 1.

Dataset	Total subjects	Selected	Resective surgery	Seizure-free	Recording duration (min)
Open iEEG	185	185	164	114	5–120
Zurich	20	20	15	15	1–12
HUP	58	15	15	10	5
SourceSink	39	35	21	16	1–17
Total	**302**	**255**	**215**	**155**	—

Each source was cross-referenced with its associated publication, and recordings were prioritized when reliable alignment between the signal and clinical channel metadata could be established. Board-certified clinicians reviewed the data and reconciled metadata into a unified schema, including seizure-onset zone, resection, outcome, and available anatomical information. Original sampling rates are retained in the released recordings to preserve provenance. For benchmark model inputs, signals are reproducibly resampled to 1,000 Hz, with the original rate stored in the metadata and a shared resampling procedure supplied. Training and test data are separated at the patient level using an approximately 60/40 split, with contributing sources distributed across both partitions and clinically relevant attributes (including outcome, channel count, and recording modality) considered during stratification.

Harmonization does not make the centers identical, nor should it erase clinically meaningful variation. Its function is to document that variation and prevent representational differences introduced by naming, formatting, or preprocessing

conventions from being confused with biological differences. The same principle applies to data splitting: a standardized split does not prove external generalization, but it makes comparisons among models on the available cohort reproducible.

3.2.3 Annotations and Clinically Grounded Benchmarks

Omni-iEEG supplies event-level evidence that is rarely available at this scale. STE, MNI, and Hilbert detectors were used to construct a broad candidate HFO pool across institutions. Four board-certified epileptologists then annotated candidates as artifact, pathological spkHFO, or non-spkHFO. The resulting benchmark contains 36,177 annotated events from 49 patients: 9,288 artifacts, 7,709 non-spkHFOs, and 19,180 spkHFOs. Because the classes are imbalanced, pathological-event classification is evaluated with macro-averaged precision, recall, F1 score, and one-versus-rest AUC. This task asks whether a model can recover an expert-defined event distinction across a multi-center candidate pool; it does not claim that spkHFO is the only possible definition of pathological activity.

The second primary benchmark moves from events to brain regions. Its labels are constructed from clinical evidence with different levels of certainty. Seizure-onset-zone channels are treated as positive examples of pathological tissue. Preserved channels in patients who became seizure-free are treated as high-confidence negative examples, because the successful resection left those regions intact. Resected non-SOZ channels remain ambiguous, since surgical margins reflect anatomical and practical considerations, while residual pathological tissue in non-seizure-free patients cannot be localized exactly. The task therefore evaluates models using two complementary criteria: discrimination between high-confidence pathological and normal channels, and

association between predicted pathological burden in the resection and postoperative outcome.

The channel-level benchmark contains 252 patients and 21,250 channels, including 2,162 pathological and 19,088 normal channels under the benchmark definitions. Precision, recall, specificity, F1 score, and AUC quantify different error consequences at the channel level. Patient-level outcome evaluation computes the proportion of predicted pathological activity located in resected channels and tests whether that resection ratio is associated with seizure freedom. Reporting both levels is necessary because a model can correlate with outcome while still producing clinically undesirable false-positive or false-negative channel predictions.

Three exploratory tasks are included: anatomical-location classification, ictal-period identification, and sleep–awake classification, extending the shared resource to other clinically meaningful distinctions. Because some required annotations are available from only one or two contributing sources, these tasks do not provide the same degree of cross-center evidence as the primary benchmarks. Their inclusion illustrates an important design constraint: benchmark scope must follow the actual coverage of the metadata, not only the questions researchers would like to ask.

Figure 3.1 summarizes the assembled resource, its primary and exploratory tasks, and the clinical evidence used to interpret model outputs.

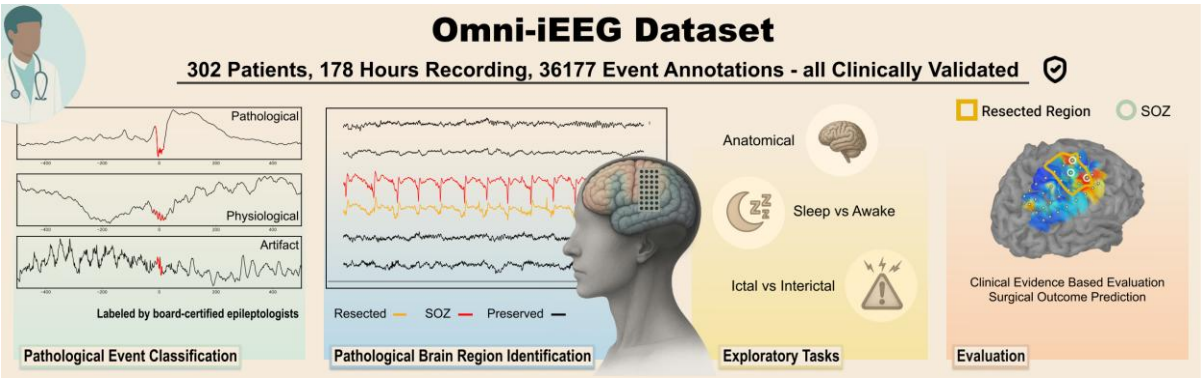

Figure 3.1. Overview of the Omni-iEEG dataset and benchmark. The resource comprises 302 patients, 178 hours of recordings, and 36,177 expert-annotated events. Its primary tasks evaluate pathological-event classification and pathological-brain-region identification; exploratory tasks address anatomical location, sleep versus wakefulness, and ictal versus interictal activity. Clinical evaluation relates predicted pathological activity to seizure-onset zones, resected and preserved regions, and postoperative surgical outcome. Reproduced from Duan et al. (2026), Figure 1.

3.2.4 What the Benchmark Design Reveals

The pathological-event benchmark compares models operating on raw time-domain traces with a clinically informed time-frequency pipeline. PyHFO-Omni achieves the strongest reported event-level performance, with a macro-F1 score of 0.8061 and macro-AUC of 0.9390. PatchTST, the strongest of the reported raw time-series baselines, reaches a macro-F1 score of 0.7726 and macro-AUC of 0.9311. Table 3.3 reports the shared comparison. These results do not imply that one representation class is universally superior. They show that, under a common candidate pool, annotation protocol, split, and metric, the contribution of representation choice can be evaluated without changing the clinical target at the same time.

Table 3.3. Omni-iEEG pathological-event classification benchmark. Reproduced from Duan et al. (2026), Table 4.

Event representation/model	Macro precision	Macro recall	Macro F1	Macro AUC
LSTM + attention	0.7352	0.7359	0.7338	0.9109
PatchTST	0.7757	0.7686	0.7726	0.9311
TimesNet	0.7589	0.7726	0.7652	0.9221
PyHFO-Omni	**0.8025**	**0.8110**	**0.8061**	**0.9390**

For pathological-region identification, the benchmark compares event-rate biomarkers with models trained directly on one-minute iEEG segments. Table 3.4 retains the three metrics needed to distinguish channel-level discrimination from patient-level outcome association.

Table 3.4. Selected metrics from the Omni-iEEG pathological-brain-region benchmark. Channel AUC and specificity evaluate discrimination between clinically defined pathological and normal channels, whereas outcome AUC evaluates whether predicted pathological burden in resected channels is associated with postoperative seizure freedom. Adapted from Duan et al. (2026), Table 5.

Model	Representation	Channel AUC	Specificity	Outcome AUC
eHFO	Event-based	0.6611	0.4101	0.4521
PyHFO-spkHFO	Event-based	0.6557	0.4089	0.4972
PyHFO-Omni-spkHFO	Event-based	0.7351	0.6951	**0.7438**
SEEG-NET	Segment-based	0.7850	0.6049	0.5952
CLAP	Segment-based	0.7684	0.7823	0.6770
TimeConv-CNN	Segment-based	**0.8061**	**0.8230**	0.7380

PyHFO-Omni attains an outcome AUC of 0.7438, while the end-to-end TimeConv-CNN reaches a similar outcome AUC of 0.7380 but a higher channel-level AUC and specificity. The audio-pretrained CLAP baseline also produces competitive channel-level performance after fine-tuning. These comparisons show why a single summary metric is insufficient. Similar outcome association can coexist with materially different channel-level discrimination, and a representation learned outside neurophysiology can serve as

a useful baseline without establishing that it captures the same clinical structure as an HFO-based model.

The multi-center benchmark also exposes poor transport of some publicly available event classifiers trained on single-center data. This observation strengthens the central argument: performance within a local cohort does not by itself identify which aspects of the representation will remain informative under institutional heterogeneity. At the same time, Omni-iEEG does not eliminate this uncertainty. Its held-out test split contains data from the same assembled sources as training; it provides stronger within-resource multi-center comparison, not prospective validation at a previously unseen hospital.

3.2.5 From Comparable Clinical Targets to Observable Cognitive Content

Omni-iEEG demonstrates how harmonization, clinical metadata, shared annotations, and standardized benchmarks can make an existing clinical target comparable across institutions. This solution, however, assumes that the relevant recordings, candidate events, and clinical reference variables can already be defined. More abstract cognitive targets do not begin with such an event catalog. A particular episodic memory has no detector-defined start and end interval in a continuous neural recording, and its later occurrence cannot be confirmed from morphology alone. The next design problem is therefore not how to harmonize already recognized instances, but how to create the observations that connect neural activity to the content of a specific experience. In terms of the four-part framework, Omni-iEEG standardizes the operational definition and supervision of existing clinical targets, separates participants used for training and testing, and evaluates models across heterogeneous contributing sources

under common tasks and metrics. Its relevant variation is nevertheless bounded by the assembled resource: it supports multi-source comparison, not prospective generalization to a hospital absent from both training and test data.

3.3 Making Episodic Content Observable

3.3.1 Anchoring Neural Activity to a Single Naturalistic Experience

Episodic content presents a different data-design problem. Repeated presentations of isolated images or words can generate many labeled trials, but repetition changes the experience being studied and may encourage learning of a stimulus category rather than a unique episode. A naturalistic movie provides a compromise between experimental structure and episodic richness: all participants encounter a temporally defined experience containing interacting people, places, actions, and abstract themes, while the continuous audiovisual sequence can be annotated and synchronized with the neural recording.

In the memory-decoding study *Reading Specific Memories from Human Neurons Before and After Sleep* (Ding et al., 2025b), ten participants undergoing clinical monitoring for drug-resistant epilepsy watched a single 42-minute episode of the television series *24*. Two participants were female; ages ranged from 24 to 55 years, with a reported mean of 36.9 ± 3.9 years. The protocols were approved by the institutional review boards of UCLA and the University of Iowa, and all participants provided written informed consent. The episode was manually annotated for eight recurring concepts spanning characters, locations, and abstract content. Labels were defined as multi-label vectors because several concepts could be present simultaneously. Annotation was initially performed at one-second resolution and mapped to 250-ms intervals aligned with the neural inputs.

Episode audio recorded during the experiment was aligned with the original soundtrack so that the semantic labels and neural recordings shared a participant-specific time base. Concept selection balanced two requirements: sufficient presence during the movie to support learning and sufficient mention during recall to permit later evaluation. The selected vocabulary therefore does not exhaust the episode's content and should not be interpreted as a complete ontology of the memory. Importantly, the grouping of recall expressions into the eight concepts was completed before model training and was not revised in response to model output. This temporal separation reduces the risk that the target definition is adjusted to favor a particular decoder.

3.3.2 Behavioral Evidence of Encoding and Retrieval

Movie annotations identify what was externally present, but they do not establish that a participant encoded or later retrieved that content. The paradigm therefore includes three complementary memory tests before and after overnight sleep. During free recall, participants verbally described everything they could remember about the episode. During recognition, they judged 150 short clips presented in random order: 75 clips from the viewed episode and 75 foils from another episode of the same series. During cued recall, participants answered six image-based questions; two participants completed free recall but not the cued-recall component.

The tests play different evidential roles. Recognition verifies memory for presented material without requiring spontaneous verbal production. Free recall provides an unconstrained sample of internally retrieved content. Cued recall increases the opportunity to observe retrieval of material that may not arise spontaneously. Recall audio was recorded at 32 kHz, and automated alignment followed by manual curation supplied

timestamps for individual vocalizations. References, synonyms, and unambiguous pronouns were mapped to the prespecified concept vocabulary, producing sparse behavioral markers tied to the original experience.

Participants recognized the episode reliably: mean old/new accuracy was $81.0\% \pm 7.0\%$, and every participant had $d' > 1$, with a group mean of 1.94 ± 0.54. These behavioral results establish that the episode was encoded well enough to support a later neural decoding question. They do not label the entire recall period. A vocalization provides evidence that a concept was retrieved by the time it was spoken, but it cannot specify the exact onset or duration of the preceding internal process. The design therefore provides sparse and temporally indirect retrieval evidence rather than frame-level ground truth.

3.3.3 Neural Observations Across Experience and Memory Tests

Electrode placement was determined entirely by clinical need. Participants were implanted with Behnke–Fried hybrid depth electrodes containing macrocontacts along the shaft and bundles of microwires at the electrode tips. Macro-iEEG was recorded from 0.1–500 Hz at 2 kHz, while broadband microwire potentials were recorded from 0.1 Hz–8 kHz at 32 kHz. The microwires provide access to population spiking with high temporal resolution across available medial temporal, frontal, and other cortical regions, but the number and anatomical distribution of channels vary across participants.

The study used thresholded, clusterless voltage-spike amplitude trains rather than requiring every event to be assigned to a manually curated single unit. Positive- and negative-going events were preserved separately, coincident bundle-wide artifacts were removed, and the resulting activity was summarized in 250-ms windows aligned with the

movie annotations. This choice makes a much larger fraction of the acquired microwire signal available and supports participant-specific models despite variable electrode coverage. It also defines the object learned in the next chapter: a representation of population-level threshold-crossing activity, not a complete census of identified neurons or a direct measurement of a memory engram.

Together, the movie, behavioral tests, and neural recordings form a chain of observations: the movie supplies a content-labeled experience; recognition verifies encoding at the participant level; vocal reports provide sparse evidence of later retrieval; and the neural recordings permit population activity during viewing and retrieval to be compared. None of these observations alone makes specific episodic content empirically identifiable; their coordination makes the question testable. Under the comparative framework, the movie operationalizes a limited content vocabulary and supplies dense source supervision, whereas vocal reports provide target-state evidence that is not used to train the movie decoder. The relevant variation is a change from audiovisual experience to internally generated recall within each participant. The paradigm does not test transfer to a new participant, a new episode, or an unrestricted set of memories.

Figure 3.2 presents the coordinated experimental design, behavioral evidence, recording coverage, clusterless population input, and semantic annotations that make the later decoding analysis possible.

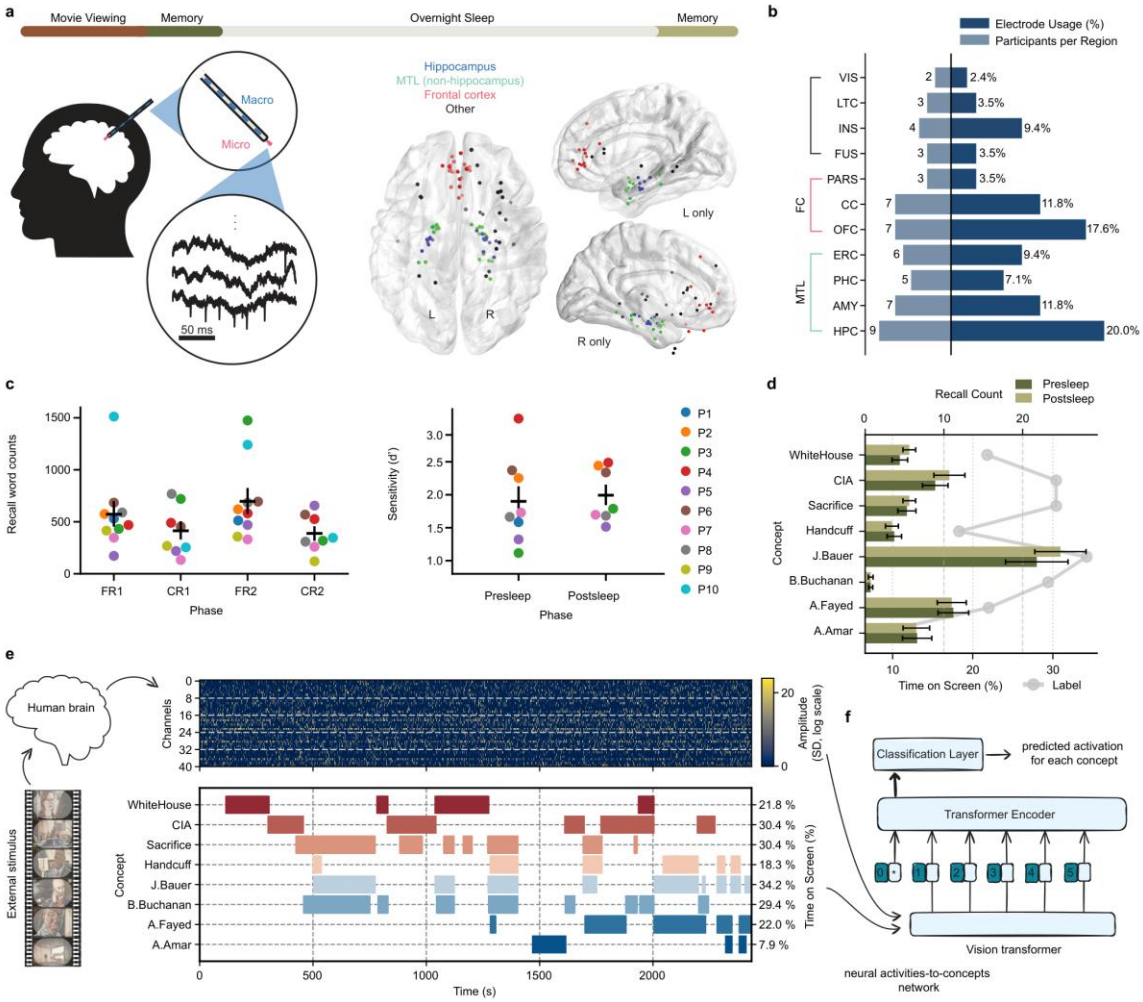

Figure 3.2. Experimental design, recordings, behavioral evidence, and content annotations in the single-episode memory study. (a) Participants viewed a movie, completed presleep memory tests, slept overnight, and repeated the tests the following morning while activity was recorded from clinically placed Behnke–Fried electrodes. (b) Microwire coverage varied across medial temporal, frontal, and other cortical regions. (c) Free- and cued-recall word counts and recognition sensitivity establish behavioral access to the episode before and after sleep. (d) Recall frequency and screen time characterize the selected concept vocabulary. (e) Clusterless microwire activity is synchronized with time-resolved, multi-label concept annotations during movie viewing. (f) A high-level schematic links neural activity to concept outputs; the decoder architecture and its evaluation are developed in Chapter 4. Reproduced from Ding et al. (2025b), Figure 1.

