CONCEPT-ALIGNED NEURONAL REPRESENTATION TRANSFER ACROSS BRAIN STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Neuronal representations formed during a given experience (e.g. place cells) reinstate during subsequent consolidation of that experience despite changes in behavioral state (e.g. place cell replay during sleep). Pinpointing when consolidation of specific memories occurs, especially those formed during one-shot, naturalistic experiences, remains a grand challenge. Detecting such internally-generated reinstatement in humans is particularly challenging with current methods. 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. Here, using intracranial microelectrode recordings comprising 452 curated single-neuron units from 13 neurosurgical participants across densely-labeled source (i.e. movie watching) and target states where labels remain sparse (i.e. vocalized memory recall) or unavailable (i.e. sleep), we address this problem using cross-state domain adaptation and held-out neuronal validation during sleep. We build participant-specific models using region-wise attention encoders trained with multi-label classification and contrastive alignment to learned concept embeddings. To improve adaptation across brain states, we introduce Recall-Adapted Sleep Alignment (RASA), which combines sleep-domain alignment with recall-based semantic regularization. Crucially, we leverage the rare presence of concept-selective neurons, known to reactivate during both perceptually driven and internally generated reinstatement of their given concept, to test our decoding framework. Each reference neuron's recording wire is excluded from model input throughout training, adaptation, and sleep evaluation. These neurons allow evaluation of decoding success during sleep by testing bidirectional temporal alignment between model-derived concept scores and their concept-selective activity even when behavioral labels are unavailable. Tested across 62 neurons selective to concepts during movie watching, RASA produces sleep-time concept scores that are more strongly concentrated around concept-matched held-out neuronal events and more precisely aligned with their timing than those from the movie-trained baseline and alternative neural decoders, providing evidence of sleep-time concept-aligned structure consistent with concept-specific reinstatement. We provide a general framework for transferring concept-aligned representations under severe cross-state distribution shift.

1 INTRODUCTION

The ability to decode internal representations—namely, memories—from neural data has potential applications in ameliorating memory dysfunction, targeted memory modulation, and brain-computer interfaces (Andersen et al., 2010; Mankin & Fried, 2020). Methods of decoding such internal representations, particularly for one-shot, naturalistic experience, remain absent. Decoding during stimulus-driven experience, such as movie watching, uses external event timing and semantic content as dense training labels (Gerken et al., 2025; Zhang et al., 2023). In contrast, during internally generated states such as memory recall (Ding et al., 2025; Gelbard-Sagiv et al., 2008) or sleep consolidation (Rasch & Born, 2013), the same memory content reinstates without precise behavioral reports or any behavioral labels at all. This creates a cross-state transfer problem: representations learned from densely labeled perception must generalize to target states in which supervision is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sparse, indirect, or absent. The challenge is especially severe for human single-neuron recordings, with limited units and sparse firing. Sleep is the most stringent target state, because it combines a lack of behavioral labels with substantial shifts in neural dynamics, including slow waves and spindles (Staresina et al., 2023). Thus, sleep decoding requires solving two coupled problems: how to transfer concept-aligned representations across brain states, and how to evaluate that transfer when target-state labels remain unavailable. In our setting, models learn from labeled movie-viewing activity and are evaluated on unlabeled NREM sleep.

While decoding of specific representations from neural data recorded during internally-driven states has been achieved in select use cases, these methods require unique data not practical for one-shot training in humans. In rodents, hippocampal ensemble activity during offline states can reinstate recent experience, whether through large-scale population firing rates (Gridchyn et al., 2020; Gulati et al., 2014; Maboudi et al., 2024; Peyrache et al., 2009) or temporally-compressed replay of place-cell sequences (Carr et al., 2011; Foster & Wilson, 2006; Nádasdy et al., 1999; Wilson & McNaughton, 1994). In humans, fMRI has reconstructed internally-generated semantic content, including continuous language representations during imagined speech and silent video viewing, but required many hours of training data (Tang et al., 2023). While intracranial recordings have shown that population spiking patterns can decode categorical information in human cortex, they do not pinpoint specific memories (Xie et al., 2024). In sum, decoding of out-of-distribution representations during memory consolidation of specific experiences remains limited to cases with high-density neuronal recordings, repeated structures, or data collection of a different kind than single-exposure naturalistic memory in humans.

Single neurons responsive to particular stimuli provide an especially direct form of content-specific neural reference. In particular, "concept cells" found in the human medial temporal lobe (MTL) provide an ideal marker of memory reinstatement, as individual neurons selectively and invariantly respond to specific people, places, or objects across images, text, and internally-generated recall (Gelbard-Sagiv et al., 2008; Quian Quiroga et al., 2009; 2007; 2005). Importantly, concept cells also reactivate during sleep (Schwimmbeck et al., 2026; Niediek et al., 2026), where they play a putative role in memory consolidation with increases in sleep activity relating to improvements in subsequent postsleep memory (Kehl et al., 2026). However, finding neurons selective to particular concepts requires extensive pre-screening, and even then only $\sim$5% of MTL neurons show visual selectivity and $\sim$2% the invariant responses characteristic of true concept neurons (Bausch et al., 2021). Thus, in a typical neurosurgical recording limited to a few dozen units, concept-selective cells may validate one or a few specific concepts, but cannot themselves provide dense labels for the many characters, places, and other abstract concepts present in a naturalistic experience.

Our central question is therefore: *How can we learn neural representations from stimulus-driven supervision capable of transferring to target states with sparse or absent labels, and can we evaluate the success of such domain transfer in out-of-distribution states like sleep?*

Cross-state transfer requires more than maximizing source-state decoding accuracy (Ben-David et al., 2010) as the representation must also remain informative when the same concepts reactivate in a different brain state (Polyn et al., 2005; Horikawa et al., 2013; Gelbard-Sagiv et al., 2008). Many neural decoders train as direct classifiers or regressors on task labels (Pereira et al., 2009), an approach that works well when training and evaluation occur under matched stimulus-driven conditions (Quian Quiroga et al., 2007; Liu et al., 2009; Pasley et al., 2012; Zhang et al., 2023). However, direct prediction ties each concept to domain decision boundaries learned at training that do not necessarily transfer under domain shift (Ben-David et al., 2010), producing independent scalar scores rather than a reusable concept space for internally-generated activity (Bengio et al., 2013). Here, we train a participant-specific decoder with a region-wise attention encoder (Vaswani et al., 2017) that models both local brain region structure and cross-region interactions, producing time-resolved concept scores for sleep evaluation. We use a classification objective that preserves direct decoding performance in the densely labeled source domain while a contrastive objective learns an embedding space in which neural activity organizes around reusable concept anchors for cross-state transfer and target-state validation (Khosla et al., 2020; Radford et al., 2021).

Even with this concept-aligned source representation, direct transfer from stimulus-driven movie viewing to sleep remains difficult because reinstated concepts occur amidst substantially different neural dynamics (Rasch & Born, 2013; Staresina et al., 2023). RASA addresses this problem by separating distribution alignment from semantic preservation. First, an unlabeled sleep segment

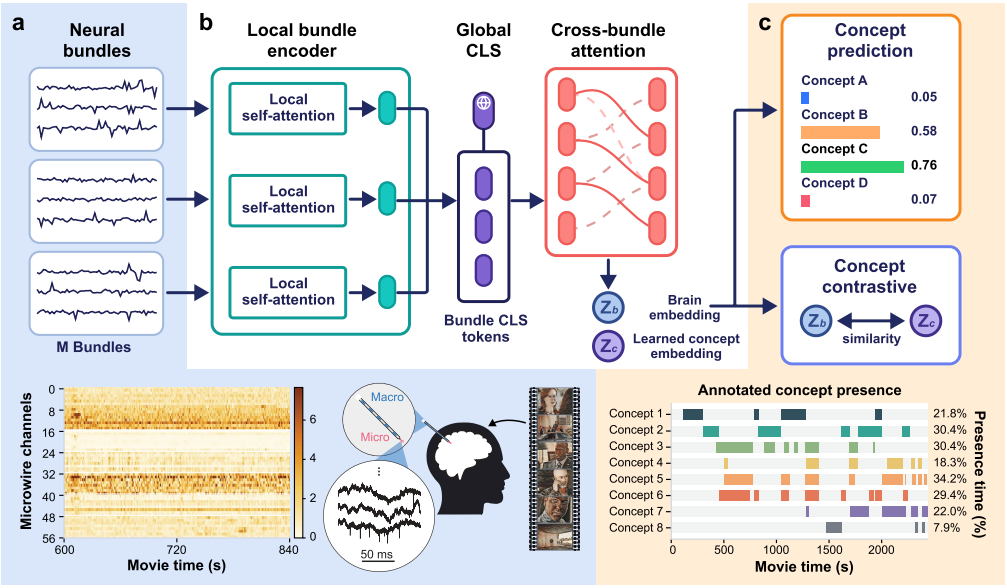


Figure 1: **Model overview.** **(a)** Movie-viewing inputs are constructed by aligning multi-label concept annotations from the episode to 250-ms neural windows. Neural input is represented as voltage-spike amplitude tensors with separate polarity channels; an example positive-polarity slice is shown with microwire bundles separated by horizontal lines. Concept occurrence tracks show when each annotated concept is present. **(b)** The shared neural encoder processes each input window using attention modules designed to capture both temporal structure within channels and interactions across channels. Cross-scale/channel attention blocks integrate information across microwires, while temporal self-attention summarizes activity within the 250-ms neural window. **(c)** Two complementary heads supervise the encoder output. A classifier head predicts the multi-label movie concept vector using a binary cross-entropy objective, while an embedding head aligns brain embeddings with concept embeddings using a contrastive alignment objective. Together, the two heads provide direct concept logits and embedding-based concept similarity readouts for downstream sleep decoding.

reduces the movie-to-sleep embedding distribution gap via domain adaptation (Sun & Saenko, 2016). Second, free recall acts as a semantic safeguard for the movie-trained concept geometry through distillation of recall embeddings against a frozen movie-trained teacher. Recall acts as a natural intermediate between movie viewing and sleep, since post-experience memories reinstate via similar processes as in sleep (Yang et al., 2024) with content tied to the movie.