3.4 Observing Content Without Continuous Behavioral Reports

The pre- and postsleep tests allow memory-related behavior to be observed on both sides of an overnight interval. Sleep itself creates a more difficult evidential gap. Participants cannot provide a continuous report of which content, if any, is active at each

moment, and neural dynamics differ substantially from movie viewing and awake recall. A sleep recording is therefore not a labeled extension of the recall dataset. It supplies observations from a theoretically important state while removing the behavioral reference that made recall events interpretable.

The expanded study *Concept-Aligned Neuronal Representation Transfer Across Brain States* (Ding et al., 2026) extends the common paradigm to 13 participants with drug-resistant epilepsy, including five female participants and an age range of 21–55 years. The protocols were approved by the institutional review boards of UCLA and the University of Iowa, and all participants provided written informed consent. Participants viewed four movies, yielding 10.9 hours of movie viewing, 2.8 hours of free recall, and 58.4 hours of identified NREM sleep. Across participants, the recordings include 952 microwires, averaging approximately 73 per participant, and 452 curated single units. NREM intervals were identified from a clean macro-iEEG channel using slow-wave- and spindle-band power calculated in 30-second windows; a two-component Gaussian mixture separated NREM from REM and wake. This expanded cohort should not be merged numerically with the ten-participant single-episode cohort used in Chapter 4. The studies share a design logic, but they contain different participant sets, movie exposure, recording totals, preprocessing requirements, and downstream questions.

Content-specific neuronal activity provides one limited way to bridge the absence of continuous sleep reports. In this study, a concept-selective neuron is an individual unit whose firing is significantly associated with a particular person, place, or object under a prespecified movie-viewing test. Selectivity was assessed during events in which a single concept was visible. Units were required to show sufficient post-onset firing, pass a

permutation-based selectivity test, and respond significantly to only one tested concept. Sixty-two units met these criteria.

These units are valuable precisely because they are sparse and independent. They cannot label every concept in the movie, and their absence does not imply that a concept is missing from the broader population representation. For the concepts they do encode, however, their firing supplies a content-specific signal that is not derived from a participant's verbal report. To preserve independence, the channel containing an evaluated concept-selective unit is excluded from the population given to the model. Its activity is not used as a model input or as dense training supervision. This separation creates the possibility of asking whether a signal inferred from the remaining population coincides with an independently recorded unit selective for the same content.

At the data-design level, this held-out reference addresses leakage and circularity. It does not itself establish that every firing event during sleep is a memory reinstatement, nor does it specify how activity from the remaining population should be aligned across experimental conditions. Those are modeling and evaluation questions developed in Chapter 5. The contribution of the present section is to make clear why continuous sleep recordings alone are insufficient and what kind of independent observation is required for a content-specific claim to be testable in the absence of behavior. Here, target operationalization depends on linking NREM activity to the same experience and concept vocabulary established during movie viewing. Learning uses dense movie labels and unlabeled recall and sleep recordings, while evaluation uses a sparse concept-selective neuronal reference that is excluded from the model input and assessed on sleep data not used for adaptation. The relevant variation is therefore cross-state transfer within

participant; the design does not establish transfer across people or arbitrary semantic content.

3.5 A General Principle for Neural Data Design

The clinical and memory cases reveal a common structure. A representation target is not made meaningful by its label name alone. The study must specify how the target is operationalized, what evidence supervises the model, how evaluation evidence is separated from that supervision, and which changes the resulting claim is expected to survive. As the target becomes more abstract, these roles are less likely to be served by a single annotation.

Three principles follow. First, **the target and its evidence must be designed jointly**. Harmonized clinical metadata make cross-center comparisons possible because they define common clinical variables; movie annotations make experience-related content learnable because they anchor concepts to time. Second, **training supervision and scientific validation should be separated whenever possible**. Resection and outcome test an event representation at a different level from the morphological input used to learn it; vocal reports test a movie-trained signal outside the viewed episode; held-out selective neurons preserve an independent neural reference. Third, **increasing abstraction requires complementary observations rather than labels alone**. Clinical events can be proposed from local signal morphology, whereas episodic content requires coordinated experience, behavior, recording state, and content-specific references.

This framework also clarifies what the chapter does not claim. Omni-iEEG standardizes available clinical evidence but does not create a definitive ground truth for epileptogenic tissue. Movie annotations specify external content but do not determine the

neural code for that content. Vocalization times mark overt reports but not the exact temporal boundary of internal retrieval. Concept-selective neuronal firing offers a sparse reference but not a complete description of a distributed memory. Empirical identifiability is therefore graded: better data design narrows the set of plausible interpretations and enables stronger tests, but it does not remove all biological ambiguity.

3.6 Remaining Limitations and the Need for Content-Level Testing

The clinical resource remains retrospective and inherits selection effects from publicly available recordings and surgical practice. Metadata are incomplete for some centers, expert spkHFO annotations remain subjective despite standardized review, and not every task can be evaluated across every contributing source. The shared test set supports reproducible multi-center comparison within Omni-iEEG but is not equivalent to prospective deployment at a new institution. Clinical labels such as seizure-onset zone, preserved tissue in seizure-free patients, and resection ratio remain evidence-based proxies rather than direct measurements of the epileptogenic network.

The memory datasets are rare but small. Participants all underwent epilepsy monitoring, electrode placement was clinically determined, and coverage differs across individuals. The principal decoding cohort uses one episode and eight selected concepts, while the expanded cohort contains four movies but remains limited in participant number and recording diversity. Semantic annotations depend on human judgment; verbal recall incompletely samples remembered content; sleep lacks direct behavioral labels; and concept-selective neurons are available for only a subset of concepts and regions. These constraints limit population-level generalization and prevent causal claims about memory storage or consolidation.

The data designs developed here make stronger questions possible, but they do not answer those questions by themselves. In particular, the Movie–Recall–Sleep paradigm links a single experience to semantic annotations, subsequent behavior, and neuronal activity, but it does not establish that population activity contains sufficient information to recover which content a participant is retrieving. Chapter 4 therefore asks: **can specific content from a single naturalistic episode be decoded from neuronal population activity during subsequent recall?**

Chapter 4 Decoding Specific Episodic Content from Human Neuronal Populations

4.1 Scientific Question and Prior Gap

Chapter 3 established an experimental chain connecting a single naturalistic experience to semantic annotations, later behavioral reports, and simultaneous neuronal recordings. This design makes content-specific recall empirically testable, but it does not establish that a population model can recover that content when the external stimulus is no longer present. The central question of this chapter is therefore: **can specific content from a single naturalistic episode be decoded from distributed human neuronal populations during subsequent internally generated recall, and does the population evidence supporting that decoding change after overnight sleep?**

Previous studies provide several important but incomplete precedents. Neural activity can distinguish visual objects and movie characters when training and evaluation both occur during stimulus presentation (Horikawa and Kamitani, 2017; Zhang et al., 2023; Gerken et al., 2024). Human medial temporal neurons selective for particular people or places can reactivate before free recall, establishing that content-specific

neuronal signals accompany internally generated retrieval (Gelbard-Sagiv et al., 2008). Other decoding approaches recover visual or semantic content from neural recordings, but often require repeated stimuli, many hours of training data, pooled participants, preselected responsive neurons, or a constrained trial structure (Quian Quiroga et al., 2007; Li et al., 2024; Tang et al., 2023). These conditions do not answer whether a model trained during one exposure to an extended episode can identify the content of later verbal recall in the same individual.

The study *Reading Specific Memories from Human Neurons Before and After Sleep* addresses this gap (Ding et al., 2025b). Each participant-specific model is trained only on neuronal activity and semantic labels recorded during a single movie viewing. The model is then applied without retraining to neuronal activity recorded while the participant verbally recalls the episode. Recall labels are used to evaluate the temporal relationship between model output and behavior, not to train the decoder. This separation distinguishes the problem from held-out classification within the movie: the test condition contains internally generated retrieval without the audiovisual input that supplied the training labels. The claim pursued here is deliberately narrower than reconstructing a complete memory. The decoder operates over eight prespecified concepts that recur in the episode and in participants' verbal reports. Evidence for successful decoding means that model activity for one of these concepts rises preferentially before the participant recalls that concept relative to matched unrelated vocalizations. It does not mean that every detail of the episode, the subjective quality of recollection, or an unobserved memory trace has been read from the brain.

4.2 Formalizing One-Shot Memory Decoding

Here, *one-shot* refers to one uninterrupted viewing of the naturalistic episode, not to one labeled neural interval in the machine-learning sense. The single exposure still yields many temporally annotated 250-ms training windows, but the participant never repeats the episode to create additional presentations of the same events.

For participant p, the labeled movie-viewing data can be written as

$$\mathrm{D}_{\mathrm{movie}}^{(p)} = \{(X_t^{(p)}, \mathbf{y}_t^{(p)})\}_{t=1}^{T_{\mathrm{movie}}},$$

where $X_t^{(p)}$ contains neuronal population activity in a 250-ms window and $\mathbf{y}_t^{(p)} \in \{0,1\}^K$ is a multi-label vector indicating the presence of $K = 8$ episode concepts. Multiple entries of $\mathbf{y}_t$ may be active because characters, locations, and abstract content can occur together. A participant-specific decoder f_{θ_p} is trained to estimate

$$\hat{\mathbf{y}}_t^{(p)} = f_{\theta_p}(X_t^{(p)}).$$

Recall provides a different type of dataset:

$$\mathrm{D}_{\mathrm{recall}}^{(p)} = \{X_t^{(p)}\}_{t=1}^{T_{\mathrm{recall}}},$$

accompanied by sparse behavioral timestamps $\mathrm{V}_c^{(p)} = \{\tau_{c,1}, \ldots, \tau_{c,n_c}\}$ at which participant p vocalized concept c. Free- and cued-recall recordings are pooled when both are available; P1 and P2 contribute free recall only. The recall recordings do not contain dense labels for every 250-ms interval. The trained movie decoder is applied across the recall period to generate activation $a_c^{(p)}(t) \in [0,1]$ for each concept independently. The scientific hypothesis is not that movie and recall activity are identical. It is that some content-related population structure learned during the experience remains sufficiently informative that $a_c(t)$ rises before later vocalization of the matching concept.

This formulation creates a stringent distribution change. Movie viewing combines perceptual responses, audiovisual structure, attention, and memory formation. Recall lacks the original stimulus and includes internal search, speech planning, and vocal production. Postsleep recall may add physiological reorganization or recording drift. Because the available recall labels are sparse and behaviorally delayed, standard frame-level accuracy cannot serve as the primary evaluation. The decoder instead requires a statistic that asks whether its continuous output is selectively related to subsequent matching recall.

Chapter 3 described the ten participants, episode, behavioral tasks, semantic annotations, and recording hardware. The present chapter uses the resulting clusterless microwire activity. Broadband microwire signals are notch-filtered to attenuate line noise and high-pass-filtered at 300 Hz. For each channel, signal amplitude SD is calculated separately in each task epoch; the minimum SD across epochs defines a shared symmetric threshold of ± 3SD, which is then applied to every epoch. Positive- and negative-going threshold crossings are retained as separate amplitude trains. Events occurring synchronously on at least six of eight microwires within a bundle and a 4-ms interval are removed as likely artifacts. Amplitudes are discretized in 0.5-SD intervals from 3.5 to 30.5 times the detection SD, summed within nonoverlapping 5-ms bins, and represented as 50 temporal bins within each 250-ms interval. The binned amplitudes are z-scored within each recording bundle over the experimental recording, separately for the two polarities. The resulting input for participant p is

$$X_t^{(p)} \in \mathrm{R}^{2\times N_e^{(p)}\times 50},$$

where $N_e^{(p)}$ varies with clinically determined electrode coverage. Participant-specific modeling is therefore a design requirement rather than a claim that neural codes are identical across people.

4.3 Participant-Specific Population Decoder

The decoder is organized around the anatomical grouping of the microwires. Channels are partitioned into recording bundles, and each bundle is processed by a region-embedding module that converts local spike-amplitude patterns into a sequence of feature patches with positional information. A learnable class token summarizes each region. Region-wise self-attention then models relationships among the temporal and channel features within a bundle, while cross-region attention allows each regional summary to incorporate information from the other recorded regions. The final regional representations are averaged by a combiner to produce a unified population representation $H_t^{(p)}$.

For region r, let $E_r(X_{t,r})$ denote its embedded feature sequence. Region-wise self-attention produces summary $h_{t,r}$, and cross-region attention updates it using summaries from the remaining regions:

$$h_{t,r} = RSA_r\left(E_r(X_{t,r})\right), \tilde{h}_{t,r} = CRA_r\left(h_{t,r}, \{h_{t,j} : j \neq r\}\right).$$

The unified representation is

$$H_t = \frac{1}{R}\sum_{r=1}^{R} \tilde{h}_{t,r},$$

and a linear classification head with sigmoid activation produces eight independent concept probabilities:

$$\hat{\mathbf{y}}_t = \sigma(WH_t + b).$$

The model is optimized with binary cross-entropy. Concept co-occurrence is imbalanced across the movie, so sampling is stratified by unique label combinations; combinations appearing for fewer than 50 seconds are excluded from training. The reported architecture uses six region-wise and cross-region attention layers, six attention heads, a hidden dimension of 396, and a feed-forward dimension of 792. Models are trained for 49 epochs with Adam, an initial learning rate of 10^{-4}, and a StepLR schedule. Region embeddings use convolutional patches of 1×5 for the clusterless microwire inputs.

All available labeled movie-viewing intervals are used for final training. No movie-viewing fold is reserved as the principal test set because the intended generalization target is not another segment of the same perceptual episode; it is the subsequent recall condition. This choice concentrates the limited one-shot training signal but also changes how performance must be interpreted. The study demonstrates association with out-of-condition recall, not a conventional estimate of held-out movie classification accuracy. Hyperparameters were selected using four participants with rich pre- and postsleep narratives and then applied to all ten participants. Because those four participants remain in the final analysis, hyperparameter selection is a potential source of optimistic bias and is considered in the limitations.

Figure 4.1 summarizes the participant-specific encoder and the division between within-region and cross-region attention.

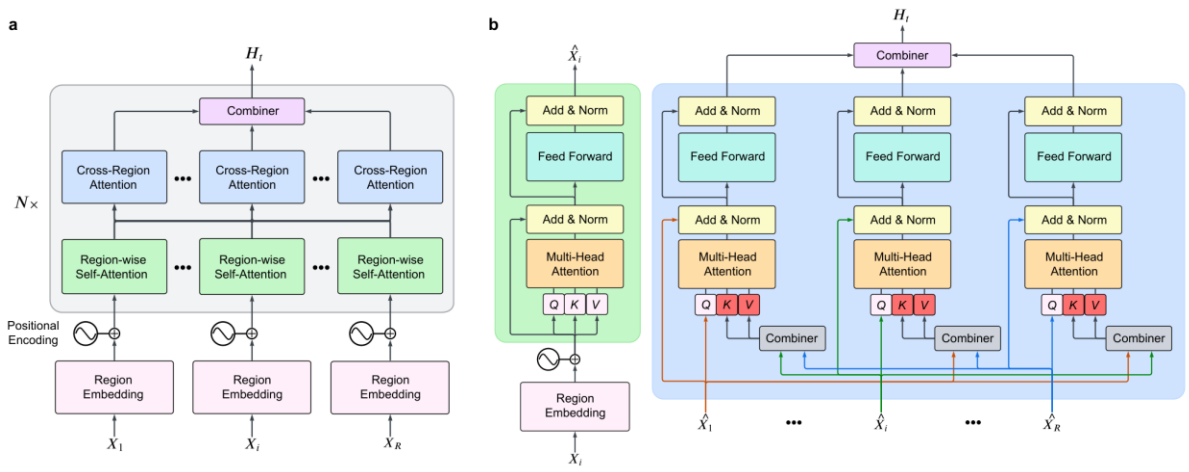

Figure 4.1. Multi-region transformer architecture for participant-specific neuronal population decoding. (a) Neural signals from each recorded region are mapped through region-specific embedding layers with positional encoding, processed by region-wise self-attention and cross-region attention, and integrated by a combiner into the population representation H_t. The attention blocks are repeated across N layers before the representation is mapped to concept probabilities by the classification head described in the text. (b) Detailed structure of the region-wise and cross-region attention mechanisms. Region-wise attention models dependencies within each regional input, whereas cross-region attention exchanges information through the regional class-token representations before combination. Reproduced from Ding et al. (2025b), Extended Data Figure 1.

4.4 Evaluating Internally Generated Recall

4.4.1 Memory Confidence Score

For concept c, the time-resolved activation $a_c(t)$ is first smoothed with a one-second moving average. Let b_c be the mean activation for that concept across the complete recall period. For each matching vocalization at time $\tau \in \mathrm{V}_c$, the analysis measures activation above this baseline during the four seconds preceding speech. The observed score is

$$A_{\text{real}}^{(c)} = \frac{1}{|\mathrm{V}_c|} \sum_{\tau \in \mathrm{V}_c} \int_{\tau-4}^{\tau} max\,(a_c(t) - b_c, 0)\, dt.$$

If the same concept is mentioned more than once within five seconds, only the first mention is retained so that primary evaluation windows do not overlap. An equal number of unrelated-word vocalizations, including other concept mentions and non-concept

words, is then sampled, and the same quantity is computed. Repeating this procedure 1,500 times produces the surrogate distribution

$$\{A_{\text{surrogate}}^{(c,k)}\}_{k=1}^{1500}.$$

The Memory Confidence Score (MCS) is the percentile rank of $A_{\text{real}}^{(c)}$ within this distribution:

$$MCS_c = 100 \cdot \frac{1}{1500} \sum_{k=1}^{1500} \mathbf{1}\,[A_{\text{surrogate}}^{(c,k)} \leq A_{\text{real}}^{(c)}].$$

An MCS above 50 indicates that activation preceding matching vocalizations tends to exceed activation preceding unrelated words. The statistic is computed only for concepts recalled at least three times by a participant. Group-level tests compare the resulting participant–concept MCS values with 50 using a Wilcoxon signed-rank test.

MCS is a relative event-associated statistic, not timepoint-level classification accuracy. It does not require the model output to cross 0.5, which is important because internally generated recall activity may contain only a component of the population pattern learned during audiovisual experience. It also does not determine the precise onset of a remembered thought. The -4 to 0-second interval captures activity preceding a verbal report, including possible retrieval, speech planning, and other correlated processes. Temporal controls and alternative models are therefore required before interpreting a high MCS as content-related reinstatement.

Figure 4.2 illustrates the continuous model output, matching and unrelated vocalization windows, surrogate construction, and percentile calculation underlying MCS.

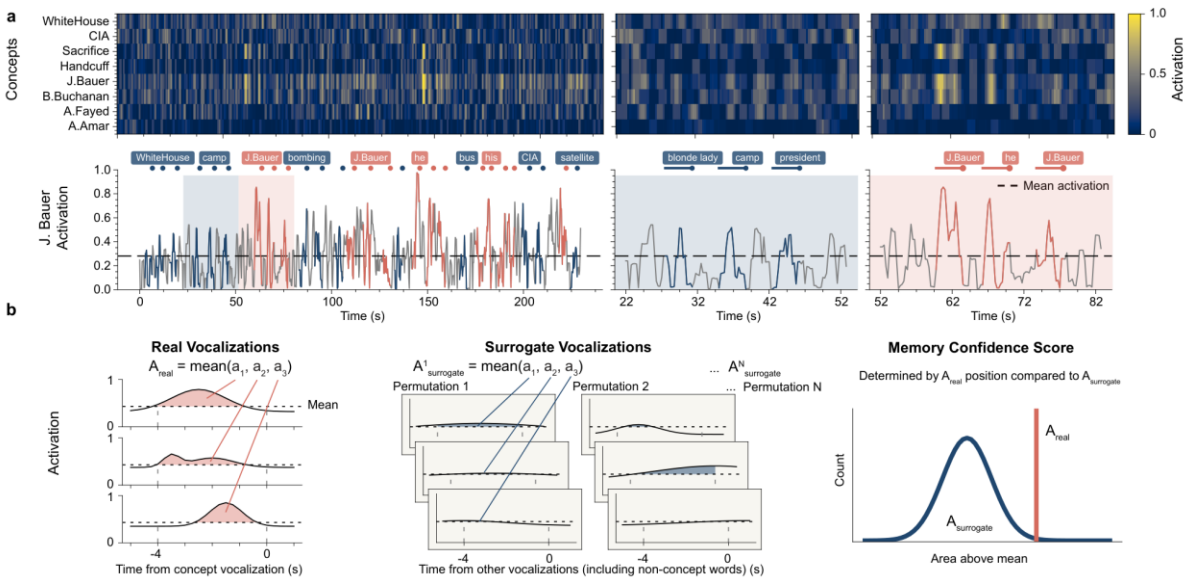

Figure 4.2. Construction of the Memory Confidence Score. (a) An example participant's continuous model output during presleep recall is shown for all concepts and for the J. Bauer output alone. Matching and unrelated vocalizations define four-second pre-vocalization windows; enlarged examples illustrate stronger activation before matching mentions than before unrelated words. (b) For each concept, activation above the recall-period mean is integrated within matching windows and averaged to obtain A_{real}. Equally sized samples of unrelated vocalizations are drawn 1,500 times to form $A_{\text{surrogate}}$, and MCS is the percentile rank of A_{real} within that distribution. Adapted from Ding et al. (2025b), Figure 2a–b.

4.4.2 Controls Against Coincident Temporal Structure

The movie contains recurring scenes and temporally correlated concepts, so a flexible model could exploit accidental alignment between neural activity and the annotation sequence. A circular-shift control preserves the internal temporal structure of both the movie labels and neural recording while breaking their correct alignment. For each of 100 shifts, all microwire channels are displaced by the same random temporal offset with wraparound; a participant-specific model is retrained on the shifted pairing, applied to the unaltered recall data, and evaluated with the same MCS procedure.

Two simpler input–model combinations test whether the result can be obtained without nonlinear population integration or high-resolution microwire activity. In the first, the attention encoder is replaced by a linear logistic-regression mapping from the

flattened clusterless input to the concept labels. In the second, the same general decoding framework is applied to bipolar macrowire iEEG represented by Morlet-wavelet power in eight frequency bands from 3 to 180 Hz. Stepwise evaluation windows before and after the primary interval additionally test whether the effect is concentrated near the subsequent vocalization rather than distributed uniformly through recall.

4.5 Can Population Activity Identify Recalled Content?

Across all ten participants, the full population decoder produces MCS values significantly above the 50th-percentile null during presleep recall ($Z = 3.26$, $p = 0.0011$, $D = 0.470$). The effect is also present during postsleep recall ($Z = 3.46$, $p = 0.0005$, $D = 0.516$). Figure 4.3 shows examples of the event-aligned activation and the participant–concept heatmaps. The heatmaps reveal heterogeneity in which concepts are recalled and decoded, but the pooled effects occur on both sides of the overnight interval. These results support the claim that a decoder trained only during the single movie viewing recovers content-related population structure during later internally generated recall.

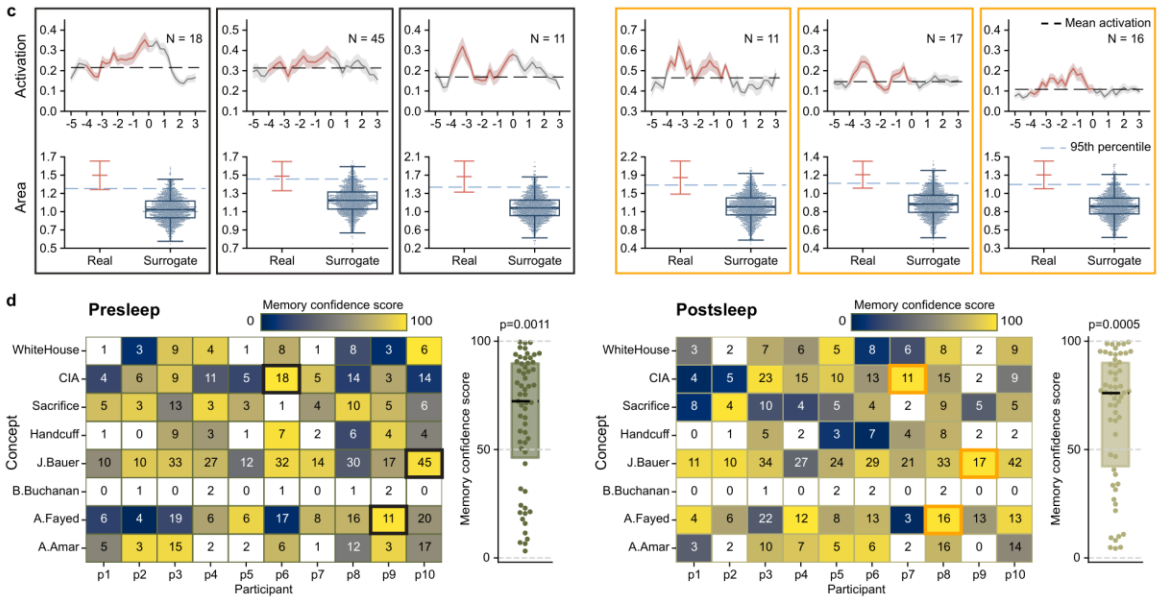

Figure 4.3. Example and population-level recall-decoding results. (c) Six concepts with high MCS illustrate mean activation aligned to vocalization onset and the position of the observed activation area relative to its surrogate distribution; three examples are from presleep recall and three from postsleep recall. (d) Participant-by-concept heatmaps report MCS for concepts mentioned at least three times, with cell entries giving the number of matching vocalizations. Pooled participant–concept scores exceed the 50th-percentile null both presleep ($p = 0.0011$) and postsleep ($p = 0.0005$). Adapted from Ding et al. (2025b), Figure 2c–d.

The temporal analysis strengthens this interpretation. Moving the evaluation interval earlier or later than the primary -4 to 0-second window reduces performance, with peak decoding occurring approximately one to two seconds before vocalization (Figure 4.6c). A signal that emerged only after speech onset would be more readily explained by auditory feedback or articulation. A pre-vocalization peak is consistent with retrieval-related activity, although it cannot completely separate mnemonic content from concept-specific speech preparation.

The simpler baselines do not show significant pooled decoding. Logistic regression yields $Z = 0.743$, $p = 0.4572$ presleep and $Z = 0.878$, $p = 0.3800$ postsleep. The macrowire model yields $Z = -0.108$, $p = 0.9137$ presleep and $Z = 0.573$, $p = 0.5667$ postsleep. In addition, the correctly aligned model exceeds every circular-shift surrogate

in the reported comparison; shifted models produce mean group statistics near the null in both conditions. Figures 4.4 and 4.5 show these complementary controls. Together, they argue against an explanation based only on linear concept tuning, coarse macroscopic spectral activity, or the recurring temporal structure of the movie.

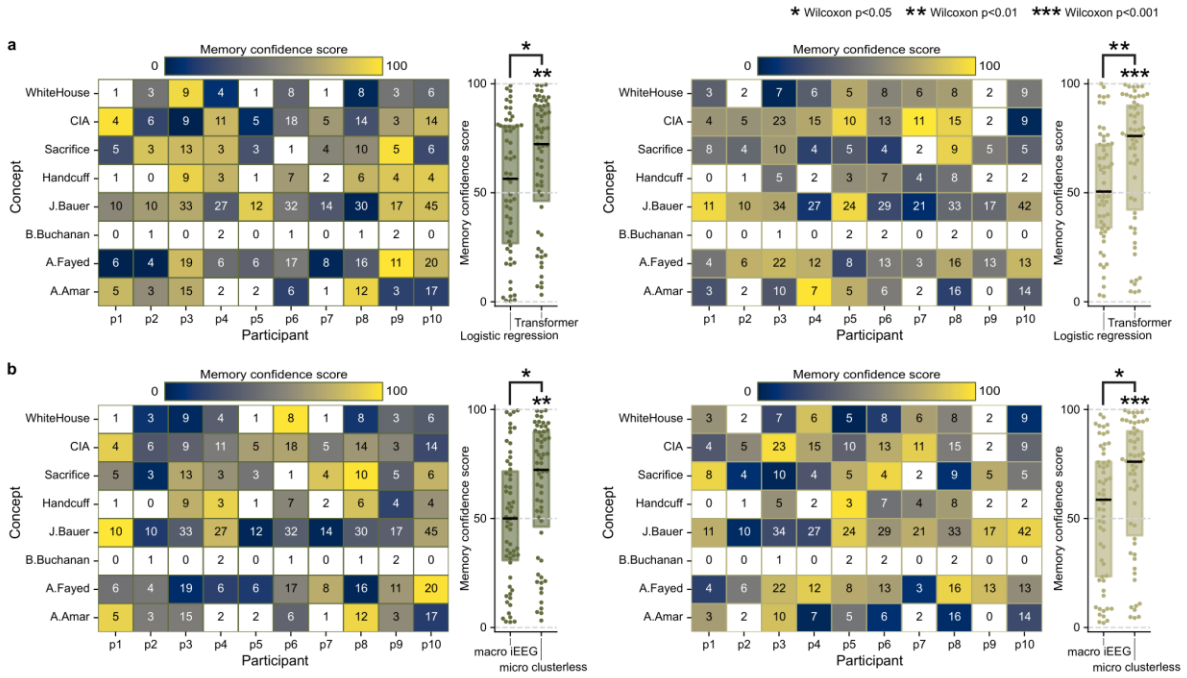

Figure 4.4. Model and input controls for recall decoding. (a) Participant–concept MCS heatmaps and pooled score distributions compare the population transformer with logistic regression during presleep recall (left) and postsleep recall (right). (b) The corresponding comparison contrasts the transformer applied to bipolar macrowire iEEG with the clusterless microwire population input. Brackets report paired Wilcoxon signed-rank comparisons, while asterisks over individual distributions denote tests against the 50th-percentile null. Reproduced from Ding et al. (2025b), Extended Data Figure 2.

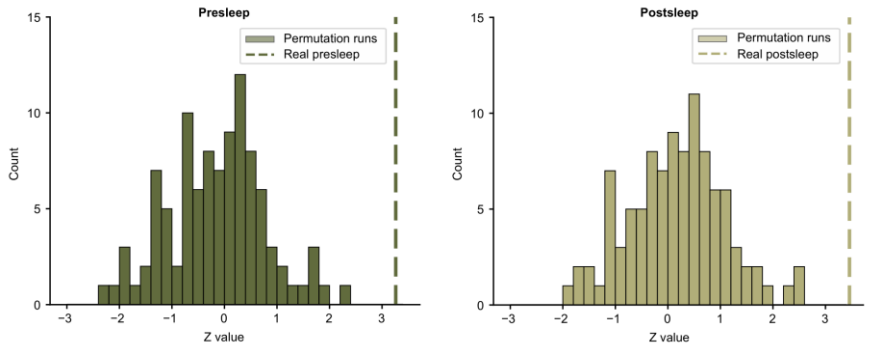

Figure 4.5. Circular-shift controls for recall decoding. Histograms show group-level Z-statistics from 100 models trained after circularly shifting movie-period neural activity relative to the concept annotations. Dashed lines mark the correctly aligned results for presleep ($Z = 3.261$) and postsleep ($Z = 3.457$); both real values exceed every shifted surrogate and therefore lie at the 100th percentile of their respective control distributions. Reproduced from Ding et al. (2025b), Extended Data Figure 3.

Table 4.1 collects the principal recall-decoding statistics and the complementary linear, macrowire, and circular-shift controls.

Table 4.1. Recall-decoding and control results. The circular-shift entries report the mean Z-statistic across 100 surrogate models rather than a single inferential test and are therefore not assigned a p-value in this table.

Model or control	Presleep Z	Presleep p	Postsleep Z	Postsleep p
Population attention decoder	**3.260**	**0.0011**	**3.460**	**0.0005**
Logistic regression	0.743	0.4572	0.878	0.3800
Macrowire spectral decoder	-0.108	0.9137	0.573	0.5667
Circular-shift models, mean statistic	-0.138	—	0.174	—

The evidence supports concept-specific decoding at the level defined by MCS: activation learned during movie viewing preferentially precedes later matching verbal reports. It does not establish continuous identification of mental content throughout recall, and the comparison of a significant nonlinear model with nonsignificant baselines does not by itself test the statistical significance of the difference between model classes. The

control results should therefore be interpreted as convergent evidence rather than proof of a unique computational mechanism.

4.6 Which Regions Support Decoding Before and After Sleep?

The full model integrates all clinically available recording sites, but successful decoding does not show whether the same regional population information supports presleep and postsleep recall. Electrode locations were obtained by coregistering postoperative CT with preoperative structural MRI, followed by automated cortical parcellation, dedicated medial-temporal segmentation, expert labeling, and visual quality control (Davis et al., 2021; Fischl et al., 2002; Yushkevich et al., 2015). MTL comprises hippocampus (HPC), amygdala (AMY), parahippocampal cortex (PHC), and entorhinal cortex (ENT); FC comprises prefrontal cortex (PFC) and anterior cingulate cortex (ACC). Table 4.2 reports the participant-level coverage underlying the regional analyses.

Table 4.2. Microwire counts by participant and anatomical region. MTL is the sum of HPC, AMY, PHC, and ENT; FC combines PFC and ACC. Full-model counts can additionally include channels in other recorded regions. Adapted from Ding et al. (2025b), Table 1.

Participant	Full	MTL	HPC	AMY	PHC	ENT	PFC	ACC
P1	48	24	8	0	16	0	8	8
P2	64	32	16	0	8	8	16	8
P3	72	24	16	8	0	0	24	8
P4	40	40	16	8	8	8	0	0
P5	72	32	16	16	0	0	24	16
P6	80	48	16	16	8	8	8	8
P7	72	48	16	16	0	16	8	16
P8	40	16	8	0	8	0	16	0
P9	96	24	0	8	0	16	8	16
P10	96	40	24	8	0	8	24	0

Region-specific analyses address the scientific question in two ways. Exclusion models remove MTL, HPC, or FC channels from the full input. Restriction models use only MTL or only FC channels. Because these operations also change channel number and the anatomical composition of the remaining input, they test the information available to a retrained decoder rather than the causal necessity of a region. Table 4.3 summarizes the resulting condition-specific tests.

Removing MTL channels abolishes significant presleep decoding but leaves postsleep decoding significant. Removing hippocampal channels yields no significant decoding in either condition. Removing FC channels leaves presleep decoding significant but not postsleep decoding. The restricted models show a complementary pattern: Only MTL decodes presleep but not postsleep recall, while Only FC decodes postsleep but not presleep recall.