Prior sleep-decoding studies rely on proxy references, including retrospective dream reports after awakening, externally delivered cues in targeted memory reactivation, or task-specific replay model templates (Belal et al., 2018; Bendor & Wilson, 2012; Horikawa et al., 2013; Rubin et al., 2022; Schönauer et al., 2017). An ideal reference for reactivated concepts during continuous sleep should be content-specific, observable in the target state, and independent of model input. Concept-selective neurons provide such a reference for the specific concepts they encode (Bausch et al., 2021; Gelbard-Sagiv et al., 2008; Kehl et al., 2026; Niediek et al., 2026; Quian Quiroga et al., 2005). The concept vocabulary is defined from movie annotations independently of subsequent concept-selective neuron identification; these neurons are used only as sparse, held-out references for sleep evaluation. For each held-out neuron, its recording wire is excluded from model input throughout training, adaptation, and sleep evaluation. We formalize this comparison with Bidirectional Event-Conditioned Alignment (BECA), which tests whether held-out selective-neuron activity and the corresponding model-derived concept score exhibit bidirectional event-locked alignment beyond circular-shift surrogates. We extend the discussion of related work in Appendix A.

This work makes three contributions:

- **Concept-aligned representation learning for neural decoding.** We propose a joint brain–concept embedding framework trained on stimulus-driven movie data, combining multi-label classification

with contrastive alignment to learned concept embeddings. This dual-head formulation supports direct concept decoding and learns a concept-aligned embedding space for cross-state transfer.

- **Recall-Adapted Sleep Alignment (RASA).** We propose RASA, an adaptation strategy for transferring movie-trained representations to sleep without target-state concept labels. RASA combines sleep-domain distribution alignment with a recall-guided semantic safeguard.
- **Held-out neuronal validation in unlabeled sleep.** We use held-out concept-selective neurons as independent neural references for evaluating concept transfer during sleep, where dense behavioral labels are unavailable.

2 METHOD

2.1 CROSS-STATE CONCEPT REPRESENTATION

We learn concept representations across three brain states with different supervision levels: movie watching $\mathcal{D}_{\text{movie}}$, free recall $\mathcal{D}_{\text{recall}}$, and NREM sleep $\mathcal{D}_{\text{sleep}}$. Movie watching provides dense time-aligned multi-label annotations over K concepts: for samples $(x, y) \sim \mathcal{D}_{\text{movie}}$, $y \in \{0, 1\}^K$. Free recall provides an internally-generated memory state over the same concept set, but not dense time-aligned labels. Sleep provides no dense stimulus-aligned or behavioral concept labels, making standard label-based target supervision and evaluation unavailable.

The neural encoder maps each neural signal input x to classifier logits $\ell(x) \in \mathbb{R}^K$ and a normalized embedding $\tilde{z}(x) \in \mathbb{R}^d$. Together with learned concept embeddings $\{c_1, \ldots, c_K, c_{\text{null}}\}$, where c_{null} anchors time bins with no annotated concept, we define the concept similarity score $s_k(x) = \cos(\tilde{z}(x), c_k)/\tau$, where τ is a temperature parameter. The logits support direct source-domain concept prediction, while the embedding space exposes structure not directly available from logits alone, including relative concept similarities, cross-concept geometry, and temporal trajectories. This space supports cross-state alignment and target-domain evaluation without dense labels.

Each participant-specific decoder is trained on annotated movie-viewing data and applied continuously to unlabeled NREM sleep, producing time-resolved scores for a closed set of movie concepts. BECA evaluates these scores using independent neuronal references.

2.2 DATASET AND PREPROCESSING

We analyze a rare intracranial microelectrode dataset from 13 epilepsy patients, all of whom underwent a standardized protocol spanning movie watching, free recall, and overnight sleep (Table 1). Patients viewed movie episodes while neural activity was recorded (30–45 min per viewing session), completed two free-recall blocks of previously viewed content, and were recorded during overnight sleep (5–10 hours per patient). The audiovisual stimuli were naturalistic segments from four movies or television series. Each decoder used a movie-specific vocabulary of 5–8 concepts, with 25 distinct concepts across all four stimuli (Appendix Table 9).

Table 1: Dataset overview. Intracranial micro-electrode recordings from 13 epilepsy patients spanning movie watching, free recall, and overnight sleep.

Name	Description	Value	Name	Description	Value
N_{subj}	Number of subjects	13	T_{recall}	Total recall duration (hr)	2.8
N_{movie}	Unique movies viewed	4	N_{elec}	Electrodes (microwires)	952
T_{movie}	Total movie viewing time (hr)	10.9	$\bar{N}_{\text{elec}}$	Avg. electrodes / subject	73
T_{NREM}	NREM sleep duration (hr)	58.4	N_{cells}	Concept-selective neurons	62

The analysis included microwire recordings from medial temporal lobe structures and other distributed brain regions (Appendix C.1.1). Coverage differed across participants because electrode placement was determined solely by clinical criteria. The encoder uses all eligible non-held-out microwire recordings to model local bundle structure and interactions across the distributed recording coverage.

Movie annotations were manually curated over time and represented as multi-label concept vectors on the fixed, non-overlapping 250-ms grid used by the neural input windows. Source annotations

were mapped or, where required, upsampled to this grid. Multiple concepts could co-occur within a window, while windows containing none of the predefined concepts received an all-zero label. The label distribution is imbalanced, with dominant concepts appearing in a majority of frames. Source-domain training therefore uses concept-balanced sampling.

The decoder input consists of clusterless voltage-spike amplitude signals from the retained microwires. The concept-selective neurons are identified from movie-viewing responses following the procedure of Ding et al. (2025). In brief, unit responses were tested on single-concept, character-face-visible events, and units were retained if they showed significant post-onset firing for exactly one concept under a sign-flip permutation criterion. These units serve as independent, held-out references for sleep evaluation; their firing activity is not used as model input or dense concept supervision. Overall, 62 of 452 curated units met the concept-selectivity criterion and entered the final evaluation cohort. Appendix C.2 details spike sorting, sleep scoring, unit curation, annotation, and unit selection. To prevent information leakage, each evaluated concept-selective neuron's recording wire is excluded from the model input throughout training, adaptation, and sleep evaluation.

2.3 Hierarchical Attention Neural Encoder

Each 250 ms input window is represented as $x \in \mathbb{R}^{2\times W\times T}$, where W is the number of recording wires, T is the number of temporal samples within the window, and the two input channels preserve positive- and negative-polarity spike components. Figure 1 summarizes the encoder.

The encoder uses a hierarchical attention design reflecting the anatomical grouping of the recordings. Wires are partitioned into anatomical 8-wire microwire bundles, and channel-specific attention (CSA) encodes temporal patches via convolution and self-attention to produce bundle-level tokens $h^{(b)}$. Cross-channel attention (CCA) then integrates the tokens across R bundles into a global CLS representation $z \in \mathbb{R}^d$. This two-stage design preserves local region-specific structure before modeling cross-region interactions. Two heads are applied to z: a classifier head produces source-domain concept logits $\ell = W_{\text{cls}} z \in \mathbb{R}^K$, and an embedding head produces the normalized neural embedding $\tilde{z} = W_{\text{emb}} z / \|W_{\text{emb}} z\|_2$. The embedding is compared with the learned concept embeddings and null embedding defined in Section 2.1; architecture hyperparameters are provided in Appendix C.3.

Source-domain objective. We train on samples $(x, y) \sim \mathcal{D}_{\text{movie}}$ with two complementary objectives. The classifier head uses standard multi-label binary cross-entropy over the K annotated concepts; to address label imbalance, we use concept-balanced sampling. The embedding head uses a multi-label contrastive objective, inspired by supervised contrastive learning (Khosla et al., 2020) but adapted to multi-label annotations where multiple concepts can co-occur within a single sample. Each positive pair comprises a 250-ms movie-window neural embedding and the learned embedding of a concept present in that window. When multiple concepts co-occur, each forms a separate positive pair; windows with no active concept are paired with the null anchor. For a sample x, let $\mathcal{P}(x)$ denote the active concepts, using the null concept when no annotated concept is active, and let $\mathcal{N}(x)$ denote the remaining inactive concept anchors. We minimize

$$\mathcal{L}_{\text{ctr}}(x) = -\sum_{k\in\mathcal{P}(x)} \log \frac{\exp(s_k(x))}{\sum_{j\in\mathcal{N}(x)\cup\{k\}} \exp(s_j(x))}.$$

This objective encourages each active concept to individually outrank all inactive concepts, rather than collapsing multiple active labels into a single target. The source objective is $\mathcal{L}_{\text{src}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}}$, with hyperparameters and concept-embedding initialization provided in Appendix C.4.

2.4 Recall-Adapted Sleep Alignment (RASA)

Recall-Adapted Sleep Alignment builds on a movie-trained concept representation and adapts it toward sleep with two embedding-level regularizers. During adaptation, the source objective remains active. Figure 3 illustrates these roles separately: sCoral primarily reduces the movie–sleep domain gap, whereas distillRecall alone does not perform global domain alignment but helps limit recall-domain displacement when combined with sleep alignment. Recall carries movie-related content in an internally generated state, providing a natural domain for semantic regularization during sleep alignment.

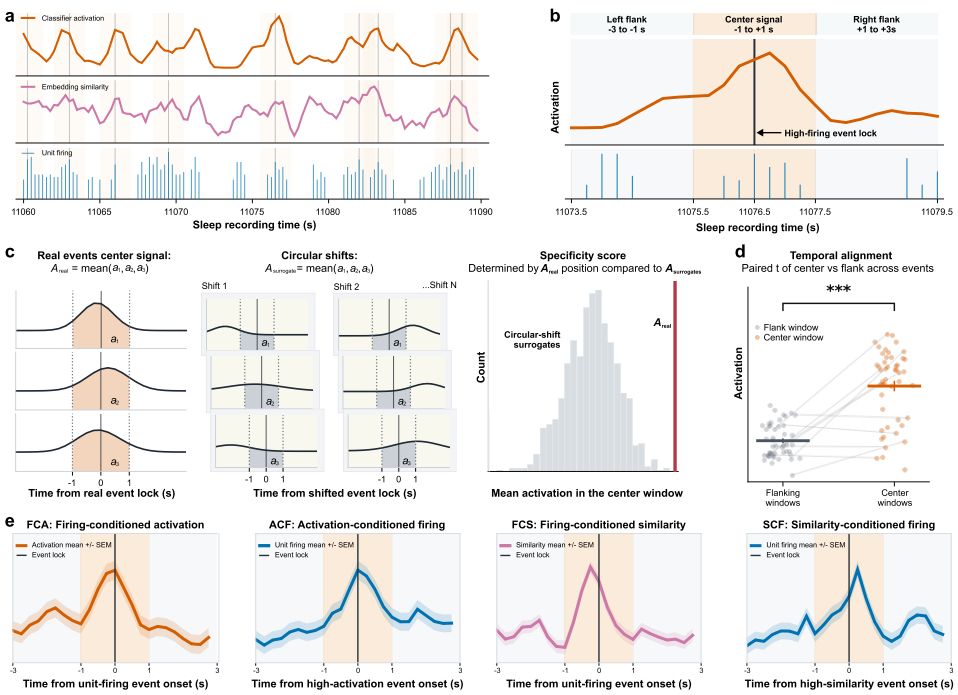


Figure 2: **Bidirectional Event-Conditioned Alignment (BECA) quantifies alignment between model-derived concept signals and concept-selective neuron activity during sleep. (a)** Example sleep segment from one participant showing classifier activation (top) and embedding similarity (middle) for a concept-selective unit's associated concept, together with that unit's firing rate. Shaded windows mark high-firing events. **(b)** Zoomed high-firing event from **a**. The event defines a center window (-1 to $+1$ s) and two flanking windows (-3 to -1 s and $+1$ to $+3$ s). **(c)** We assess event specificity by comparing the real event-locked center signal (left) against a circular-shift surrogate distribution (center). Each surrogate uses the same number of events after circularly shifting the event times. The percentile rank of the real center signal within this surrogate distribution defines the event-specificity score (right). **(d)** Temporal alignment is assessed with a paired test comparing the center-window signal with the flanking-window signal across events. For each event, the flanking value is the average of the left and right flank windows. A paired t-statistic quantifies peak sharpness, with larger values indicating stronger local alignment to the unit-firing event. This example shows 50 selected events. **(e)** The same BECA event-locking statistic is applied to four complementary directions: firing-conditioned activation, activation-conditioned firing, firing-conditioned similarity, and similarity-conditioned firing. Lines and shading denote mean event-locked traces and SEM.

Sleep-domain alignment. To adapt the source-trained encoder to the sleep domain without concept labels, we use an unlabeled NREM sleep segment from each patient. This sleep-adaptation segment is defined as the first contiguous 30-minute NREM interval and is excluded from all subsequent BECA evaluation. We align movie and sleep embeddings using CORAL (Sun & Saenko, 2016). Let $Z_M = \{\tilde{z}(x_m)\}$ denote movie-domain embeddings from the current minibatch and $Z_S = \{\tilde{z}(x_s)\}$ denote embeddings from the sleep-adaptation segment. Let μ_M, μ_S and Σ_M, Σ_S denote their empirical means and covariance matrices. The sleep-alignment loss is

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

This loss performs distribution-level alignment: it matches first and second moments of movie and sleep embeddings, but does not by itself preserve concept-specific geometry.

Recall-guided semantic safeguard. To preserve semantic structure during sleep alignment, we distill recall-domain embeddings against a frozen baseline teacher trained only on movie data. Let $\tilde{z}_{\text{student}}(x_r)$ and $\tilde{z}_{\text{teacher}}(x_r)$ denote the student and teacher embeddings for the same recall-domain

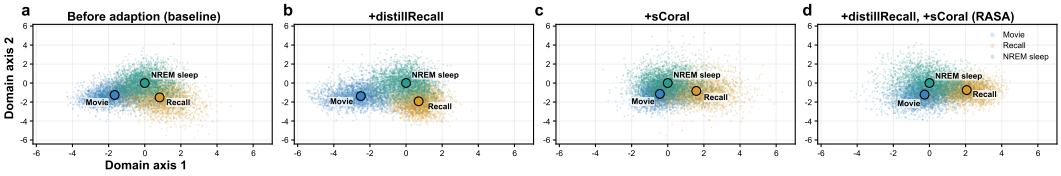


Figure 3: **Representative embedding-domain shifts under RASA adaptation components.** Sleep-centered two-dimensional Linear Discriminant Analysis (LDA) projections of the model's 128-dimensional, L2-normalized neural embeddings, fitted using domain labels for movie, recall, and NREM sleep. Each dot represents one 250-ms neural window, and outlined markers denote domain medians. Domains are equally subsampled, and coordinates are centered on the NREM median. **(a)** Before adaptation, the three domains occupy separated embedding regions because the representation is trained using only movie-domain labels. **(b)** Separately adding distillRecall leaves the global domain geometry similar to baseline, consistent with its role as a recall-domain semantic consistency constraint rather than a distribution-alignment loss. **(c)** Separately adding sCoral reduces the movie–sleep domain gap by matching movie and sleep embedding moments, but does not explicitly protect recall-domain semantic geometry. **(d)** Full RASA combines sleep-domain alignment with the recall safeguard: movie and sleep embeddings occupy a more shared region, while recall shows less displacement than under sCoral alone, consistent with distillRecall constraining recall geometry to the frozen baseline.

input $x_r \sim \mathcal{D}_{\text{recall}}$. We define the recall safeguard loss as

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

This objective acts as a cross-domain consistency regularizer: it encourages recall-domain embeddings to remain compatible with the concept geometry learned from labeled movie data, while avoiding direct supervision from sparse or incomplete recall labels. Thus, the full adaptation objective is

$$\mathcal{L}_{\text{RASA}} = \mathcal{L}_{\text{src}} + \lambda_{\text{sCoral}} \mathcal{L}_{\text{sCoral}} + \lambda_{\text{distill}} \mathcal{L}_{\text{distillRecall}}.$$

The three terms respectively maintain source-domain concept learning, align movie and sleep distributions, and preserve recall-domain semantic structure.

2.5 Measuring Sleep Transfer with Neuronal References

For each evaluated concept-selective neuron, we train a separate participant-specific model after excluding the neuron's recording wire from all model inputs. The wire remains excluded throughout movie training, cross-state adaptation, and sleep inference. The decoder uses the retained clusterless population signal, and the excluded neuron serves as an independent, content-specific validation reference. Sleep-domain transfer is assessed by testing whether model-derived concept signals are temporally coupled with independent neuronal references during sleep. We thus propose Bidirectional Event-Conditioned Alignment (BECA). For a neuron associated with concept k, let $a_k(t)$ denote the held-out unit firing activity, and let $m_k^{\text{cls}}(t) = \ell_k(t)$ and $m_k^{\text{emb}}(t) = s_k(t)$ denote the classifier-head activation and embedding-head similarity for the same concept (Figure 2a). Specifically, BECA measures this event-conditioned coupling in both directions. In the unit-conditioned direction, high-firing events are defined from the held-out unit, $\mathcal{E}_a = \{t : a_k(t) \geq \theta_a\}$, and the model response is evaluated around those events. In the model-conditioned direction, events are defined from the model signal, $\mathcal{E}_m = \{t : m_k(t) \geq \theta_m\}$, where $m_k(t)$ is either $\ell_k(t)$ or $s_k(t)$, and held-out unit firing is evaluated around those model events. The event thresholds select high-activity NREM periods separately for each neuron and model readout. Specifically, θ_a is the 90th percentile of nonzero held-out-unit firing during NREM, whereas θ_m is the 90th percentile of the corresponding NREM model signal. These unit- and readout-specific percentiles accommodate differences in firing rates and model-score scales while focusing the analysis on the strongest candidate events. Unless otherwise specified, the event-selection threshold is the 90th percentile.