Table 4.3. Region-exclusion and region-restriction results. Bold entries are significantly above the MCS null within the corresponding condition. The participant counts differ because an exclusion model requires coverage in the excluded region and at least one remaining channel, whereas a restriction model requires coverage in the retained region.

Model	Participants	Presleep Z	Presleep p	Postsleep Z	Postsleep p
No MTL	9	-0.961	0.3368	**3.233**	**0.0012**
No hippocampus	9	1.249	0.2116	0.190	0.8496
No FC	9	**2.324**	**0.0201**	1.248	0.2121
Only MTL	10	**2.359**	**0.0183**	0.337	0.7363
Only FC	9	-0.518	0.6045	**3.033**	**0.0024**

The Only FC pre/post comparison is nominally significant ($Z = 2.060$, $p = 0.0391$, paired $D = 0.332$), whereas the corresponding Only MTL pre/post comparison is not ($Z = 1.120$, $p = 0.2695$, paired $D = 0.181$). These reported regional tests were not adjusted for the multiple exclusion, restriction, condition-versus-null, and pre–post comparisons and

should therefore be treated as exploratory. The results are consistent with changing regional contributions after sleep, but the phrase “double dissociation” should be used cautiously because significance in one condition and nonsignificance in another does not by itself establish a significant interaction.

The hippocampal exclusion result further complicates a simple MTL-to-FC replacement account. Although FC-only activity supports postsleep decoding, removing hippocampal channels from the broader population eliminates significant postsleep decoding. The available evidence is therefore more consistent with a redistribution of decodable information and changing interactions among recorded regions than with complete hippocampal independence after one night. The observational design cannot determine whether sleep caused this pattern, and alternative explanations include participant heterogeneity, electrode coverage, recording stability, and differences in the content produced during the two recall sessions.

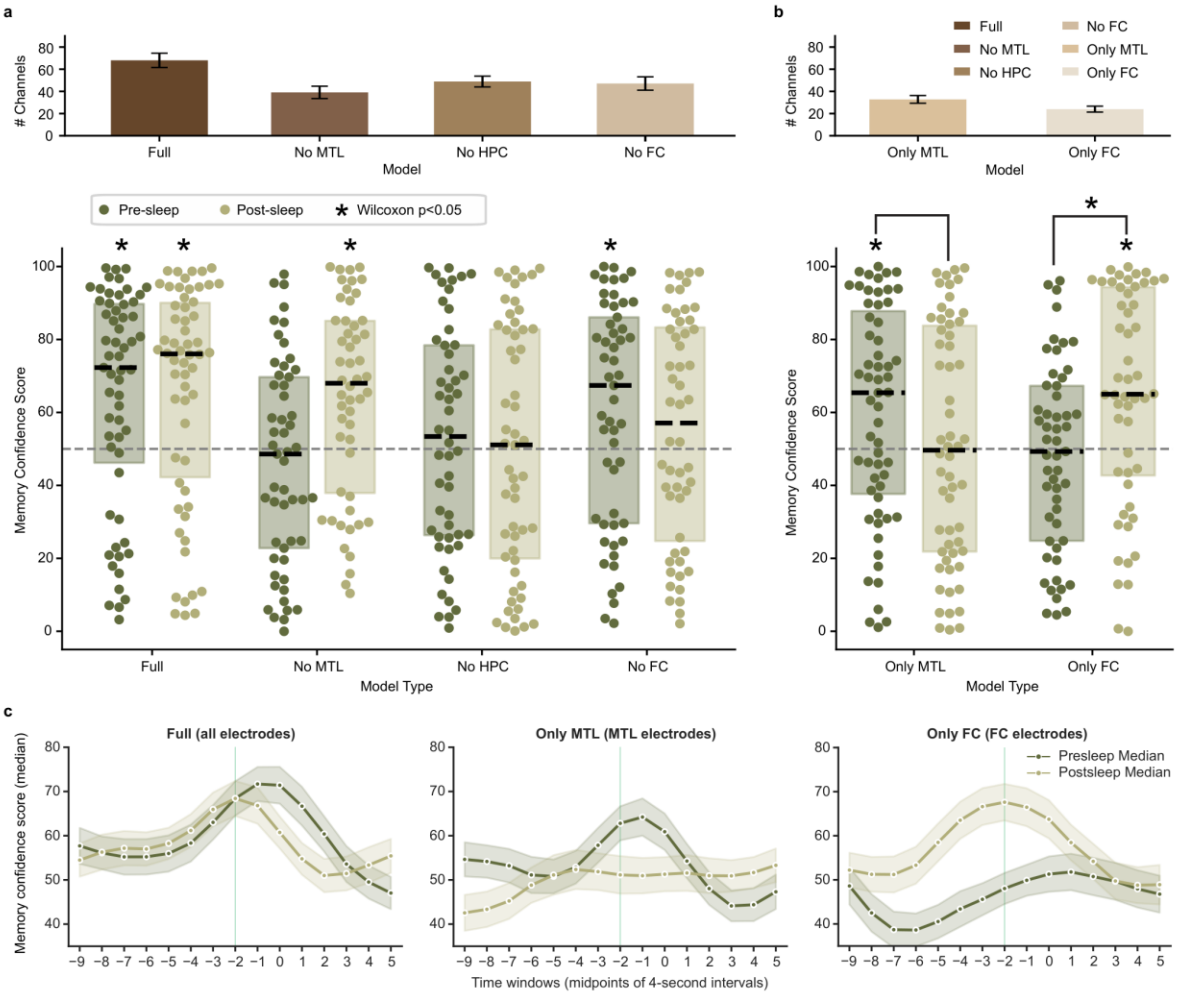

Figure 4.6. Regional contributions and temporal profiles of recall decoding before and after sleep. (a) Channel counts and MCS distributions are shown for the Full model and models retrained after excluding MTL, hippocampal (HPC), or FC channels. (b) The corresponding analyses use models restricted to MTL or FC channels. Dots denote participant–concept observations, dashed black lines denote medians, boxes denote interquartile ranges, and asterisks indicate Wilcoxon signed-rank tests against the 50th-percentile null; brackets mark paired presleep–postsleep comparisons. (c) Median MCS across stepwise four-second windows for the Full, Only MTL, and Only FC models peaks before vocalization, showing that the regional decoding patterns are temporally concentrated in the putative retrieval interval rather than after speech onset. Shaded bands denote variability as reported in the source figure. Reproduced from Ding et al. (2025b), Figure 3.

4.7 Does Decoding Depend on Identified Selective Neurons?

The population decoder is trained from clinically determined microwire recordings without first selecting neurons that respond to the target concepts. Nevertheless, some recorded units may be character selective, raising the possibility that a small number of

conventionally tuned cells dominate decoding (Quian Quiroga et al., 2005; Gelbard-Sagiv et al., 2008). To test this account, 626 manually curated single and multi-units with movie-period firing rates of at least 0.5 spikes per second were examined using frame-level annotations of ten characters in the episode. A unit was considered selective when it fired in more than 30% of the eligible post-onset windows, its observed pre-to-post firing-rate increase exceeded 99% of 1,000 sign-permutation surrogates, and it responded significantly to only one tested character. Thirty-seven units, approximately 6%, met these criteria.

If these units were the principal source of model information, concepts associated with more selective units would be expected to have higher MCS. The observed correlations instead trend negative and are not statistically significant: presleep $r = -0.182$, $p = 0.24$, and postsleep $r = -0.289$, $p = 0.06$. Relaxing the definition to include units responsive to more than one character yields 45 units and 56 character responses, but still produces no positive relationship with MCS: presleep $r = -0.167$, $p = 0.28$, and postsleep $r = -0.313$, $p = 0.04$. MCS also does not differ significantly between concepts with and without an identified selective unit. Presleep medians are 65.10 versus 75.60 ($Z = -1.34$, $p = 0.18$); postsleep medians are 70.50 versus 82.40 ($Z = -0.57$, $p = 0.57$).

These results argue against the narrow explanation that decoding success primarily reflects the number of identified character-selective units available for each concept. They do not show that selective neurons make no contribution. Selectivity estimates are limited by the stimuli, firing-rate threshold, recorded sample, and number of observations, and a tuned neuron may participate in a broader nonlinear population pattern without producing a simple positive relationship with MCS. The appropriate

conclusion is that the decoder accesses information beyond what can be explained by the measured availability of individually identified selective units.

Figure 4.7 shows the selective-unit criterion and the observed relationship between the number of identified units and participant–concept MCS.

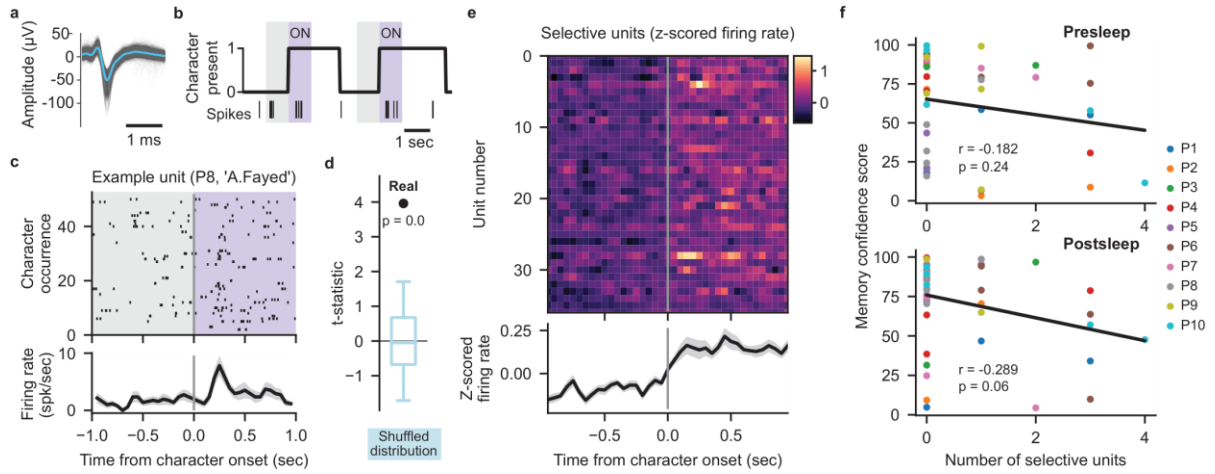

Figure 4.7. Relationship between identified character-selective units and population-decoding performance. (a) Waveforms from an example manually sorted single unit. (b) Frame-level character-presence labels and spikes illustrate the pre- and post-onset analysis windows. (c) Raster and peri-stimulus time histogram (PSTH) show the example unit's response across 51 appearances of the character. (d) The observed firing-rate t-statistic is compared with a sign-permutation surrogate distribution to determine selectivity. (e) PSTHs summarize the 37 units selective for a single character at $p < 0.01$. (f) Across participant–concept observations, the number of identified selective units is not positively correlated with MCS during presleep ($r = -0.182$, $p = 0.24$) or postsleep ($r = -0.289$, $p = 0.06$) recall. Reproduced from Ding et al. (2025b), Figure 4.

4.8 Scientific Interpretation

The central result is that relatively small, clinically determined neuronal populations contain information that links a single externally experienced episode to later internally generated recall. Models trained separately for each participant on one movie viewing produce concept-related model activations before matching verbal reports without being trained on recall labels or restricted to preidentified responsive neurons. The nonlinear attention decoder, temporal controls, regional analyses, and selective-unit analysis

support a population-level account whose performance is not explained by the measured number of individually identified tuned cells. This finding is consistent with neural reinstatement, a broader account in which retrieval reactivates aspects of neural activity engaged during prior experience (Danker and Anderson, 2010; Gelbard-Sagiv et al., 2008; Vaz et al., 2020). The word *reinstatement* should nevertheless be interpreted at the resolution of the experiment. The model detects a statistical correspondence between movie-trained signals and activity preceding recall. It does not establish that the complete encoding state is reproduced, that the same neurons carry identical information in both conditions, or that the decoded signal is sufficient for the subjective memory.

The presleep and postsleep regional results add a second conclusion. Specific content remains decodable after sleep, but the information available from MTL- and FC-restricted populations differs across the two recall sessions. This pattern is compatible with theories in which memory representations are reorganized across hippocampal and neocortical systems during consolidation (Wang and Morris, 2010; Squire et al., 2015; Brodt et al., 2023) and with evidence for a shift toward neocortically centered retrieval after consolidation (Takashima et al., 2009). It does not establish a causal effect of sleep, because there is no wake-control interval matched for elapsed time and repeated testing, and the regional analyses are influenced by clinical electrode placement and unequal channel counts.

4.9 Remaining Limitations and the Need for Cross-State Representation Transfer

The study contains ten participants with drug-resistant epilepsy and clinically determined electrode placement. Participant-specific models accommodate individual coverage but do not establish transfer to a new person, recording system, or institution.

The target vocabulary contains eight concepts selected for sufficient movie exposure and recall frequency, and the single principal episode constrains the range of experiences tested. Verbal recall samples only content that participants choose and are able to express; vocalization onset is delayed relative to internal retrieval and is correlated with speech preparation. MCS is a within-recall percentile statistic rather than continuous classification accuracy, and the requirement of at least three vocalizations selects concepts with relatively rich behavioral evidence. Several analytical choices also constrain inference. All movie data are used for training, so conventional held-out movie performance is not estimated. Hyperparameters were selected using four participants who are included in the final analysis. Threshold and normalization parameters use unlabeled statistics calculated across task epochs, so recall evaluation is independent of recall labels but is not fully inductive with respect to the target recording distribution. The study lacks an independent replication cohort, and multiple participant–concept observations may share participant-specific dependencies. Region-exclusion and restriction models alter both anatomy and input dimensionality, limiting causal localization. Finally, the comparison with selective neurons depends on the units and characters that could be identified in the available recordings and should not be interpreted as evidence against a role for concept-selective neurons in memory.

Most importantly, successful recall decoding does not show that the learned representation provides a stable concept space across experimental conditions. Each concept is represented by a classifier output learned during movie viewing. The model is applied directly to recall despite differences in sensory input, behavior, and neural distribution, and the regional results indicate that the population information supporting

decoding is state dependent. During sleep, continuous behavioral labels disappear entirely. Chapter 5 therefore asks: **how can concept-aligned neural representations be learned and transferred across movie viewing, recall, and sleep when target-state supervision becomes sparse or absent?**

Chapter 5 Transferring Concept-Aligned Neuronal Representations Across Brain States

5.1 From Recall Decoding to Representation Transfer

Chapter 4 showed that a participant-specific decoder trained during one movie viewing produced concept-related activations before later verbal recall. That result established that neuronal information associated with an experience remained accessible outside the original perceptual condition. It did not establish that the learned representation preserved a reusable concept structure across brain states. Each concept was represented by a separate classifier output, and evaluation depended on recall vocalizations. During sleep, the neural distribution changes further and time-resolved behavioral labels disappear.

Prior work establishes that neural activity can reinstate aspects of recent experience during offline states, including replay of rodent hippocampal population patterns and place-cell sequences (Wilson and McNaughton, 1994; Foster and Wilson, 2006). In humans, internally generated semantic content has been reconstructed from fMRI, but such approaches can require many hours of participant-specific training data (Tang et al., 2023). Concept-selective neurons provide a more direct content-specific reference during recall and sleep, yet they are sparse and cannot densely label the people, places, and abstract concepts contained in a one-shot naturalistic experience

(Quian Quiroga et al., 2005; Gelbard-Sagiv et al., 2008). These precedents leave unresolved the combined problem addressed here: transferring a concept-organized population representation from a densely labeled experience to unlabeled sleep and evaluating the transferred signal without treating either model output or global domain alignment as ground truth.

This chapter addresses two coupled problems. The first is **representation transfer**: how can concept structure learned from densely labeled movie viewing remain informative during internally generated recall and NREM sleep? The second is **target-state evaluation**: how can that transfer be tested when the sleep state provides no continuous behavioral report of content? A model can reduce a global difference between movie and sleep activity without retaining any concept-specific information. Conversely, a concept-related signal cannot be interpreted during sleep unless an observation independent of the model provides evidence about its content. The central scientific question is therefore: **how can concept-aligned neuronal representations learned from densely labeled experience be transferred to internally generated states with sparse or absent labels, and how can successful transfer be evaluated using independent neural evidence?** The study *Concept-Aligned Neuronal Representation Transfer Across Brain States* addresses this question with three connected components (Ding et al., 2026): a dual-head representation that learns both direct concept logits and a concept-organized embedding space; Recall-Adapted Sleep Alignment (RASA), which combines unlabeled sleep adaptation with a recall-based semantic safeguard; and Bidirectional Event-Conditioned Alignment (BECA), which evaluates model signals against held-out concept-selective neurons.

The form of transfer studied here is specific. Models are trained separately for each participant and transferred across that participant's movie-viewing, recall, and sleep recordings. The study does not demonstrate transfer across participants, institutions, electrode systems, or arbitrary semantic vocabularies. Likewise, “sleep decoding” refers to content-specific alignment between model-derived signals and independent neuronal references for a limited set of concepts. It is not continuous reconstruction of every thought, dream, or memory expressed during sleep.

The Chapter 4 and Chapter 5 studies share a Movie–Recall–Sleep logic but should not be treated as the same cohort or as a single model with an added loss. Table 5.1 compares their cohorts, inputs, selective-unit analyses, scientific questions, and encoder terminology.

Table 5.1. Relationship between the Chapter 4 recall-decoding study and the expanded Chapter 5 cross-state transfer study. The studies share a design logic but use different cohorts, model objectives, and evaluation targets.