As shown in Figure 2b, each event time t_0 defines a center window, $\mathcal{W}_c(t_0) = [t_0 - 1, t_0 + 1]$, and two flanking windows, $\mathcal{W}_f(t_0) = [t_0 - 3, t_0 - 1] \cup [t_0 + 1, t_0 + 3]$. For FCA and FCS, $q(t)$ denotes the corresponding classifier activation or embedding similarity; for ACF and SCF, $q(t)$ denotes held-out-unit firing activity. We use $\overline{q}_{\mathcal{W}}$ to denote the temporal mean of $q(t)$ within window $\mathcal{W}$. For evaluated

Table 2: **Comparison with SOTA decoding baselines under BECA.** Values are median ± SEM.

Evaluation setting	Model	Specificity	Temporal alignment
Classifier	EEG Conformer	52.80 ± 3.69	0.61 ± 0.18
	EEGNet	54.85 ± 3.37	0.77 ± 0.29
	SPaRCNet	51.00 ± 3.59	0.45 ± 0.21
	Ours (RASA)	$\mathbf{79.40 \pm 3.20}$	$\mathbf{1.29 \pm 0.13}$
Embedding	CEBRA	52.10 ± 3.15	-0.22 ± 0.14
	SIMPL	48.90 ± 3.50	-0.50 ± 0.22
	Ours (RASA)	$\mathbf{83.85 \pm 3.10}$	$\mathbf{1.03 \pm 0.15}$

signal $q(t)$, define the center-flank response $\Delta_q(t_0) = \overline{q}_{\mathcal{W}_c(t_0)} - \overline{q}_{\mathcal{W}_f(t_0)}$, where the flank value averages the left and right flanking windows. BECA summarizes event-conditioned alignment using two complementary statistics. The **specificity score** is the percentile rank of the real center-window response relative to a circular-shift surrogate distribution, where event times are circularly shifted to preserve temporal structure (Figure 2c). The **temporal-alignment score**, denoted z_{peak}, is the paired center-versus-flank t-statistic across events, measuring whether the evaluated signal shows a reliable local peak around the conditioning event (Figure 2d). Under the BECA null hypothesis, the model-derived concept signal and held-out neuronal reference have no concept-specific local coupling beyond the temporal structure captured by circular shifts. The specificity score therefore has a null median of 50, whereas temporal alignment is centered at zero.

BECA applies the event-conditioned statistics in four complementary perspectives (Figure 2e). Firing-conditioned activation (FCA) measures classifier activation around high unit-firing events, whereas activation-conditioned firing (ACF) measures unit firing around classifier-activation peaks. Firing-conditioned similarity (FCS) measures embedding similarity around high unit-firing events, whereas similarity-conditioned firing (SCF) measures unit firing around embedding-similarity peaks. Bidirectional alignment requires both unit-conditioned and model-conditioned directions to show concept-specific event locking. This provides complementary evidence of coupling, but does not by itself exclude shared physiological confounds such as arousal-linked activity. Threshold choices, surrogate construction, and event-window parameters are specified in Appendix C.5.

3 Results

3.1 Comparison with State-of-the-Art Neural Decoding Models

We compare our RASA-adapted model with five state-of-the-art neural decoding methods relevant to our setting: EEG Conformer (Song et al., 2023), EEGNet (Lawhern et al., 2018), SPaRCNet (Jing et al., 2023), CEBRA (Schneider et al., 2023), and SIMPL (George et al., 2025). We reproduce these methods and adapt their differing decoding objectives to our dataset and evaluation protocol. EEG Conformer, EEGNet, and SPaRCNet use the classifier-activation benchmark; CEBRA and SIMPL use the embedding-similarity benchmark. All analyses are performed on the common cohort of 62 concept-selective cells across 13 patients using the BECA evaluation framework defined in Section 2.5. Implementation details are provided in Appendix C.

RASA achieves the highest specificity and temporal alignment in both benchmarks (Table 2): 79.40 and 1.29 for classifier activation versus 54.85 and 0.77 for EEGNet, and 83.85 and 1.03 for embedding similarity versus 52.10 and -0.22 for CEBRA, the strongest competing models in their respective benchmarks.

3.2 RASA Recovers Concept-Aligned Structure During Sleep

We next compare RASA with the movie-trained baseline and sleep-domain alignment alone during NREM sleep. Across pooled unit-readout observations, median specificity increases from 64.75 ± 2.36 SEM for the movie-trained baseline to 78.20 ± 2.25 with sleep-domain alignment and 82.48 ± 2.23 with RASA (Table 3). Pooled median temporal alignment increases from 0.96 ± 0.11 under +sCoral to 1.17 ± 0.10 under RASA. These results indicate that sleep-domain alignment substantially improves transfer to NREM sleep, while recall-guided safeguarding further strengthens both concept specificity

Table 3: **RASA improves sleep BECA specificity and temporal alignment across the three-stage adaptation progression.** Values are median ± SEM across pooled unit-readout observations from FCA, ACF, FCS, and SCF.

	Specificity			Temporal alignment		
	Baseline	+sCoral	RASA	Baseline	+sCoral	RASA
Median ± SEM	64.75 ± 2.36	78.20 ± 2.25	**82.48 ± 2.23**	0.99 ± 0.08	0.96 ± 0.11	**1.17 ± 0.10**
Δ vs baseline	–	+13.45	**+17.73**	–	-0.03	**+0.17**
Δ vs +sCoral	–	–	**+4.28**	–	–	**+0.21**

and local event locking. Relative to +sCoral, RASA improves specificity and temporal alignment in all four directions (Appendix Table 4).

To test whether sleep-time alignment depends on concept identity, we compared each held-out neuron's matched-concept readouts with mismatched-concept readouts, keeping its firing activity fixed. We then randomized concept mappings within each movie vocabulary; matched mappings yielded higher pooled specificity and temporal alignment than the permutation null (Holm-adjusted $p = 0.00280$ for both outcomes; Appendix B.4). These results support concept-identity-specific alignment.

Population-level distributions and across-stage visualizations are provided in Appendix B.1 (Figures 4 and 5). These analyses show that the table-level gains are broadly reflected across cells rather than being driven by a single summary statistic.

3.3 Additional Results and Ablations

Dual-head ablations show that joint classifier and contrastive training improves overall specificity, while the embedding head contributes event-locked similarity structure (Appendix B.2). Complementing the sleep BECA results, recall-domain teacher consistency and geometry analyses support the interpretation that distillation preserves movie-trained representations during sleep alignment more effectively than recall-domain CORAL (Appendix B.3); they do not directly validate remembered content. BECA outcomes were also robust to event thresholds from the 85th to the 95th percentile (Appendix B.5).

4 Conclusion and Discussion

We study cross-state concept representation transfer across movie watching, free recall, and overnight NREM sleep using rare human intracranial microelectrode recordings. To address the absence of behavioral labels during sleep, we introduce BECA, which evaluates transferred representations through held-out concept-selective neurons as independent neural references. Empirically, the proposed hierarchical attention neural encoder trained with RASA achieves higher concept specificity and temporal alignment than alternative neural decoding models across classifier-activation and embedding-similarity benchmarks. Recall-domain teacher consistency and representational-geometry analyses further show that movie-trained semantic structure is better preserved under RASA than under sleep alignment alone or recall-domain CORAL. Together, these results provide evidence that participant-specific, concept-aligned neural structure learned during movie viewing can be transferred to NREM sleep and validated through concept-matched temporal agreement with independent neuronal references.

Limitations and future work. The present framework operates over a closed set of annotated movie concepts, rather than an open-vocabulary semantic space, and BECA can validate only concepts for which a held-out concept-selective neuron is available. Thus, our evaluation provides content-specific evidence for transferred representations, but not dense verification of all latent sleep content. Our analyses also use fixed 250 ms input windows, which may miss reinstatement dynamics at finer or longer temporal scales. Future work should extend concept-aligned transfer to open-vocabulary semantic embeddings, evaluate richer temporal or sequence-level memory structure, and test cross-subject generalization beyond patient-specific models.

AI USE STATEMENT

In this work, we used generative AI tools for [design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis]. We have not used generative AI tools for [Generate synthetic data sets, propose or refine hypotheses, interpret results], and [help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs] are not applicable to this work. Additionally, we used generative AI tools for [create or modify scientific figures or images, suggest experimental parameters, create or edit software code, draft parts of a research paper, summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, propose a title or keywords for a research paper]. We have reviewed all AI-assisted work. Specifically, the authors reviewed AI-assisted methodological suggestions and made the final decisions on experimental design and analysis. AI-assisted code was inspected and tested against the intended methods and evaluation protocols. Numerical values and statistical results presented in the text, tables, and figures were checked against the underlying analysis outputs. AI-assisted literature summaries and suggested references were checked against the original publications. The authors also reviewed AI-assisted text and figures for technical accuracy and consistency with the empirical evidence. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The research protocol was approved by the institutional review boards of both participating institutions, and all participants provided written informed consent prior to participation. Intracranial electrodes were implanted for clinical seizure monitoring, with placement determined solely by clinical criteria (Appendix C.1). The clinical recordings and copyrighted audiovisual stimuli cannot be publicly redistributed because of patient-privacy and copyright constraints. Our analyses provide evidence of concept-aligned neural structure during sleep; they do not establish access to subjective sleep content. The framework is intended for research and is not a deployable system for clinical decision-making or general-purpose monitoring of mental content.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the neural representation learning framework, RASA adaptation objectives, and held-out-wire evaluation protocol in Section 2. Recording procedures and preprocessing, including concept annotation, unit selection, and sleep scoring, are documented in Appendices C.1.1 and C.2. Architecture and training settings are provided in Appendices C.3 and C.4, and baseline implementations are described in Appendix C.7. The BECA event-selection rules, evaluation windows, and surrogate construction are specified in Appendix C.5. Concept-identity controls and event-threshold sensitivity analyses, including their randomization and bootstrap procedures, are reported in Appendices B.4 and B.5. The underlying clinical recordings and commercial video stimuli cannot be publicly redistributed because of patient-privacy and copyright restrictions.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush R. Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Richard A. Andersen, Eun Jung Hwang, and Grant H. Mulliken. Cognitive neural prosthetics. *Annual Review of Psychology*, 61:169–190, 2010. doi: 10.1146/annurev.psych.093008.100503.

Marcel Bausch, Johannes Niediek, Thomas P. Reber, Sina Mackay, Jan Boström, Christian E. Elger, and Florian Mormann. Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts. *Nature Communications*, 12:6164, 2021. doi: 10.1038/s41467-021-26327-3.

Suliman Belal, James Cousins, Wael El-Deredy, Laura Parkes, Jules Schneider, Hikaru Tsujimura, Alexia Zoumpoulaki, Marta Perapoch, Lorena Santamaria, and Penelope Lewis. Identification

of memory reactivation during sleep by eeg classification. *NeuroImage*, 176:203–214, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.04.029. URL https://www.sciencedirect.com/science/article/pii/S1053811918303380.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. URL http://www.springerlink.com/content/q6qk230685577n52/.

Daniel Bendor and Matthew A Wilson. Biasing the content of hippocampal replay during sleep. *Nature Neuroscience*, 15(10):1439–1444, 2012. ISSN 1546-1726. doi: 10.1038/nn.3203. URL http://dx.doi.org/10.1038/nn.3203.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50. URL https://doi.org/10.1109/TPAMI.2013.50.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV*, pp. 139–156, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01263-2. doi: 10.1007/978-3-030-01264-9_9. URL https://doi.org/10.1007/978-3-030-01264-9_9.

Margaret F. Carr, Shantanu P. Jadhav, and Loren M. Frank. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, 14(2):147–153, 2011. doi: 10.1038/nn.2732.

Fernando J. Chaure, Hernan G. Rey, and Rodrigo Quian Quiroga. A novel and fully automatic spike-sorting implementation with variable number of features. *Journal of Neurophysiology*, 120(4):1859–1871, October 2018. ISSN 1522-1598. doi: 10.1152/jn.00339.2018. URL http://dx.doi.org/10.1152/jn.00339.2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, Oct 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00714-5. URL https://doi.org/10.1038/s42256-023-00714-5.

Yuanyi Ding, Soraya L. S. Dunn, John J. Sakon, Zahra M. Aghajan, Chenda Duan, Yipeng Zhang, Joel I. Berger, Ariane E. Rhone, Kirill V. Nourski, Hiroto Kawasaki, Matthew A. Howard, Vwani P. Roychowdhury, and Itzhak Fried. Reading specific memories from human neurons before and after sleep. 2025. doi: 10.1101/2025.07.01.662486.

Chenda Duan, Yipeng Zhang, Sotaro Kanai, Yuanyi Ding, Atsuro Daida, Pengyue Yu, Tiancheng Zheng, Naoto Kuroda, Shaun A. Hussain, Eishi Asano, Hiroki Nariai, and vwani Roychowdhury. Omni-iEEG: A large-scale, comprehensive iEEG dataset and benchmark for epilepsy research. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rv9lQpY5cG.

David J. Foster and Matthew A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680–683, 2006. doi: 10.1038/nature04587.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, January 2016. ISSN 1532-4435.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. Rankme: assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Hagar Gelbard-Sagiv, Roy Mukamel, Michal Harel, Rafael Malach, and Itzhak Fried. Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, 322 (5898):96–101, 2008. doi: 10.1126/science.1164685. URL https://www.science.org/doi/abs/10.1126/science.1164685.

Tom George, Pierre Glaser, Kim Stachenfeld, Caswell Barry, and Claudia Clopath. SIMPL: Scalable and hassle-free optimisation of neural representations from behaviour. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9kFaNwX6rv.

Franziska Gerken, Alana Darcher, Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Johannes Niediek, Marcel S Kehl, Thomas P Reber, Stefanie Liebe, Jakob H Macke, Florian Mormann, and Laura Leal-Taixé. Decoding movie content from neuronal population activity in the human medial temporal lobe. *eLife*, August 2025. doi: 10.7554/elife.106758.1. URL http://dx.doi.org/10.7554/eLife.106758.1.

Maya Geva-Sagiv, Emily A. Mankin, Dawn Eliashiv, Shdema Epstein, Natalie Cherry, Guldamla Kalender, Natalia Tchemodanov, Yuval Nir, and Itzhak Fried. Augmenting hippocampal–prefrontal neuronal synchrony during sleep enhances memory consolidation in humans. *Nature Neuroscience*, 26(6):1100–1110, June 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01324-5. URL http://dx.doi.org/10.1038/s41593-023-01324-5.

Joshua I. Glaser, Ari S. Benjamin, Raeed H. Chowdhury, Matthew G. Perich, Lee E. Miller, and Konrad P. Kording. Machine learning for neural decoding. *eNeuro*, 7(4), 2020. doi: 10.1523/ENEURO.0506-19.2020. URL https://www.eneuro.org/content/7/4/ENEURO.0506-19.2020.

Igor Gridchyn, Philipp Schoenenberger, Joseph O'Neill, and Jozsef Csicsvari. Assembly-specific disruption of hippocampal replay leads to selective memory deficit. *Neuron*, 106(2):291–300.e6, 2020. doi: 10.1016/j.neuron.2020.01.021.

Tanuj Gulati, Dhakshin S. Ramanathan, Chelsea C. Wong, and Karunesh Ganguly. Reactivation of emergent task-related ensembles during slow-wave sleep after neuroprosthetic learning. *Nature Neuroscience*, 17(8):1107–1113, 2014. doi: 10.1038/nn.3759.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani. Neural decoding of visual imagery during sleep. *Science*, 340(6132):639–642, 2013. doi: 10.1126/science.1234330. URL https://www.science.org/doi/abs/10.1126/science.1234330.

Jin Jing, Wei Ge, Shenda Hong, Matheus B. Fernandes, Zhen Lin, Chao Yang, Shuo An, Aaron F. Struck, Aline Herlopian, Ioannis Karakis, Jonathan J. Halford, Ming C. Ng, Erin L. Johnson, Balu L. Appavu, Rani A. Sarkis, Ghada Osman, Peter W. Kaplan, Monica B. Dhakar, Lakshmi Arcot Jayagopal, Zubeda Sheikh, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17):e1750–e1762, 2023. doi: 10.1212/WNL.0000000000207127.

Marcel S. Kehl, Thomas P. Reber, Valeri Borger, Rainer Surges, Florian Mormann, and Bernhard P. Staresina. Sleep ripples drive single-neuron reactivation for human memory consolidation. *bioRxiv*, 2026. doi: 10.64898/2026.03.27.714528. Preprint, version 1, posted March 31, 2026.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etiene David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul B Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:229156320.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. June 2019.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, jul 2018. doi: 10.1088/1741-2552/aace8c. URL https://doi.org/10.1088/1741-2552/aace8c.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, December 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2773081. URL https://doi.org/10.1109/TPAMI.2017.2773081.

Hesheng Liu, Yigal Agam, Joseph R. Madsen, and Gabriel Kreiman. Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, 62(2):281–290, 2009. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2009.02.025. URL https://www.sciencedirect.com/science/article/pii/S0896627309001718.

Kourosh Maboudi, Bapun Giri, Hiroyuki Miyawaki, Caleb Kemere, and Kamran Diba. Retuning of hippocampal representations during sleep. *Nature*, 629(8012):630–638, 2024. doi: 10.1038/s41586-024-07397-x.

Joseph G. Makin, David A. Moses, and Edward F. Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4):575–582, Apr 2020. ISSN 1546-1726. doi: 10.1038/s41593-020-0608-8. URL https://doi.org/10.1038/s41593-020-0608-8.

Emily A. Mankin and Itzhak Fried. Modulation of human memory by deep brain stimulation of the entorhinal-hippocampal circuitry. *Neuron*, 106(2):218–235, April 2020. doi: 10.1016/j.neuron.2020.02.024.

Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, Aug 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06443-4. URL https://doi.org/10.1038/s41586-023-06443-4.

Zoltán Nádasdy, Hajime Hirase, András Czurkó, Jozsef Csicsvari, and György Buzsáki. Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, 19(21):9497–9507, 1999. doi: 10.1523/JNEUROSCI.19-21-09497.1999.

J. Niediek, T. P. Reber, M. Bausch, F. Schwimmbeck, H. Gast, V. A. Coenen, J. Boström, C. E. Elger, and F. Mormann. Episodic memory consolidation by reactivation of human concept neurons during sleep reflects contents, not sequence of events. 2026. doi: 10.64898/2026.01.10.698827.

Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VI*, pp. 639–658, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01230-4. doi: 10.1007/978-3-030-01231-1_39. URL https://doi.org/10.1007/978-3-030-01231-1_39.

Brian N. Pasley, Stephen V. David, Nima Mesgarani, Adeen Flinker, Shihab A. Shamma, Nathan E. Crone, Robert T. Knight, and Edward F. Chang. Reconstructing speech from human auditory cortex. *PLOS Biology*, 10(1):1–13, 01 2012. doi: 10.1371/journal.pbio.1001251. URL https://doi.org/10.1371/journal.pbio.1001251.

Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209, 2009. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2008.11.007. URL https://www.sciencedirect.com/science/article/pii/S1053811908012263. Mathematics in Brain Imaging.