Study role	Chapter 4 / Movie–Recall–Sleep study	Chapter 5 / cross-state transfer study
Participants	10	13
Naturalistic exposure	One 42-minute episode	Four movies; 10.9 hours total
Primary population input	Clusterless voltage-spike activity	Clusterless voltage-spike activity
Curated-unit analysis	626 single and multi-units	452 curated single units
Selective-unit evidence	37 character-selective units used to test whether recall decoding depends on identified tuning	62 concept-selective units used as held-out references for sleep evaluation
Main scientific question	Decoding specific content during pre- and postsleep recall	Transferring and evaluating concept structure during NREM sleep
Encoder terminology	Region-wise and cross-region attention (RSA/CRA)	Channel-specific and cross-channel attention (CSA/CCA)

5.2 Formalizing Cross-State Supervision

Let D_{movie}, D_{recall}, and D_{sleep} denote participant-specific recordings from movie viewing, free recall, and NREM sleep. Movie viewing supplies dense multi-label supervision:

$$D_{movie} = \{(x_i, \mathbf{y}_i)\}, \mathbf{y}_i \in \{0,1\}^K.$$

Recall refers to the same experience and concept vocabulary, but its verbal reports are sparse and temporally delayed. The adaptation method therefore treats recall inputs as unlabeled:

$$D_{recall} = \{x_i^{(r)}\}.$$

NREM sleep contains neither stimulus-aligned nor behavioral concept labels:

$$D_{sleep} = \{x_i^{(s)}\}.$$

The label asymmetry is as important as the distribution shift. Movie labels specify when concepts appear externally. Recall provides a behaviorally meaningful intermediate condition without dense interval-level supervision. Sleep supplies long recordings from a state of theoretical interest but no continuous report of content. The learning objective must use these domains differently rather than pretending that all three provide equivalent examples.

A successful representation should satisfy three properties. Within the movie domain, its structure should discriminate the annotated concepts. Under adaptation, it should reduce state-related distribution differences without erasing the concept relationships learned from the movie. During held-out sleep, its concept-specific signals should align temporally with an independent neural reference for the same concept. None

of these properties alone is sufficient. Source discrimination can fail under state change; distribution alignment can discard semantics; and event correlation can arise from circular use of the same channels unless evaluation independence is enforced.

The expanded cohort described in Chapter 3 contains 13 participants, four movies, 10.9 hours of movie viewing, 2.8 hours of recall, and 58.4 hours of identified NREM sleep. NREM identification and cohort characteristics are defined in Section 3.4. The model input follows the clusterless preprocessing developed for the Movie–Recall–Sleep paradigm in Section 3.3 and used in Chapter 4: positive- and negative-polarity threshold-crossing amplitudes are preserved separately and summarized in 250-ms windows. Each evaluated concept-selective unit is treated separately, and its recording channel is removed from the model input before training, adaptation, and sleep evaluation. The model must therefore infer a concept-related signal from the remaining population.

5.3 Learning a Concept-Aligned Representation

5.3.1 Hierarchical Population Encoder

The encoder preserves the anatomical bundle structure of the microwire recordings. For each participant, W input wires are partitioned into eight-wire bundles. Channel-specific attention (CSA) uses convolutional patch embedding and self-attention to summarize temporal structure within each bundle. Cross-channel attention (CCA) then integrates the bundle-level tokens into a global representation $z(x) \in \mathrm{R}^{384}$. This organization is related to the population encoder in Chapter 4 but is independently specified for the expanded cohort and dual-head objective.

Each input has shape $x \in \mathrm{R}^{2\times W\times T}$, where $T = 50$ temporal samples summarize a 250-ms interval and W varies from 40 to 96 across participants. The reported architecture

uses five CSA layers, five CCA layers, six attention heads, a hidden dimension of 384, and an embedding dimension of 128.

Two heads are applied to the shared encoder. The classifier head produces direct multi-label logits:

$$\boldsymbol{l}(x)=W_{\text{cls}}z(x) \in \mathrm{R}^{K}.$$

The embedding head produces a normalized neural representation:

$$\tilde{z}(x) = \frac{W_{\text{emb}}z(x)}{\|W_{\text{emb}}z(x)\|_2}.$$

The model jointly learns concept anchors $\{c_1, \ldots, c_K\}$ and a null anchor c_{null} for intervals with no annotated concept. Similarity between neural activity and concept k is

$$s_k(x) = \frac{cos\ (\tilde{z}(x), c_k)}{\tau},$$

where $\tau = 0.07$ is the contrastive temperature. The classifier head retains an interpretable direct prediction for each concept; the embedding head exposes relative concept geometry and temporal trajectories that cannot be recovered from independent scalar logits alone.

5.3.2 Source-Domain Concept Alignment

The classifier is trained with multi-label binary cross-entropy,

$$\mathrm{L}_{\text{BCE}} = -\sum_{k=1}^{K}[y_k log\ \sigma(\ell_k) + (1 - y_k)log(1 - \sigma(\ell_k))].$$

For the embedding objective, let $P(x)$ be the active concepts for sample x, using the null anchor when none is active, and let $N(x)$ be the inactive concept set. Extending supervised contrastive learning to the multi-label setting (Khosla et al., 2020), the loss is

$$L_{ctr}(x) = -\sum_{k \in P(x)} log \frac{exp(s_k(x))}{\sum_{j \in N(x) \cup \{k\}} exp\, s_j(x))}.$$

Each active concept is required to outrank the inactive concepts separately. This avoids collapsing a time bin containing multiple concepts into one composite embedding. The complete movie-source objective is

$$L_{src} = \lambda_{BCE} L_{BCE} + \lambda_{ctr} L_{ctr},$$

with $\lambda_{BCE} = 1.0$ and $\lambda_{ctr} = 0.5$.

Concept-balanced sampling mitigates the strongly imbalanced screen time of the movie annotations. Positive samples are drawn from each concept pool, while remaining samples come from intervals without an active annotation. For sessions dominated by one concept, its label is downweighted in co-occurring intervals to reduce the risk of a constant-positive shortcut. Concept and null embeddings are Xavier-initialized and learned jointly with the encoder.

The dual-head formulation does not assume that embedding geometry is inherently more meaningful than classifier output. It creates two complementary readouts that can be tested independently. Direct logits ask whether the model expresses concept evidence in the familiar decoding space. Concept similarity asks whether population activity occupies a reusable geometry organized around learned semantic anchors.

Figure 5.1 summarizes the shared encoder, direct classification head, and concept-embedding head used to produce these complementary readouts.

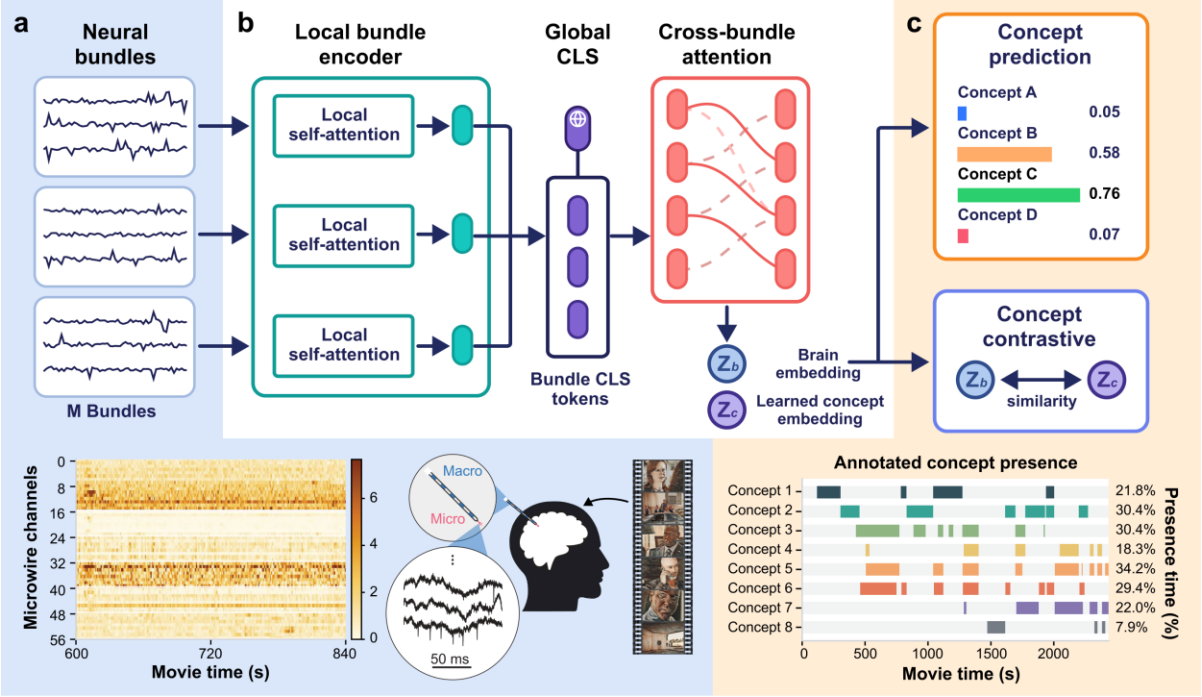

Figure 5.1. Dual-head concept-aligned neuronal population model. (a) Movie-viewing inputs pair clusterless voltage-spike amplitude activity from anatomically grouped microwire bundles with time-resolved multi-label concept annotations. The example input shows positive-polarity activity, while the model retains separate positive- and negative-polarity channels. (b) Local self-attention summarizes temporal structure within each bundle, bundle-level class tokens are integrated through a global class token and cross-bundle attention, and the encoder produces a shared brain embedding Z_b. (c) Two complementary heads map this representation either to direct multi-label concept predictions or to similarity with learned concept embeddings Z_c. The two outputs support classifier-activation and embedding-similarity readouts during subsequent cross-state evaluation. Reproduced from Ding et al. (2026), Figure 1.

5.4 Recall-Adapted Sleep Alignment

5.4.1 Why Distribution Alignment Is Not Enough

A movie-trained representation separates the source concepts, but movie, recall, and sleep embeddings occupy different regions of the learned space. Some difference is expected and biologically meaningful: sensory drive, behavior, and sleep physiology should not be made artificially identical. The objective is narrower: to reduce state-related displacement that prevents concept evidence from being expressed while protecting the source concept geometry. RASA retains L_{src} throughout adaptation and adds two embedding-level constraints. The first uses unlabeled NREM sleep to reduce the movie-

to-sleep distribution gap. The second uses recall as a semantic safeguard. Recall is useful not because it supplies dense target labels, but because it is an internally generated state tied behaviorally to the same experience.

5.4.2 Unlabeled Sleep Alignment

For each participant, the first contiguous 30-minute NREM interval is used for adaptation and excluded from all subsequent BECA evaluation. Let Z_M and Z_S denote minibatches of movie and sleep embeddings, with means μ_M, μ_S and covariance matrices Σ_M, Σ_S. Following correlation alignment for domain adaptation (Sun and Saenko, 2016), sleep CORAL matches their first two moments:

$$\mathsf{L}_{\text{sCoral}} = \|\mu_M - \mu_S\|_2^2 + \frac{1}{4d^2}\|\Sigma_M - \Sigma_S\|_F^2.$$

This loss uses no sleep concept labels. It can bring the overall distributions closer, but it cannot determine which part of a sleep embedding should remain associated with a particular concept. Global alignment may therefore improve domain similarity while distorting the semantic relationships established during movie training.

5.4.3 Recall-Guided Semantic Safeguard

A separately trained movie-only model at epoch 50 is frozen as a teacher. For recall input x_r, the adapted student is encouraged to preserve the teacher's recall embedding:

$$\mathsf{L}_{\text{distillRecall}} = \mathsf{E}_{x_r \sim \mathsf{D}_{\text{recall}}}[1 - cos\,(\tilde{z}_{\text{student}}(x_r), \tilde{z}_{\text{teacher}}(x_r))].$$

This objective does not claim that the teacher provides ground-truth recall content. It preserves the movie-trained semantic organization expressed on recall activity while

sleep alignment changes the student representation. Recall thus acts as a consistency constraint between a labeled perceptual state and an unlabeled sleep state.

The complete RASA objective is

$$L_{RASA} = L_{src} + \lambda_{sCoral} L_{sCoral} + \lambda_{distill} L_{distillRecall}.$$

Models receive 30 source-only epochs before adaptation is introduced. Sleep CORAL is ramped over 20 epochs to $\lambda_{sCoral} = 1.0$; recall distillation uses $\lambda_{distill} = 0.35$; and the final checkpoint is taken at epoch 70. Training uses Adam with learning rate 10^{-4}, weight decay 10^{-4}, batch size 128, and 2,048 samples per epoch.

Figure 5.2 provides a representative visualization of the domain geometry before adaptation, under each individual constraint, and under full RASA. The projection is descriptive; BECA supplies the content-specific evaluation.

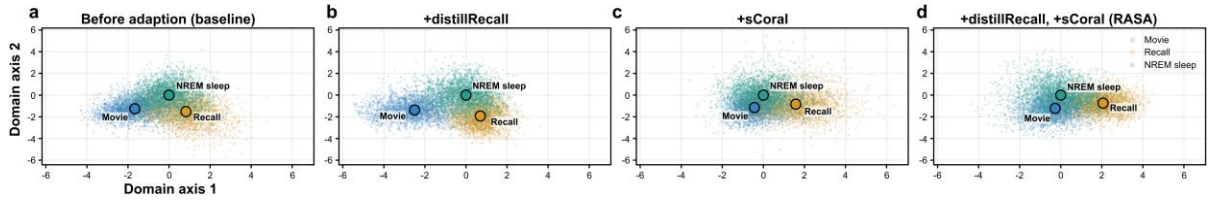

Figure 5.2. Representative embedding-domain shifts under RASA adaptation. Sleep-centered, domain-discriminant projections show the relative positions of movie, recall, and NREM-sleep embeddings in a common two-dimensional domain space. (a) Before adaptation, the movie-trained representation separates the three recording domains. (b) Recall distillation alone leaves the global domain geometry similar to baseline, consistent with its role as a semantic-consistency constraint rather than a distribution-alignment objective. (c) Sleep CORAL reduces the movie–sleep domain gap but does not explicitly protect recall-domain semantic geometry. (d) Full RASA combines sleep alignment with the recall safeguard, bringing movie and sleep embeddings into a more shared region while limiting recall displacement relative to sleep CORAL alone. The projections illustrate representative domain geometry; they do not by themselves establish preservation of concept-specific information, which is evaluated through BECA. Reproduced from Ding et al. (2026), Figure 3.

5.5 Evaluating Transfer Without Behavioral Labels

5.5.1 Independent Neuronal References

Chapter 3 introduced the 62 concept-selective units identified from movie-viewing responses. Human medial temporal neurons can respond selectively to particular people, places, and objects across sensory presentations and can reactivate during memory retrieval and sleep (Quian Quiroga et al., 2005; Gelbard-Sagiv et al., 2008; Niediek et al., 2026). Their role here is not to train the population model. For every evaluated unit, the channel containing that unit is excluded from the input throughout source training, RASA adaptation, and sleep evaluation. The 30-minute NREM segment used for CORAL is also excluded from the subsequent evaluation interval. These separations ensure that a model cannot trivially reproduce the reference neuron's signal or be evaluated on the same sleep segment used for adaptation. The held-out neuron remains an indirect reference rather than absolute sleep ground truth. Selectivity was established during movie viewing, and elevated firing during sleep is consistent with content-specific reactivation but does not prove the participant consciously experiences or recalls the concept. The evaluation question is therefore whether the population-derived signal exhibits concept-specific temporal coupling with an independent neuron known to be selective for that concept under the study's criterion.

5.5.2 Bidirectional Event-Conditioned Alignment

For a held-out neuron selective for concept k, let $a_k(t)$ denote its firing activity during NREM sleep. The model supplies two signals for the same concept: classifier activation $m_k^{\text{cls}}(t) = \ell_k(t)$ and embedding similarity $m_k^{\text{emb}}(t) = s_k(t)$. Bidirectional Event-Conditioned Alignment evaluates coupling in four directions:

Table 5.2. Four complementary BECA readouts for evaluating bidirectional coupling between population-derived concept signals and held-out concept-selective neuronal activity.

BECA readout	Conditioning events	Evaluated signal
Firing-conditioned activation (FCA)	High held-out-unit firing	Classifier activation
Activation-conditioned firing (ACF)	High classifier activation	Held-out-unit firing
Firing-conditioned similarity (FCS)	High held-out-unit firing	Concept-embedding similarity
Similarity-conditioned firing (SCF)	High embedding similarity	Held-out-unit firing

Classifier activations are smoothed with a one-second moving average before event detection and window extraction. Unit-conditioned events are the top 10% of nonzero firing bins during NREM. Model-conditioned events are the top 10% of the corresponding model signal. For event time t_0, the center window is $[t_0 - 1, t_0 + 1]$ seconds, and the flanks are $[t_0 - 3, t_0 - 1]$ and $[t_0 + 1, t_0 + 3]$ seconds. For evaluated signal $q(t)$, define the event-wise center–flank contrast

$$\Delta q(t_0) = \overline{q}_{[-1,1]}(t_0) - \frac{1}{2}(\overline{q}_{[-3,-1]}(t_0) + \overline{q}_{[1,3]}(t_0)).$$

BECA summarizes two properties. **Specificity** is the percentile of the real center-window response within 1,000 circular-shift surrogates. Each surrogate shifts the conditioning event times by at least one minute within NREM and prevents shifted events from coinciding with the originals. Values above 95 indicate stronger event locking than 95% of the surrogate alignments. **Temporal alignment** is computed as a paired t-statistic across events, with each event contributing one center value and one value obtained by averaging its two flanks. The study reports this statistic under the notation z_{peak}; despite that label, it is a paired center-versus-flank t-statistic rather than a conventional z-score. Larger values indicate a more localized peak around the conditioning event.

Bidirectionality reduces, but does not eliminate, alternative explanations. Arousal fluctuations or sleep-stage dynamics might make model activity rise near neuronal firing in one direction. Stronger evidence is obtained when neuron-defined events predict a concept-specific model signal and model-defined events predict firing of the independent neuron. BECA remains an event-coupling evaluation, not dense decoding accuracy or direct observation of subjective sleep content.