Adrien Peyrache, Mehdi Khamassi, Karim Benchenane, Sidney I. Wiener, and Francesco P. Battaglia. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12(7):919–926, 2009. doi: 10.1038/nn.2337.

Sean M. Polyn, Vaidehi S. Natu, Jonathan D. Cohen, and Kenneth A. Norman. Category-specific cortical activity precedes retrieval during memory search. *Science*, 310(5756):1963–1966, 2005. doi: 10.1126/science.1117645. URL https://www.science.org/doi/abs/10.1126/science.1117645.

Rodrigo Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005. doi: 10.1038/nature03687.

Rodrigo Quian Quiroga, Leila Reddy, Christof Koch, and Itzhak Fried. Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of Neurophysiology*, 98(4):1997–2007, 2007. doi: 10.1152/jn.00125.2007.

Rodrigo Quian Quiroga, Alexander Kraskov, Christof Koch, and Itzhak Fried. Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, 19(15):1308–1313, 2009. doi: 10.1016/j.cub.2009.06.060.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Björn Rasch and Jan Born. About sleep's role in memory. *Physiological Reviews*, 93(2):681–766, 2013. doi: 10.1152/physrev.00032.2012. URL https://doi.org/10.1152/physrev.00032.2012. PMID: 23589831.

Daniel B. Rubin, Tommy Hosman, Jessica N. Kelemen, Anastasia Kapitonava, Francis R. Willett, Brian F. Coughlin, Eric Halgren, Eyal Y. Kimchi, Ziv M. Williams, John D. Simeral, Leigh R. Hochberg, and Sydney S. Cash. Learned motor patterns are replayed in human motor cortex during sleep. *Journal of Neuroscience*, 42(25):5007–5020, 2022. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2074-21.2022. URL https://www.jneurosci.org/content/42/25/5007.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL https://doi.org/10.1038/s41586-023-06031-6.

M. Schönauer, S. Alizadeh, H. Jamalabadi, A. Abraham, A. Pawlizki, and S. Gais. Decoding material-specific memory reprocessing during sleep in humans. *Nature Communications*, 8(1):15404, May 2017. ISSN 2041-1723. doi: 10.1038/ncomms15404. URL https://doi.org/10.1038/ncomms15404.

Fabian Schwimmbeck, Johannes Niediek, Valeri Borger, Rainer Surges, Eugen Trinka, Thomas Schreiner, and Florian Mormann. Sequential coupling of sleep oscillations enables concept-neuron reactivation and supports information flow across the human hippocampal-cortical circuit. *bioRxiv*, January 2026. doi: 10.64898/2026.01.15.699122. URL http://dx.doi.org/10.64898/2026.01.15.699122.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. ISSN 1558-0210. doi: 10.1109/TNSRE.2022.3230250.

Bernhard P. Staresina, Johannes Niediek, Valeri Borger, Rainer Surges, and Florian Mormann. How coupled slow oscillations, spindles and ripples coordinate neuronal processing and communication during human sleep. *Nature Neuroscience*, 26(8):1429–1437, 2023. doi: 10.1038/s41593-023-01381-w.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26:858–866, 2023. doi: 10.1038/s41593-023-01304-9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xmcYx_reUn6.

Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254, May 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03506-2. URL https://doi.org/10.1038/s41586-021-03506-2.

Matthew A. Wilson and Bruce L. McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994. doi: 10.1126/science.8036517.

Weizhen Xie, John H. Wittig, Julio I. Chapeton, Mostafa El-Kalliny, Samantha N. Jackson, Sara K. Inati, and Kareem A. Zaghloul. Neuronal sequences in population bursts encode information in human cortex. *Nature*, 635(8040):935–942, 2024. doi: 10.1038/s41586-024-08075-8.

Wannan Yang, Chen Sun, Roman Huszár, Thomas Hainmueller, Kirill Kiselev, and György Buzsáki. Selection of experience for memory by hippocampal sharp wave ripples. *Science*, 383(6690): 1478–1483, March 2024. ISSN 1095-9203. doi: 10.1126/science.adk8261. URL http://dx.doi.org/10.1126/science.adk8261.

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with Neural Data Transformers. *Neurons, Behavior, Data analysis, and Theory*, August 2021. ISSN 2690-2664. doi: 10.51628/001c.27358. URL https://nbdt.scholasticahq.com/article/27358-representation-learning-for-neural-population-activity-with-neural-data-transformers.

Yipeng Zhang, Zahra M. Aghajan, Matias Ison, Qiujing Lu, Hanlin Tang, Guldamla Kalender, Tonmoy Monsoor, Jie Zheng, Gabriel Kreiman, Vwani Roychowdhury, and Itzhak Fried. Decoding of human identity by computer vision and neuronal vision. *Scientific Reports*, 13:651, 2023. doi: 10.1038/s41598-022-26946-w.

Yipeng Zhang, Yuanyi Ding, Chenda Duan, Atsuro Daida, Hiroki Nariai, and Vwani Roychowdhury. Self-supervised distillation of legacy rule-based methods for enhanced eeg-based decision-making. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 477–486. Springer, 2025.

A Related Work

Representation Learning Under Distribution Shift. To address distribution shift (Ben-David et al., 2010; Koh et al., 2020), domain adaptation methods include adversarial domain-invariance (Ganin et al., 2016), invariant risk minimization (Ahuja et al., 2020), and distribution alignment via second-order moment matching (Sun & Saenko, 2016). Self-supervised contrastive methods such as SimCLR and CLIP (Chen et al., 2020; Radford et al., 2021) demonstrate that similarity-structured representations can generalize beyond supervised classification. A complementary line uses knowledge distillation to constrain a student toward a teacher's predictions or representations (Hinton et al., 2015), including methods that preserve previously learned behavior during additional training (Li & Hoiem, 2018). We compose these ingredients in a setting with no target-state labels: movie-to-sleep distribution alignment for cross-state transfer and recall-domain distillation against a frozen source-trained teacher to preserve concept structure.

Neural Decoding and Brain–Computer Interfaces. Recent work in neural decoding (Glaser et al., 2020) has applied deep learning models to EEG, ECoG, and intracortical spike recordings, including compact CNNs for EEG-based BCIs (Lawhern et al., 2018; Zhang et al., 2025; Duan et al., 2026), speech decoding from ECoG (Makin et al., 2020; Metzger et al., 2023), and high-performance intracortical communication BCIs (Willett et al., 2021). Transformer-based models have further been introduced for neural population spike data (Ye & Pandarinath, 2021) and intracranial field potentials (Wang et al., 2023). Downstream evaluations in these systems are typically framed as supervised decoding problems, with performance measured by standard classification or regression metrics.

A growing line of work explores representation learning for neural data, drawing inspiration from multimodal alignment methods such as CLIP (Radford et al., 2021). Contrastive objectives have been used to learn neural representations aligned with behavior, time, speech, or visual stimulus spaces (Schneider et al., 2023; Défossez et al., 2023; Scotti et al., 2023), enabling more flexible representations than purely supervised classifiers.

Despite these advances, most neural decoding studies evaluate models within the same domain as training, leaving open the question of whether learned representations remain valid under unlabeled domain shift.

Memory Reactivation During Sleep. Neuroscientific studies have provided evidence that memory traces are reactivated during sleep (Rasch & Born, 2013). In rodents, hippocampal place cells exhibit sequential replay of spatial trajectories during sleep, supporting theories of memory consolidation (Wilson & McNaughton, 1994; Nádasdy et al., 1999; Carr et al., 2011). In humans, evidence has been obtained using fMRI and intracranial recordings, though with limited temporal resolution or content specificity (Horikawa et al., 2013; Xie et al., 2024; Schwimmbeck et al., 2026; Kehl et al., 2026) .

Intracranial studies have identified concept-selective neurons, which respond selectively to specific individuals or abstract concepts (Quian Quiroga et al., 2005; 2009; Bausch et al., 2021). These neurons have been observed to reactivate during both waking recall and sleep, providing rare content-specific neural references at the single-neuron level (Gelbard-Sagiv et al., 2008; Schwimmbeck et al., 2026; Kehl et al., 2026; Niediek et al., 2026). However, existing work primarily focuses on individual neurons and does not address whether population-level representations learned during waking experience are preserved during sleep.

Label-Free Evaluation. Approaches to evaluation without labels include label-free proxies for representation quality, clustering consistency, and downstream task transfer (Garrido et al., 2023; Caron et al., 2018; Kolesnikov et al., 2019). More recently, auxiliary modalities, such as sensor feedback or cross-modal signals, have been used to provide indirect supervision (Owens & Efros, 2018; Schneider et al., 2023; Défossez et al., 2023).

Our work contributes to this line of research by introducing BECA, a framework that leverages event-conditioned bidirectional coupling with an auxiliary signal to evaluate representations in unlabeled domains. Unlike standard metrics, BECA explicitly tests whether a representation preserves a latent variable that is observable through an independent measurement channel.

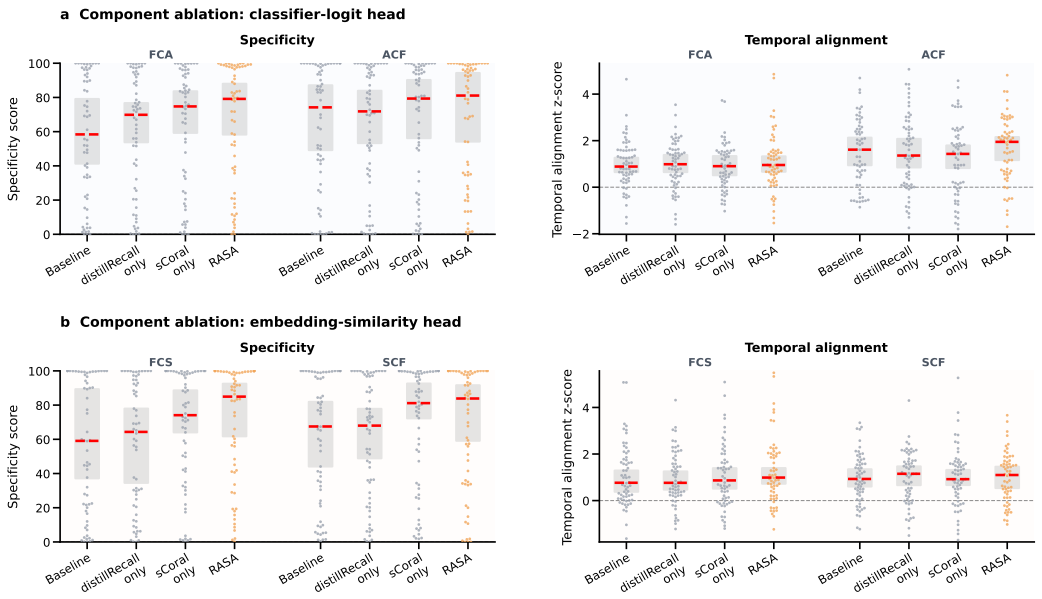


Figure 4: **Component ablation of RASA across BECA readouts. (a)** Classifier-logit head evaluation. Specificity and temporal-alignment distributions are shown for firing-conditioned activation (FCA) and activation-conditioned firing (ACF) across four adaptation configurations: movie-trained baseline, distillRecall only, sCoral only, and full RASA. **(b)** Embedding-similarity head evaluation. The same component comparison is shown for firing-conditioned similarity (FCS) and similarity-conditioned firing (SCF). Points denote individual held-out concept-selective neurons, red dashed lines denote medians, and gray bands denote 95% bootstrap confidence intervals of the median. RASA combines sleep-domain alignment with the recall-guided semantic safeguard.

Table 4: **BECA specificity and temporal alignment across adaptation stages.** Values are median ± SEM across cells for each BECA readout. Specificity values are percentile scores; temporal alignment values are center-versus-flank z_{peak} scores. FCA, firing-conditioned activation; ACF, activation-conditioned firing; FCS, firing-conditioned similarity; SCF, similarity-conditioned firing.

	Specificity			Temporal alignment		
BECA readout	Baseline	+sCoral	RASA	Baseline	+sCoral	RASA
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**

B Additional Results and Ablations

B.1 RASA Population Distributions and Across-Stage Visualizations

Figure 5 shows the unit-level RASA specificity and temporal-alignment distributions across BECA readouts. Figure 4 shows the across-stage progression from the movie-trained baseline to +sCoral and RASA, with per-readout statistics in Table 4.

B.2 Dual-Head Architecture: Rationale and Ablations

We compare classifier-only, embedding-only, and dual-head models trained on movie data to assess the contribution of each objective.

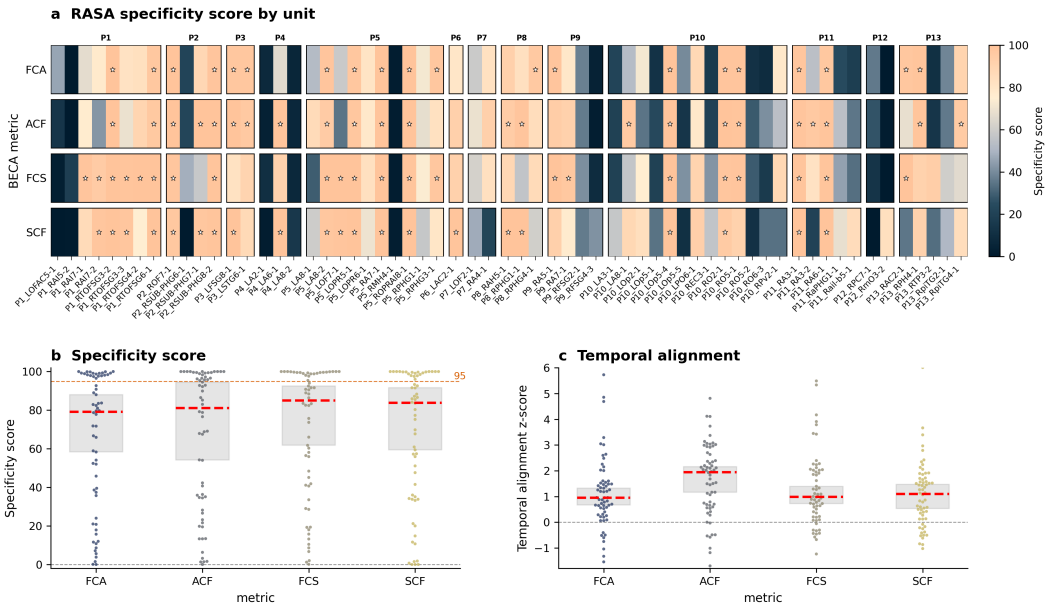


Figure 5: **Population-level BECA evaluation under RASA. (a)** Unit-level heatmap of RASA specificity scores across the four BECA readouts. Columns denote held-out concept-selective neurons, grouped by patient, electrode localization, and unit number. Stars mark unit-readout pairs exceeding the 95th percentile of the circular-shift surrogate distribution. **(b)** Unit-level distributions of RASA specificity scores across BECA readouts. **(c)** Unit-level distributions of temporal-alignment z_{peak} scores across the same readouts. In panels b and c, points denote individual units, red dashed lines denote medians, and gray bands denote 95% bootstrap confidence intervals of the median. The temporal-alignment axis in panel c is shown on a robust display range to reduce the influence of extreme outliers; summary statistics are computed from unclipped values. FCA, firing-conditioned activation; ACF, activation-conditioned firing; FCS, firing-conditioned similarity; SCF, similarity-conditioned firing.

Table 5: **Dual-head architecture benchmark.** Models are evaluated with median specificity score and median temporal-alignment z-score under BECA. Values are reported as median $\pm$ SEM across pooled unit-metric observations. Classifier-activation models are evaluated with firing-conditioned activation (FCA) and activation-conditioned firing (ACF); embedding-similarity models are evaluated with firing-conditioned similarity (FCS) and similarity-conditioned firing (SCF). Bold indicates the best model within each benchmark.

Benchmark	Model	Specificity	Temporal alignment
Classifier activation	EEG Conformer	54.60 ± 3.71	0.10 ± 0.20
	Ours, classifier-head only	56.10 ± 3.33	0.94 ± 0.11
	Ours, full dual-head (movie-only)	$\mathbf{68.40 \pm 3.31}$	$\mathbf{1.08 \pm 0.13}$
Embedding similarity	CEBRA	52.10 ± 3.15	-0.22 ± 0.14
	Ours, embedding-head only	53.45 ± 3.36	$\mathbf{1.07 \pm 0.11}$
	Ours, full dual-head (movie-only)	$\mathbf{62.25 \pm 3.37}$	0.86 ± 0.11

Table 5 summarizes matched classifier-activation and embedding-similarity benchmarks under the same spike-input setting. In the classifier-activation benchmark, the full dual-head model outperforms both EEG Conformer and the classifier-only ablation on both specificity and temporal alignment. In the embedding-similarity benchmark, the full model achieves the highest specificity, while the embedding-only ablation yields the highest temporal alignment. Thus, joint BCE+contrastive training gives the strongest overall specificity, while the embedding branch alone is sufficient to produce sharply event-locked similarity structure.

Table 6: **Recall-domain representation preservation.** Higher values indicate stronger agreement with the frozen movie-trained teacher.

Model	Teacher consistency	Recall-geometry agreement
Movie-trained baseline	1.000	1.000
+sCoral	0.773	0.766
+sCoral + rCoral	0.703	0.737
RASA	**0.949**	**0.914**

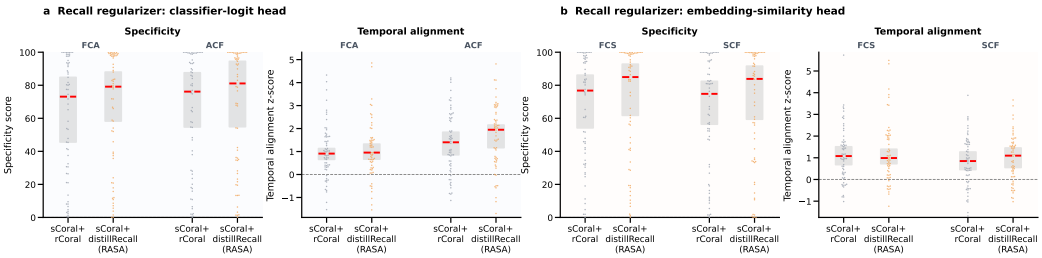


Figure 6: **Recall distillation is a more effective safeguard than recall-domain CORAL during sleep adaptation.** **(a)** Classifier-logit head BECA distributions comparing sCoral+rCoral with sCoral+distillRecall (RASA). Specificity is shown for FCA and ACF, and temporal alignment is shown for the same two readouts. **(b)** Embedding-similarity head BECA distributions for the same comparison. Specificity is shown for FCS and SCF, and temporal alignment is shown for the same two readouts. Points denote individual units, red dashed lines denote medians, and gray bands denote 95% bootstrap confidence intervals of the median. FCA, firing-conditioned activation; ACF, activation-conditioned firing; FCS, firing-conditioned similarity; SCF, similarity-conditioned firing.