Figure 5.3 summarizes the event windows, surrogate construction, temporal-alignment statistic, and four bidirectional BECA readouts.

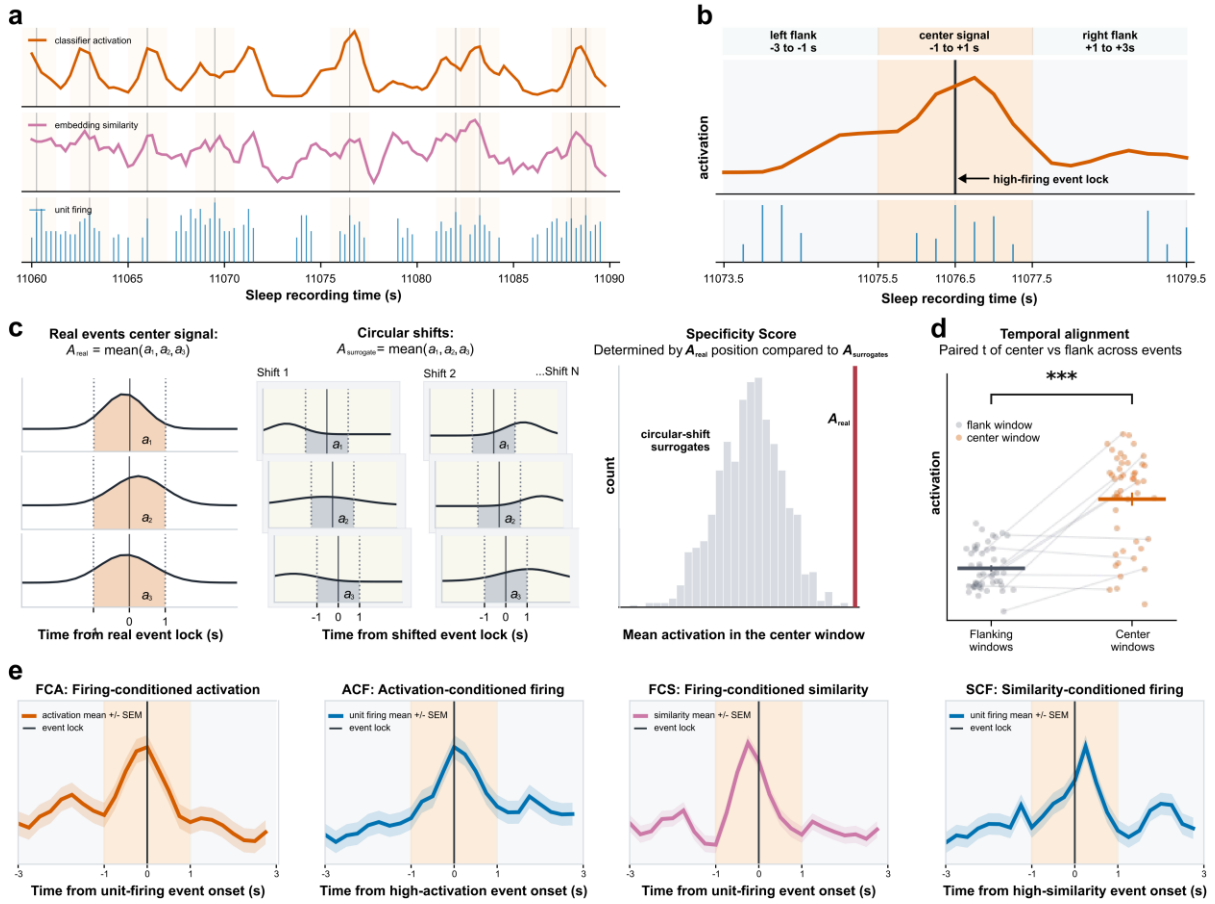

Figure 5.3. Bidirectional Event-Conditioned Alignment (BECA) for evaluating concept-related signals during sleep. (a) An example NREM segment shows classifier activation and embedding similarity for one concept together with firing of a held-out neuron selective for that concept; shaded intervals mark high-firing events. (b) Each conditioning event defines a center window from -1 to $+1$ seconds and flanking windows from -3 to -1 and $+1$ to $+3$ seconds. (c) Specificity is the percentile rank of the real event-locked center signal relative to circular-shift surrogates. (d) Temporal alignment compares the center and averaged flanking responses across events to test whether the evaluated signal forms a localized event-centered peak. (e) The same statistics are applied in four complementary directions: firing-conditioned activation (FCA), activation-conditioned firing (ACF), firing-conditioned similarity (FCS), and similarity-conditioned firing (SCF). The selective neuron's recording channel is excluded from model training, adaptation, and evaluation, preserving it as an independent reference. Reproduced from Ding et al. (2026), Figure 2.

5.6 Does Concept Alignment Transfer to NREM Sleep?

5.6.1 Comparison with Neural-Decoding Baselines

The proposed encoder is compared with output-matched baselines on the common set of 62 held-out concept-selective units. EEG Conformer (Song et al., 2023) is adapted as a multi-label classifier and evaluated through FCA and ACF. CEBRA

(Schneider et al., 2023) is adapted as an embedding model and evaluated through FCS and SCF. All methods use the same 250-ms inputs, movie annotations, participant-specific setup, channel-exclusion procedure, and BECA evaluation.

The proposed method reports higher median values for both forms of readout. In the classifier benchmark, median specificity increases from 54.60 ± 3.71 for EEG Conformer to 79.40 ± 3.20, while temporal alignment increases from 0.10 ± 0.20 to 1.29 ± 0.13. In the embedding benchmark, specificity increases from 52.10 ± 3.15 for CEBRA to 83.85 ± 3.10, and temporal alignment increases from -0.22 ± 0.14 to 1.03 ± 0.15. Table 5.3 reports the output-matched comparison.

Table 5.3. BECA comparison with output-matched neural-decoding baselines. Values are median ± SEM across pooled unit–readout observations.

Evaluation space	Model	Specificity	Temporal alignment
Classifier activation	EEG Conformer	54.60 ± 3.71	0.10 ± 0.20
Classifier activation	**Proposed method**	**79.40±3.20**	**1.29±0.13**
Embedding similarity	CEBRA	52.10 ± 3.15	-0.22 ± 0.14
Embedding similarity	**Proposed method**	**83.85±3.10**	**1.03±0.15**

These descriptive comparisons support the value of the proposed representation and adaptation pipeline under BECA. Formal participant-level inferential comparisons between architectures are not reported, and the pooled unit–readout observations are not fully independent. The results therefore do not establish universal superiority over the baseline architectures, which were originally designed for different objectives and were adapted to this dataset. They are specific to participant-level movie-to-sleep transfer and the held-out-neuron evaluation protocol.

5.6.2 Contribution of Sleep Alignment and Recall Safeguarding

The adaptation progression separates a movie-trained baseline, sleep CORAL without the recall safeguard, and full RASA. Across pooled BECA observations, median specificity rises from 64.75 ± 2.36 for the baseline to 78.20 ± 2.25 after sleep alignment and 82.48 ± 2.23 under RASA. Median temporal alignment changes from 0.99 ± 0.08 at baseline to 0.96 ± 0.11 under sleep CORAL and 1.17 ± 0.10 under RASA. Table 5.4 summarizes this pooled progression.

Table 5.4. Pooled BECA results across the adaptation progression. Values are median $\pm$ SEM across pooled unit–readout observations.

Adaptation stage	Specificity	Change from baseline	Temporal alignment
Movie-trained baseline	64.75 ± 2.36	—	0.99 ± 0.08
+ sleep CORAL	78.20 ± 2.25	+13.45	0.96 ± 0.11
Full RASA	**82.48±2.23**	**+17.73**	**1.17±0.10**

Sleep alignment accounts for most of the gain in specificity, showing that uncorrected state distribution differences limit the movie-trained representation. Its temporal-alignment statistic does not improve, however. Adding the recall safeguard produces a further 4.28-point specificity increase and raises temporal alignment above both earlier stages. This pattern supports the intended division of labor: moment matching improves access to the sleep distribution, while recall consistency helps protect the concept structure under that alignment pressure.

The improvement is not confined to one output head or conditioning direction. Relative to sleep CORAL alone, RASA increases specificity from 74.8 to 79.2 for FCA, 79.3 to 81.1 for ACF, 74.1 to 85.0 for FCS, and 81.2 to 83.9 for SCF. Temporal alignment

also increases in all four readouts. Table 5.5 places these changes alongside the movie-trained baseline and preserves the uncertainty reported for each readout.

Table 5.5. Per-readout BECA results across the movie-trained baseline, sleep CORAL, and full RASA. Values are median ± SEM across held-out concept-selective units.

BECA readout	Baseline specificity	Sleep CORAL specificity	RASA specificity	Baseline temporal alignment	Sleep CORAL temporal alignment	RASA temporal alignment
FCA	58.4 ± 4.7	74.8 ± 4.5	79.2 ± 4.5	0.89 ± 0.13	0.90 ± 0.17	0.95 ± 0.17
ACF	74.2 ± 4.7	79.3 ± 4.7	81.1 ± 4.6	1.61 ± 0.21	1.43 ± 0.24	1.95 ± 0.19
FCS	59.1 ± 4.9	74.1 ± 4.3	85.0 ± 4.4	0.77 ± 0.17	0.87 ± 0.25	0.99 ± 0.26
SCF	67.5 ± 4.7	81.2 ± 4.5	83.9 ± 4.5	0.93 ± 0.16	0.92 ± 0.16	1.10 ± 0.17

Figure 5.4 shows the full adaptation-component progression at the unit level. Across both the classifier-logit and embedding-similarity heads, sleep alignment accounts for the largest shift in specificity, while full RASA generally produces the strongest medians by combining sleep alignment with recall safeguarding.

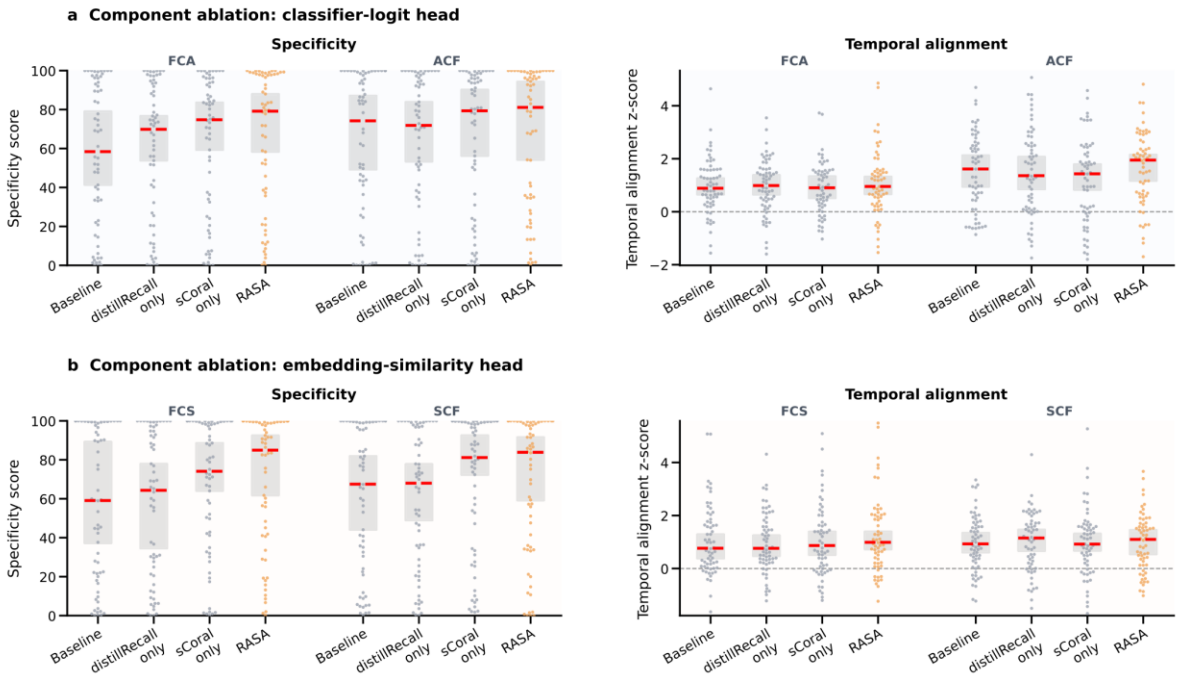

Figure 5.4. Component ablation of RASA across BECA readouts. (a) Classifier-logit evaluation shows specificity and temporal-alignment distributions for FCA and ACF under the movie-trained baseline, recall distillation alone, sleep CORAL alone, and full RASA. (b) Embedding-similarity evaluation shows the corresponding component comparison for FCS and SCF. Points denote individual held-out concept-selective units, red dashed lines denote medians, and gray bands denote 95% bootstrap confidence intervals of the median. Reproduced from Ding et al. (2026), Figure 4.

The unit-level view in Figure 5.5 further shows that RASA performance is distributed across participants and held-out units rather than attributable to one reference neuron. Nevertheless, observations from the same participant or neuron are not fully independent. The reported SEM values and bootstrap intervals describe pooled evaluation distributions and should not be interpreted as participant-level confidence intervals.

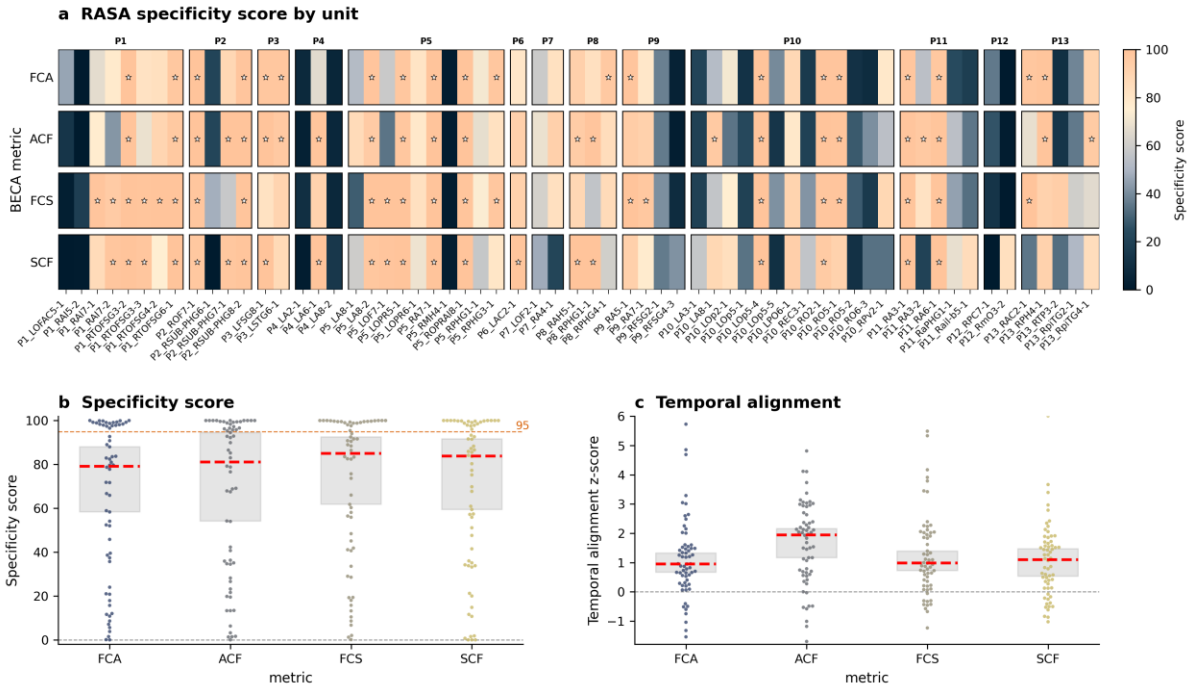

Figure 5.5. Population-level BECA evaluation under RASA. (a) A unit-level heatmap reports RASA specificity for FCA, ACF, FCS, and SCF. Columns are held-out concept-selective units grouped by participant, electrode localization, and unit number; stars mark unit–readout pairs exceeding the 95th percentile of the circular-shift surrogate distribution. (b) Unit-level specificity distributions and (c) temporal-alignment distributions summarize the same four readouts. Points denote individual units, red dashed lines denote medians, and gray bands denote 95% bootstrap confidence intervals of the median. The temporal-alignment display uses a robust plotting range, while summary statistics are calculated from the unclipped values. Reproduced from Ding et al. (2026), Figure 5.

5.6.3 Architectural and Safeguard Ablations

The dual-head ablation tests whether classifier and embedding supervision contribute complementary structure. In the classifier-activation benchmark, the full dual-head model improves specificity from 56.10 for a classifier-only model to 68.40 and temporal alignment from 0.94 to 1.08. In the embedding benchmark, the full model improves specificity from 53.45 for the embedding-only model to 62.25, although the embedding-only model has sharper temporal alignment (1.07 versus 0.86). Joint training therefore improves overall specificity but does not dominate every readout metric. The

embedding branch alone can produce temporally localized similarity structure, while the classifier objective appears to help stabilize which concept that structure represents.

The form of recall regularization also matters. Adding CORAL to both sleep and recall treats recall as another distribution to match and reduces mean specificity from 77.34 under sleep alignment alone to 75.21. Using recall distillation instead increases mean specificity to 82.26 and yields comparable or stronger temporal alignment across the classifier and embedding readouts. Figure 5.6 shows this comparison separately for the two output heads. This result supports recall's role as a semantic consistency condition rather than a second global alignment target.

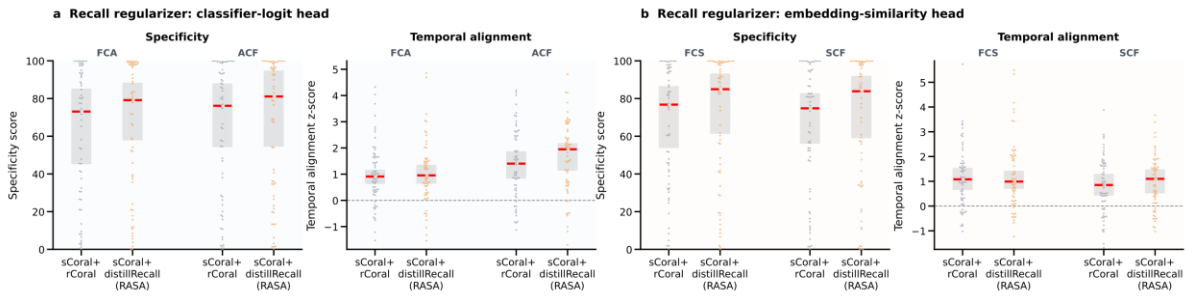

Figure 5.6. Recall distillation as a semantic safeguard during sleep adaptation. (a) Classifier-logit BECA distributions compare sleep CORAL plus recall-domain CORAL with sleep CORAL plus recall distillation (full RASA), reporting specificity and temporal alignment for FCA and ACF. (b) Embedding-similarity distributions show the same comparison for FCS and SCF. Points denote individual held-out units, red dashed lines denote medians, and gray bands denote 95% bootstrap confidence intervals of the median. Reproduced from Ding et al. (2026), Figure 6.

5.7 Scientific Interpretation

The results support a constrained form of cross-state concept transfer. A representation learned from labeled movie viewing can be adapted toward unlabeled NREM sleep so that its classifier and embedding signals show stronger event-locked coupling with independent neurons selective for the same concepts. Sleep distribution alignment provides the largest specificity improvement, while recall-based distillation

further improves concept specificity and temporal localization. This demonstrates that unlabeled intermediate and target recordings can play distinct roles: the target state can guide distribution adaptation, and a behaviorally meaningful intermediate state can limit semantic drift without supplying dense labels.

The dual-head representation also changes what can be evaluated. Chapter 4 treated each concept primarily through an independent classifier activation. Here, learned concept anchors create a geometry in which neural activity can be compared with several concepts and with a null state. The classifier and embedding readouts do not provide identical evidence, but their agreement across bidirectional BECA directions makes it less likely that the result is an artifact of a single output parameterization.

BECA contributes an equally important part of the answer. When target-state labels are unavailable, a global domain metric cannot show that a representation retains the intended latent variable. A held-out, concept-selective neuron offers an auxiliary observation of that variable for a small subset of concepts. Excluding its channel and separating adaptation from evaluation create a test of population-derived information against an independent measurement. This principle may extend to other neural problems in which dense labels disappear but a sparse physiological, behavioral, or cross-modal reference remains available. The interpretation must remain bounded by the reference: event coupling with a selective neuron is evidence that the model preserves a content-related signal during sleep, not a transcript of sleep mentation or direct proof of episodic replay. The state-specific activity of a concept-selective unit may reflect processes other than conscious memory retrieval, and the model may align with neural

factors correlated with that firing. Bidirectional, concept-matched event locking and circular-shift controls narrow these explanations but do not eliminate them.

5.8 Remaining Limitations and the Need for Broader Validation

The study uses 13 participants with drug-resistant epilepsy, clinically determined electrode placement, and participant-specific models. It does not establish generalization to a new participant, hospital, recording modality, or electrode configuration. The semantic vocabulary is closed and defined by the viewed movies. Concept-selective references are available for only 62 units and a subset of concepts, so BECA cannot verify the complete latent content of sleep. Concepts without an identified reference remain unevaluable by this method, and a failure to detect a selective unit is not evidence that the population lacks the corresponding representation. Several methodological assumptions remain. Inputs use fixed 250-ms windows, which may miss faster or sequence-level replay and longer consolidation dynamics. CORAL aligns only first- and second-order moments and may leave nonlinear state differences unresolved or suppress biologically meaningful state structure. Recall distillation preserves the output of a movie-trained teacher, including any source-specific biases. Adaptation and evaluation intervals are temporally disjoint but come from the same participant and night. BECA depends on event thresholds, window definitions, and the interpretation of concept-selective firing as an auxiliary reference; pooled unit–readout statistics do not fully account for participant-level dependence.

The framework is observational. It does not show that an aligned sleep event causes later memory change, that RASA discovers the biological mechanism of consolidation, or that the participant consciously experiences the decoded concept.

Prospective replication, participant-level hierarchical inference, cross-night evaluation, open-vocabulary concepts, richer temporal models, and causal perturbation would be required to support broader claims. Across Chapters 2–5, the dissertation has progressed from learning within noisy clinical candidate spaces to designing identifiable data, decoding content during internally generated recall, and transferring concept structure when behavioral labels disappear. Chapter 5 demonstrates a route to within-participant transfer and independent target-state evaluation, but generalization across people, institutions, tasks, modalities, and open semantic spaces remains unresolved. Chapter 6 therefore synthesizes which principles have been established, which conclusions remain study-specific, and what data and evaluation structures will be required for broader models of human neural activity.

Chapter 6 General Discussion

6.1 A Unified Progression of Neural Representation Learning

Across this dissertation, the representation target progresses from clinically proposed neural events to naturalistic cognition, specific episodic content, and concept-related population activity across wake and sleep. The scientific continuity does not come from applying one architecture, loss function, or latent space to every problem. It comes from asking the same question under changing evidential conditions: what structure can be learned from human neural activity, what makes that structure relevant to the intended target, and what observations justify its interpretation? As the target becomes less directly observable, the answer depends increasingly on the joint design of the candidate space, learning supervision, representation, and evaluation. A useful representation is therefore

not defined here by compression, separability, or predictive accuracy alone. It is defined operationally and comparatively by whether it relates to the intended target, remains informative under the variation relevant to the claim, and is supported by evidence not reducible to the information used to train it.

Three principles emerge from the progression. First, limited supervision is addressed by redistributing expert and experimental knowledge rather than by eliminating it. In the clinical studies, legacy detectors define a broad candidate space, self-supervised learning captures recurring morphology, resection-related information weakly orients the learned structure, and patient-level outcome and expert review supply additional evidence. In the memory studies, dense annotations during movie viewing supervise a source-domain representation, later vocal reports test that representation during internally generated recall, and unlabeled NREM recordings guide adaptation while recall-based consistency constrains semantic drift. The amount and location of supervision change across these settings, but none of the resulting scientific claims is independent of domain knowledge. Self-supervision and domain adaptation reduce dependence on exhaustive target labels; they do not make target definition or interpretation automatic.

Second, data design is part of the model of the scientific problem. Omni-iEEG demonstrates this point for an existing clinical target: without harmonized signals, metadata, annotations, patient splits, and clinically grounded metrics, apparent model differences can be confounded with institutional conventions. The Movie–Recall–Sleep paradigm demonstrates the same point for a more abstract target: a single annotated experience, later behavioral tests, recordings across states, and held-out concept-selective neurons create the relations through which episodic content can be learned and

evaluated. In both cases, more observations are useful only when their relationship to the target is explicit. Empirical identifiability is thus not a property of data volume alone; it is a property of the correspondence among the scientific construct, what is recorded, how supervision is supplied, and how the resulting claim can be tested.

Third, evaluation evidence must be designed at the same time as the representation. Clinical outcome, resection anatomy, expert annotation, recall vocalization, temporal controls, and held-out neuronal firing are not interchangeable ground truths. Each supports a different part of an interpretation and carries different limitations. Their value lies partly in their separation from training: surgical outcome is not an exact event label, recall reports are not movie annotations, and a held-out selective neuron is excluded from the population model whose sleep signal it evaluates. Across the dissertation, the strongest conclusions arise not from one definitive metric but from converging evidence with known roles and known boundaries. This is the practical meaning of a meaningful representation in the present work. It does not imply unique recovery of a biological code; it identifies a learned structure whose scientific interpretation is better supported than plausible alternatives under the tested conditions.

6.2 From Biomarkers to Memories

The movement from HFOs to episodic memories may initially appear to cross unrelated scientific domains, but it exposes a continuous change in the representation-learning problem. An HFO candidate has an operational time, channel, and morphology, yet its pathological meaning is uncertain. A recalled concept has semantic identity, but its internal occurrence cannot be enumerated from the neural signal alone. During sleep, even the behavioral report that anchors recall disappears. The progression is therefore

not simply from a clinical topic to a cognitive topic. It is from targets whose candidate instances can be proposed before learning to targets whose instances and evaluation evidence must be created by the experimental paradigm. Abstraction in this dissertation refers to this increasing distance between the underlying scientific construct and the observations directly available to the model.

This change alters where prior knowledge enters the analysis. For HFOs, rule-based detectors encode assumptions about frequency, duration, amplitude, and local signal statistics. Their output can be treated as a candidate pool without treating it as ground truth, allowing a VAE to learn morphological regularities and clinical variables to orient the discovered structure (Zhang et al., 2025). Because the VAE decoder maps the latent variables back to event morphology, the representation can also be inspected rather than treated only as an intermediate feature vector. For episodic content, the prior structure is not a detector threshold but a designed semantic vocabulary tied to a naturalistic experience. Movie annotations identify when people, places, and objects are externally present; vocal reports identify sparse later expressions of the same content; concept-selective neurons provide a limited internal reference when behavior is absent. Both settings therefore begin with a structured observation space. What differs is whether that structure is inherited from an established signal-processing workflow or constructed through an experiment.

The relevant scale also changes. HFO learning operates on localized event windows whose morphology can be visualized, clustered, and related to clinical anatomy. Memory decoding combines activity from distributed neuronal populations because the target may be expressed across regions and timescales rather than by one preidentified

unit. In Chapter 4, the model is trained from only one movie exposure, uses clusterless voltage-spike trains from 40–96 clinically determined recording sites per participant, and does not require prior identification of concept-selective neurons (Ding et al., 2025b). This design moves the problem beyond reading out a small set of known selective cells: cross-channel attention can learn nonlinear interactions and timing relations distributed across the recorded population. The participant-specific models produce content-related activations before matching recall reports, including after sleep, while regional restriction analyses suggest that the information available to the decoder changes across recall conditions. Chapter 5 extends the target from independent classifier outputs to a concept-organized geometry and tests whether that structure can be adapted toward NREM sleep (Ding et al., 2026). These studies do not imply that the same latent variables describe HFO morphology and episodic memory. They show that representation learning provides a common way to separate raw observations from the target-related structure needed for a scientific task.

The comparison also changes the meaning of generalization. In the clinical setting, generalization primarily concerns patients, institutions, detectors, recording conventions, and outcome associations. In the memory setting, the principal challenge is transfer within a participant across perception, recall, and sleep, where neural distributions and available labels change together. A representation should not be invariant to all of these differences: sleep and wake are biologically distinct, and institutional or anatomical variation may contain information rather than nuisance alone. The appropriate goal is selective stability. Target-related relations should remain recoverable while state-, participant-, and acquisition-specific structure is modeled rather than indiscriminately

removed. RASA operationalizes this distinction by using unlabeled sleep to reduce distribution mismatch while using recall consistency to preserve the semantic organization learned from experience.

Viewed together, the studies replace a simple performance hierarchy with an evidence hierarchy. Event classification within one cohort is weaker evidence than consistent patient-level clinical associations across institutions; movie prediction is weaker evidence for memory than concept-specific activation preceding held-out recall reports; global movie-to-sleep alignment is weaker evidence for retained content than bidirectional event coupling with an excluded concept-selective neuron. Each additional layer of evidence narrows the possible interpretation without making it unique. This progression explains why the dissertation does not culminate in one universally shared model. Participant-specific models, task-specific representations, and distinct evaluation protocols are appropriate to the present observations. The unification lies in how claims are constructed and tested, not in asserting that one representation applies unchanged across people, targets, and states.

6.3 Scientific and Clinical Implications

For machine learning, the principal implication is that representation quality in human neuroscience cannot be evaluated independently of measurement. Standard predictive evaluation assumes that labels capture the target and that held-out samples test the intended distribution. Both assumptions are fragile here. Pathological HFO labels remain contested, vocal reports incompletely reveal remembered content, and sleep lacks continuous content labels. Under these conditions, the task is not only to optimize a model against a fixed benchmark; it is to determine which observations can legitimately

function as supervision, orientation, control, or evaluation. The four-part framework developed in Chapter 3 (target operationalization, learning supervision, evaluation evidence, and relevant variation) offers a way to expose these choices. It is not proposed as a universal checklist, but it helps prevent a high-performing decoder from being interpreted more broadly than its target and test design permit.

The results also illustrate two complementary roles for representation learning. The first is discovery under imperfect labels. SS2LD learns event morphology before assigning a clinical orientation, allowing structure in the recordings to influence the boundary rather than requiring every event to inherit a disputed label (Zhang et al., 2025). The second is transfer under asymmetric supervision. The memory models use a densely labeled source state, a sparsely observed intermediate state, and an unlabeled target state for different purposes. A direct multi-label classifier ties each concept to a decision boundary and produces scalar scores; the contrastive head instead learns a reusable geometry in which related observations for the same concept are organized in a common embedding space. RASA then separates distribution alignment from semantic preservation by using sleep for the former and recall as an intermediate semantic constraint for the latter (Ding et al., 2026). These distinctions are relevant beyond the specific architectures used here. Neural datasets often contain abundant unlabeled signals but only narrow windows of reliable behavioral or physiological evidence. A productive learning design can assign each source of evidence a limited role instead of forcing all observations into a single nominal label space.

For neuroscience, the dissertation provides evidence that distributed human neuronal populations contain content-related structure linking a single naturalistic

experience to later recall. The activations preceding matching vocalizations, their temporal controls, and the persistence of population-level decoding across the overnight interval support a constrained interpretation of neural reinstatement: aspects of a movie-trained population pattern recur when related content is internally retrieved. Differences between MTL- and FC-restricted decoding before and after sleep are compatible with reorganization across memory systems, but they do not isolate a causal effect of sleep or establish that one region replaces another. Chapter 5 further shows that a concept-aligned population representation can be adapted so that its sleep activity exhibits stronger content-matched temporal coupling with independently defined selective neurons. This is evidence for content-related cross-state structure, not direct observation of a remembered episode during sleep, conscious sleep mentation, or the mechanism of consolidation.

For clinical neural-signal analysis, the work argues for systems that preserve the distinction among candidate generation, learned prioritization, and expert interpretation. PyHFO 2.0 makes this separation visible in an open, inspectable workflow that combines established detectors, learned classifiers, model replacement, and interactive expert annotation (Ding et al., 2025a). SS2LD reduces the demand for exhaustive event annotation, and Omni-iEEG provides a common basis for testing models across heterogeneous sources. The Omni-iEEG experiments further show why no single performance number is sufficient: models with similar patient-outcome discrimination can differ in their ability to distinguish pathological channels, while end-to-end long-segment models and transferred audio representations can be competitive with explicitly designed biomarker pipelines (Duan et al., 2026). These findings do not establish that audio and

iEEG share one biological code. They show that useful structure may be transferred or learned outside a pre-specified biomarker representation and must then be tested against the clinical relation of interest. Together, the studies suggest that clinically useful machine learning need not begin by replacing established detectors or expert judgment. It can first make large candidate spaces more manageable, expose the basis of model outputs, and test whether learned prioritization is associated with independent clinical evidence. The present results are not sufficient for autonomous diagnosis or surgical planning: the clinical proxies are imperfect, external prospective validation remains limited, and patient outcomes depend on factors beyond the modeled signals. Human review and provenance therefore remain parts of the scientific and clinical design rather than deployment details.

The memory studies support a longer-term translational roadmap while also clarifying the distance that remains. A model-guided closed-loop system for enhancing memory during sleep would require access to stable human neural signals, an estimator of content-related population activity, verification when memory can be behaviorally observed, a target-state validation strategy when behavior is unavailable, and finally a causal intervention linked to later memory outcome. The present dissertation addresses the middle of this chain: intracranial recordings provide access, Chapters 4 and 5 develop population estimators, recall provides wake verification, and held-out concept-selective neurons provide a limited form of sleep verification. Human sleep stimulation studies show that appropriately timed perturbation can influence neural synchrony and later memory (Geva-Sagiv et al., 2023), but this dissertation does not design or test stimulation. Moving from observational decoding to intervention will require prospective

target selection, real-time uncertainty estimation, stimulation controls, safety constraints, and evidence that changing the model-defined event changes subsequent behavior.

6.4 Limitations

The broadest limitation is the observation regime shared by human intracranial research. Participants underwent monitoring for drug-resistant epilepsy, electrode placement was determined by clinical need, and anatomical coverage therefore differs systematically across individuals. The recordings provide rare access to local field potentials and single-neuron activity, but they are neither random samples of the brain nor representative samples of the general population. Participant-specific models accommodate variable coverage and avoid assuming channel correspondence that the data do not provide, yet they do not establish transfer to a new participant. Omni-iEEG expands the clinical evidence across institutions, but its cohort remains geographically and clinically incomplete, and its channel and anatomical distributions reflect clinical implantation rather than balanced sampling (Duan et al., 2026). The memory studies remain smaller, use closed semantic vocabularies, and draw their strongest conclusions from within-participant transfer. The dissertation therefore supports general principles about designing and evaluating representations more strongly than it supports a single model that generalizes across people.

Each target also inherits uncertainty from the evidence used to operationalize it. Detector-defined HFO candidates omit clinically relevant activity that falls outside their rules, while resection and surgical outcome are indirect and patient-level indicators rather than authoritative labels for individual events. Recall vocalizations reveal only memories that participants choose and are able to express, and their timing follows internal retrieval

while also reflecting speech planning. Concept-selective neurons provide sparse, content-specific sleep references, but their absence does not imply absence of a concept representation, and their firing during NREM may reflect processes other than episodic replay. The controls in Chapters 2, 4, and 5 reduce particular alternatives, but no evaluation source uniquely identifies the biological construct. Converging evidence should therefore be understood as increasing support for a target-related interpretation, not transforming a proxy into ground truth.

The modeling choices impose additional boundaries. PyHFO 2.0 currently supports a limited set of recording formats and offline rather than streaming analysis; its learned classifiers depend on how well their training data represent the recording and patient population to which they are applied (Ding et al., 2025a). SS2LD learns from fixed-duration time-frequency images, excludes signal content outside the analyzed frequency range, and remains bounded by a detector-defined candidate pool (Zhang et al., 2025). Omni-iEEG does not contain every proposed detector, artifact strategy, interchannel method, or epilepsy biomarker, and even high-consensus spkHFO annotations do not remove institutional and conceptual uncertainty about the event definition (Duan et al., 2026). The recall decoder uses a single principal episode, a small set of selected concepts, and all movie data for training; its Memory Confidence Score is a relative event-associated statistic rather than continuous recall accuracy. The cross-state model uses fixed 250-ms windows, a closed concept vocabulary, first- and second-moment CORAL alignment, and a movie-trained teacher whose biases may be preserved by recall distillation. Hyperparameter selection, threshold definitions, and pooled unit-level summaries create dependencies that are not fully represented by simple observation-

level uncertainty estimates. Richer models may improve performance, but greater capacity will not by itself correct incomplete target definitions, leakage between model development and evaluation, or insufficient participant-level replication.

The dissertation is also a conceptual synthesis of studies with different cohorts, signals, and models. The clinical HFO, recall-decoding, and cross-state-transfer results should not be read as stages of one end-to-end system or as evidence that a single latent representation was progressively refined across the PhD. Chapter 3 links them through the scientific roles of data and evidence, not through shared participants or identical inputs. This distinction limits direct quantitative comparisons across chapters but protects the central argument from an equally problematic inference: that a method validated for one target can be transferred to another merely because both use neural data. The common conclusions concern how supervision is allocated, how targets are made observable, and how interpretations are evaluated.

Finally, the results are observational and associational. Region-restriction models change both anatomy and input dimensionality and cannot identify causal necessity. Presleep and postsleep comparisons lack a matched wake-control interval and cannot attribute change uniquely to sleep. BECA measures event coupling between model signals and held-out neuronal firing; it does not show that either event causes consolidation or later behavior. Clinical outcome associations do not establish that acting on the predicted events will improve treatment. Causal claims will require perturbation, prospective intervention, and behavioral or clinical endpoints. Any move toward real-time cognitive decoding or stimulation would also amplify ethical concerns involving privacy, consent, autonomy, unintended memory effects, and unequal access. These questions

are not resolved by model accuracy and must be treated as design constraints before translation.

6.5 Future Directions

The first priority is broader and more prospective validation. Clinical event models should be tested on institutions, devices, and patient populations excluded from model development, with preprocessing and decision thresholds fixed before evaluation. Reproducible data organization, public model implementations, and preserved checkpoints, as provided with Omni-iEEG, make such tests possible but do not substitute for prospective validation. Future clinical resources should broaden demographic and institutional coverage, include additional annotators and biomarkers, and support graph-based or other interchannel methods alongside event-based and end-to-end baselines (Duan et al., 2026). Memory studies require larger cohorts, additional experiences, repeated nights, and participant-level statistical models that distinguish within-person replication from pooled event counts. Cross-person learning is an important goal, but it should not be framed as simply concatenating channels from different brains. Shared models will need representations of anatomy, electrode geometry, signal quality, and participant-specific adaptation, together with evaluations that separate transfer to new recordings from transfer to new people. Omni-iEEG provides part of this infrastructure for clinical iEEG; comparable multi-site resources with naturalistic behavior, microwire activity, and longitudinal memory tests remain rare.

A second direction is to expand the target beyond a small, closed concept vocabulary, one of the principal limitations identified by the cross-state study (Ding et al., 2026). Movie-derived labels can be connected to language, audio, vision, and structured

knowledge representations, allowing neural activity to be compared with semantic relations that were not manually enumerated for each participant. Multimodal foundation models may provide useful priors, but their semantic fluency should not be mistaken for neural validity. Their value must be tested against held-out behavior, physiology, or perturbation rather than against agreement with their own embeddings. Open-vocabulary evaluation will also require careful controls for correlated concepts, narrative context, perceptual similarity, and language-model priors. At the temporal level, sequence models should test whether memory-related activity is organized as brief events, ordered trajectories, compressed replay, or longer state-dependent dynamics rather than assuming that independent 250-ms windows are sufficient.

A third direction is to develop learning objectives around the actual asymmetry of available evidence. In clinical workflows, active learning could direct expert review toward events whose annotation would most change patient-level decisions, while generative or self-supervised models could learn from the much larger unlabeled recording archive. In cognitive experiments, behaviorally anchored recall can function as an intermediate domain, but future work should compare distillation, contrastive consistency, multimodal alignment, and nonlinear adaptation under preregistered target-state tests. Evaluation references need not be dense to be useful: selective neurons, physiological events, behavioral reports, stimulation responses, and cross-modal measurements can each provide sparse but independent constraints. The critical requirement is to keep their role separate from the information used to tune the model.

The ultimate translational direction is model-guided closed-loop stimulation during sleep. The clusterless input and participant-specific inference used for one-shot recall

decoding were chosen partly because they avoid the delay and selectivity assumptions of spike sorting and are compatible in principle with rapid online estimation (Ding et al., 2025b). PyHFO 2.0 is currently offline, but its reported computational profile likewise motivates future streaming event analysis rather than establishing that capability in the present system (Ding et al., 2025a). A credible closed-loop memory system would first detect a candidate content-related event in real time, estimate both its semantic identity and uncertainty, and determine whether the current sleep state is appropriate for intervention. Stimulation would then be delivered under randomized sham, mistimed, or content-mismatched controls, with subsequent memory behavior as the primary endpoint and neuronal dynamics as mechanistic evidence. The initial objective should be causal testing, whether perturbing a model-defined event changes consolidation, rather than an immediate claim of memory enhancement. Only after replication should the system optimize stimulation timing or content. Because a false positive, an incorrectly targeted memory, or excessive stimulation could have consequences not captured by average accuracy, abstention, participant-specific calibration, conservative stopping rules, and independent safety monitoring will be essential.

These directions preserve the central lesson of the dissertation. Larger datasets and more expressive models are necessary for broader neural representations, but they are not sufficient. Each expansion of scale or capability must be accompanied by a clearer target, an explicit generalization axis, and evidence that remains meaningful outside the training objective. The next generation of neural models should therefore be judged not only by how much neural activity they can encode, but by which scientific questions their representations make answerable and which causal tests they enable.

Chapter 7 Conclusion

This dissertation asked how machine learning can recover and evaluate meaningful representations from human neural activity as the target progresses from clinical neural events to episodic content across experience, recall, and sleep. The answer is not a single model or a claim to have recovered the unique biological code underlying these phenomena. Meaningful representations become learnable when the target is operationalized through appropriate observations, when supervision is allocated according to what each source of evidence can support, and when the learned structure is tested under the variation relevant to the scientific claim. Their meaning is empirical: it is strengthened when target-related predictions persist beyond the training condition and agree with evidence that was not used to define the model.

The clinical studies established the first form of this answer. Legacy detectors, self-supervised morphology learning, weak clinical orientation, and patient-level evaluation can be separated so that clinically useful event structure is learned without exhaustive event annotation. PyHFO 2.0 preserves candidate provenance and expert review in an open workflow, SS2LD demonstrates label-efficient and inspectable morphology learning, and Omni-iEEG makes multi-center comparisons more reproducible through harmonized data and clinically grounded benchmarks while testing both biomarker-based and end-to-end representations. These contributions show that a clinical representation is supported not by detector agreement or one aggregate metric alone, but by the convergence of morphology, anatomy, patient outcome, institutional variation, and human interpretation.

The memory studies extended the same logic to targets that cannot be proposed directly by a signal-level detector. A purpose-designed Movie–Recall–Sleep paradigm

connected one naturalistic experience to semantic annotations, later behavior, and neuronal activity across states. Participant-specific population models learned from a single movie viewing and clusterless activity across clinically determined sites, without requiring preselection of concept-responsive neurons, and produced content-related activations before matching recall reports. This result demonstrates that information about a single episode can be recovered during later internally generated cognition under the tested conditions. A dual-head model then transformed independent concept scores into a reusable concept-organized space, and recall-adapted sleep alignment improved the correspondence between population-derived model signals and held-out neurons selective for the same concepts during NREM sleep. The resulting evidence supports a constrained form of cross-state concept transfer; it does not establish direct access to subjective sleep content or causal enhancement of memory.

Taken together, the dissertation shows that representation learning for human neural activity is as much a problem of scientific design as of model architecture. Candidate spaces, annotations, auxiliary domains, controls, benchmarks, and independent references determine what a neural representation can be claimed to capture. The progression from biomarkers to memories therefore leads to a general conclusion: as targets become more abstract and target-state labels disappear, stronger representational claims require more deliberate observation and evaluation, not merely larger models. This principle provides a foundation for future work on cross-person neural models, open semantic spaces, prospective clinical validation, and eventually closed-loop experiments that test whether model-defined memory events can be causally influenced during sleep.

REFERENCES

Bellier, L., Llorens, A., Marciano, D., Gunduz, A., Schalk, G., Brunner, P., and Knight, R. T. (2023). Music can be reconstructed from human auditory cortex activity using nonlinear decoding models. *PLOS Biology, 21*(8), e3002176.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning, 79*, 151–175.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Brodt, S., Inostroza, M., Niethard, N., and Born, J. (2023). Sleep—a brain-state serving systems memory consolidation. *Neuron, 111*(7), 1050–1075.

Buccino, A. P., Garcia, S., and Yger, P. (2022). Spike sorting: New trends and challenges of the era of high-density probes. *Progress in Biomedical Engineering, 4*(2), 022005.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*.

Danker, J. F., and Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin, 136*(1), 87–102.

Davis, T. S., Caston, R. M., Philip, B., Charlebois, C. M., Anderson, D. N., Weaver, K. E., Smith, E. H., and Rolston, J. D. (2021). LeGUI: A fast and accurate graphical

user interface for automated detection and anatomical localization of intracranial electrodes. *Frontiers in Neuroscience, 15*, 769872.

Ding, Y., Zhang, Y., Duan, C., Daida, A., Zhang, Y., Kanai, S., Lu, M., Hussain, S. A., Staba, R. J., Nariai, H., et al. (2025a). PyHFO 2.0: An open-source platform for deep learning–based clinical high-frequency oscillations analysis. *Journal of Neural Engineering, 22*(5), 056040.

Ding, Y., Dunn, S. L. S., Sakon, J. J., Aghajan, Z. M., Duan, C., Zhang, Y., Berger, J. I., Rhone, A. E., Nourski, K. V., Kawasaki, H., Howard, M. A. III, Roychowdhury, V. P., and Fried, I. (2025b). Reading specific memories from human neurons before and after sleep. *bioRxiv*. https://doi.org/10.1101/2025.07.01.662486

Ding, Y., Duan, C., Dunn, S. L. S., Sakon, J. J., Zhang, Y., Berger, J. I., Rhone, A. E., Garcia, C. M., Nourski, K. V., Kawasaki, H., Howard, M. A. III, Fried, I., and Roychowdhury, V. P. (2026). Concept-aligned neuronal representation transfer across brain states. Research manuscript.

Duan, C., Zhang, Y., Kanai, S., Ding, Y., Daida, A., Yu, P., Zheng, T., Kuroda, N., Hussain, S. A., Asano, E., Nariai, H., and Roychowdhury, V. P. (2026). Omni-iEEG: A large-scale, comprehensive iEEG dataset and benchmark for epilepsy research. In *The Fourteenth International Conference on Learning Representations*.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron, 33*(3), 341–355.

Foster, D. J., and Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440*(7084), 680–683.

Frauscher, B., von Ellenrieder, N., Zelmann, R., Rogers, C., Nguyen, D. K., Kahane, P., Dubeau, F., and Gotman, J. (2018). High-frequency oscillations in the normal human brain. *Annals of Neurology, 84*(3), 374–385.

Fried, I., Wilson, C. L., Maidment, N. T., Engel, J. Jr., Behnke, E., Fields, T. A., MacDonald, K. A., Morrow, J. W., and Ackerson, L. (1999). Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients: Technical note. *Journal of Neurosurgery, 91*(4), 697–705.

Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., and Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science, 322*(5898), 96–101.

Gerken, F., Darcher, A., Gonçalves, P. J., Rapp, R., Elezi, I., Niediek, J., Kehl, M. S., Reber, T. P., Liebe, S., Macke, J. H., Mormann, F., and Leal-Taixé, L. (2024). Decoding movie content from neuronal population activity in the human medial temporal lobe. *bioRxiv*. https://doi.org/10.1101/2024.06.13.598791

Geva-Sagiv, M., Mankin, E. A., Eliashiv, D., Epstein, S., Cherry, N., Kalender, G., Tchemodanov, N., Nir, Y., and Fried, I. (2023). Augmenting hippocampal–prefrontal neuronal synchrony during sleep enhances memory consolidation in humans. *Nature Neuroscience, 26*(6), 1100–1110.

Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., and Kording, K. P. (2020). Machine learning for neural decoding. *eNeuro, 7*(4).

Gotman, J. (2010). High frequency oscillations: The new EEG frontier? *Epilepsia, 51*(Suppl. 1), 63.

Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications, 8*, 15037.

Jacobs, J., Zijlmans, M., Zelmann, R., Chatillon, C.-E., Hall, J., Olivier, A., Dubeau, F., and Gotman, J. (2010). High-frequency electroencephalographic oscillations correlate with outcome of epilepsy surgery. *Annals of Neurology, 67*(2), 209–220.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016* (pp. 694–711). Springer.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In *Advances in Neural Information Processing Systems, 33*.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., et al. (2021). WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*.

Li, Y., Pazdera, J. K., and Kahana, M. J. (2024). EEG decoders track memory dynamics. *Nature Communications, 15*, 2787.

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., et al. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. *Nature, 620*(7976), 1037–1046.

Nagahama, Y., Dewar, S., Behnke, E., Eliashiv, D., Stern, J. M., Kalender, G., Fields, T. A., Wilson, C., Staba, R., Engel, J. Jr., and Fried, I. (2023). Outcome of stereo-

electroencephalography with single-unit recording in drug-refractory epilepsy. *Journal of Neurosurgery, 139*(6), 1588–1597.

Nariai, H., Wu, J. Y., Bernardo, D., Fallah, A., Sankar, R., and Hussain, S. A. (2018). Inter-rater reliability in visual identification of interictal high-frequency oscillations on electrocorticography and scalp EEG. *Epilepsia Open, 3*, 127–132.

Niediek, J., Reber, T. P., Bausch, M., Schwimmbeck, F., Gast, H., Coenen, V. A., Boström, J., Elger, C. E., and Mormann, F. (2026). Episodic memory consolidation by reactivation of human concept neurons during sleep reflects contents, not sequence of events. Research manuscript.

Quian Quiroga, R., and Panzeri, S. (2009). Extracting information from neuronal populations: Information theory and decoding approaches. *Nature Reviews Neuroscience, 10*(3), 173–185.

Quian Quiroga, R., Reddy, L., Koch, C., and Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of Neurophysiology, 98*(4), 1997–2007.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435*(7045), 1102–1107.

Rasch, B., and Born, J. (2013). About sleep's role in memory. *Physiological Reviews, 93*(2), 681–766.

Schneider, S., Lee, J. H., and Mathis, M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature, 617*(7960), 360–368.

Song, Y., Zheng, Q., Liu, B., and Gao, X. (2023). EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31*, 710–719.

Spring, A. M., Pittman, D. J., Aghakhani, Y., Jirsch, J., Pillay, N., Bello-Espinosa, L. E., Josephson, C., and Federico, P. (2017). Interrater reliability of visually evaluated high frequency oscillations. *Clinical Neurophysiology, 128*(3), 433–441.

Squire, L. R., Genzel, L., Wixted, J. T., and Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor Perspectives in Biology, 7*(8), a021766.

Sun, B., and Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops* (pp. 443–450). Springer.

Takashima, A., Nieuwenhuis, I. L. C., Jensen, O., Talamini, L. M., Rijpkema, M., and Fernández, G. (2009). Shift from hippocampal to neocortical centered retrieval network with consolidation. *Journal of Neuroscience, 29*(32), 10087–10093.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience, 26*, 858–866.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson (Eds.), *Organization of Memory* (pp. 381–403). Academic Press.

Vaz, A. P., Wittig, J. H. Jr., Inati, S. K., and Zaghloul, K. A. (2020). Replay of cortical spiking sequences during human memory retrieval. *Science, 367*(6482), 1131–1134.

Wang, S.-H., and Morris, R. G. M. (2010). Hippocampal-neocortical interactions in memory formation, consolidation, and reconsolidation. *Annual Review of Psychology, 61*, 49–79.

Wilson, M. A., and McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science, 265*(5172), 676–679.

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S.-L., Gertje, E. C., et al. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping, 36*(1), 258–287.

Zelmann, R., Mari, F., Jacobs, J., Zijlmans, M., Dubeau, F., and Gotman, J. (2012). A comparison between detectors of high frequency oscillations. *Clinical Neurophysiology, 123*(1), 106–116.

Zhang, Y., Aghajan, Z. M., Ison, M., Lu, Q., Roychowdhury, V., and Fried, I. (2023). Decoding of human identity by computer vision and neuronal vision. *Scientific Reports, 13*, 651.

Zhang, Y., Ding, Y., Duan, C., Daida, A., Nariai, H., and Roychowdhury, V. P. (2025). Self-supervised distillation of legacy rule-based methods for enhanced EEG-based decision-making. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2025* (pp. 477–486). Springer.

Zijlmans, M., Jiruska, P., Zelmann, R., Leijten, F. S. S., Jefferys, J. G. R., and Gotman, J. (2012). High-frequency oscillations as a new biomarker in epilepsy. *Annals of Neurology, 71*(2), 169–178.