B.3 Recall Safeguard Choice

We compare two ways of using recall during sleep-domain adaptation, both built on sleep-domain alignment with sCoral. The first uses recall-domain CORAL (sCoral+rCoral), which adds a second distribution-matching objective on recall embeddings. The second uses the RASA safeguard (sCoral+distillRecall), which distills recall embeddings against a frozen movie-trained teacher and therefore preserves the movie-trained semantic geometry during sleep alignment.

To quantify recall-domain preservation directly, we compute teacher consistency as the cosine agreement between recall embeddings produced by each adapted model and by the frozen movie-trained teacher on the same recall bins, averaged across the evaluated recall bins and cells. We additionally compute recall-geometry agreement using an RSA-style comparison of cross-concept embedding geometry between each adapted model and the frozen teacher on concept-relevant recall bins. Sleep alignment alone yields a teacher-consistency score of 0.773 and recall-geometry agreement of 0.766. RASA increases these values to 0.949 and 0.914, respectively, while +sCoral+rCoral yields 0.703 and 0.737. This ordering is consistent with recall distillation preserving the movie-trained representation during sleep alignment.

Figure 6 shows the unit-level BECA distributions for these two choices. Recall-domain CORAL underperforms when stacked on sCoral, reducing mean specificity from 77.34 to 75.21. In contrast, the distillation-based recall safeguard increases mean specificity to 82.26 and yields comparable or stronger temporal alignment across classifier-logit and embedding-similarity readouts. This contrast supports the central RASA design choice: when sleep alignment already operates at the distribution level, recall is more effective as a semantic consistency constraint than as an additional distribution-level alignment term.

Table 7: **Concept-identity negative control for RASA.** Classifier summaries pool FCA and ACF; embedding summaries pool FCS and SCF; all-readout summaries pool all four directions. **A:** Matched medians are compared with 5,000 within-movie concept-identity permutations. Permutation-null medians summarize the resulting null distributions. Holm correction covers only the two all-readout outcomes; classifier and embedding tests are descriptive and unadjusted. Dashes indicate no adjustment. **B:** For each unit–readout pair, the paired contrast is the matched score minus the mean score across mismatched concepts. Δ is the median of these contrasts, not the difference between the matched and permutation-null medians in panel A. Confidence intervals use 5,000 participant-clustered bootstrap resamples.

	Specificity			**Temporal alignment**		
	Classifier	Embedding	**All readouts**	Classifier	Embedding	**All readouts**
A. Concept-identity permutation test						
Matched median	79.40	83.85	82.48	1.334	0.988	1.174
Permutation-null median	52.00	57.08	54.05	0.801	0.774	0.792
Raw p	0.00780	0.00080	0.00140	0.00120	0.05540	0.00280
Holm-adjusted p	–	–	**0.00280**	–	–	**0.00280**
B. Matched–mismatched contrast						
Median paired contrast, Δ	32.08	23.62	28.65	0.550	0.340	0.464
95% CI	[12.11, 39.19]	[11.71, 36.33]	[11.91, 37.36]	[0.368, 0.809]	[0.126, 0.521]	[0.305, 0.591]

B.4 Concept-Identity Negative Control

Methods. We tested whether sleep-time BECA alignment depends on decoded concept identity in the full cohort of 62 held-out units from 13 participants. For each neuron i with movie-defined preferred concept k_i, we paired the same firing activity $a_i(t)$ with the classifier and embedding outputs for every concept in its movie-specific vocabulary. Matched (k_i) and mismatched ($j \neq k_i$) pairings were evaluated with identical settings in all four BECA directions; the null/background anchor was excluded.

We performed 5,000 cohort-level randomizations, independently permuting concept identity within each movie vocabulary and combining the assignments across movies into a cohort statistic. Concepts could remain fixed by chance. This control preserved the decoder outputs, neuronal activity, vocabulary sizes, and temporal structure while randomizing concept correspondence. One-sided randomization tests for pooled specificity and temporal alignment were Holm-corrected as two primary outcomes. Separately, we computed the median matched-minus-mean-mismatched contrast across unit-readout observations, with 95% confidence intervals from 5,000 participant-clustered bootstrap samples.

Results. Matched concept mappings yielded higher pooled specificity (82.48 versus a permutation-null median of 54.05) and temporal alignment (1.174 versus 0.792), with Holm-adjusted $p = 0.00280$ for both outcomes (Table 7; Figure 7). The matched-minus-mean-mismatched contrasts were 28.65 [95% CI: 11.91, 37.36] and 0.464 [0.305, 0.591], respectively. The contrasts were positive in 10 of 13 participants for specificity and 11 of 13 for temporal alignment after pooling readouts within participant. Participant-clustered intervals also excluded zero in each classifier/embedding summary, although the descriptive embedding temporal-alignment permutation test did not reach $p < 0.05$ (raw $p = 0.0554$). These analyses support concept-identity-specific alignment under the tested control, consistent with reinstatement, but do not provide behavioral ground truth for subjective sleep content.

B.5 Event-Threshold Sensitivity

Methods. We varied the BECA event-selection threshold from the 85th to the 95th percentile in one-percentile increments around the prespecified 90th percentile. Trained RASA models and sleep outputs were fixed; no model was retrained. FCA and FCS thresholds were computed over nonzero held-out-unit firing bins during NREM, whereas ACF and SCF thresholds used all NREM bins of the

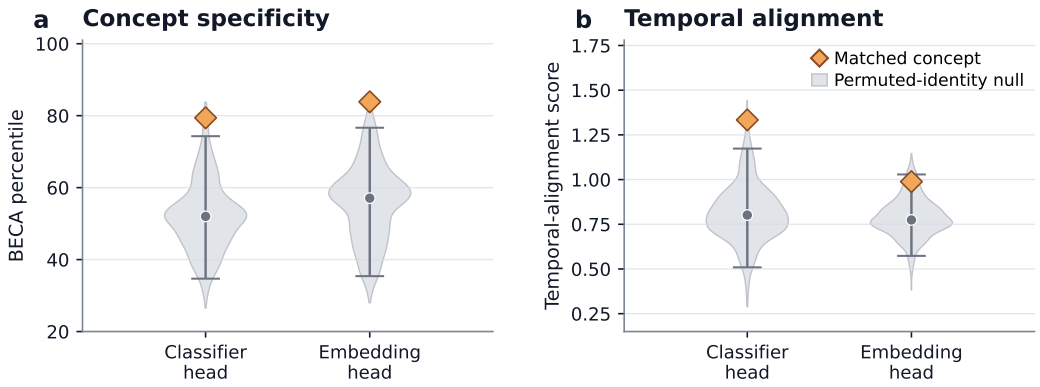


Figure 7: **Concept-identity negative control. (a,b)** Specificity and temporal alignment. Gray violins show null distributions from 5,000 cohort-level within-movie concept-identity permutations; vertical lines mark the 2.5th–97.5th percentiles, gray circles the null medians, and orange diamonds the matched RASA medians. Classifier-head summaries pool FCA and ACF; embedding-head summaries pool FCS and SCF. Matched and mismatched concepts use identical BECA settings.

corresponding model signal. At each threshold, we recorded event counts, the fraction of all NREM bins selected, specificity, and temporal alignment. We also identified adjacent thresholds selecting identical spike-conditioned event sets because of discrete spike counts. Pointwise 95% confidence intervals used 10,000 bootstrap resamples of the 13 participants, retaining all observations within each sampled participant.

The sweep retained specificity estimates for all 248 unit-readout observations at every threshold. Of 1,000 requested circular shifts per observation, at least 840 were valid at the 85th percentile, and all 1,000 were valid from the 90th percentile onward under the event-exclusion and boundary rules. Lower percentiles were not included because the analysis targets high-activity bins and denser event sets can leave no valid nonoverlapping surrogates for some observations.

Results. Pooled median specificity ranged from 80.30 to 82.73 and temporal alignment from 1.059 to 1.202 (Table 8; Figure 8); all participant-clustered temporal-alignment intervals excluded zero. Across individual readouts and thresholds, specificity medians ranged from 77.25 to 85.82. Model-conditioned event fractions decreased from approximately 15% to 5% of NREM bins, whereas spike-conditioned fractions decreased nonlinearly from a cohort median of 7.67% to 2.37%. Between adjacent thresholds, 74.2%–93.5% of units retained identical spike-conditioned event sets. Thus, the BECA results were robust within the tested high-activity threshold range, with discrete firing counts limiting how often spike-conditioned event sets changed.

C Data and Implementation Details

C.1 Participants

13 participants with drug-resistant epilepsy (5 female, age range 21-55 years) implanted with intracranial depth electrodes for seizure monitoring participated in the study. The research protocol was approved by the Institutional Review Board (IRB) of both institutions and all participants provided written informed consent prior to participation.

C.1.1 Data Collection

Participants were implanted with Behnke-Fried hybrid depth electrodes (AdTech Medical) with placement determined solely by clinical criteria. Electrophysiological data were recorded using the Neuralynx ATLAS system. Intracranial EEG (iEEG; 0.1 - 500 Hz; 2 kHz sample rate) signals were recorded from macroelectrode contacts (1.5 mm wide) spaced along the probe shaft. Broadband potentials (0.1 Hz - 8 kHz; 32 kHz sample rate) were recorded from bundles of nine 40 μm Platinum-

Table 8: **Event-threshold sensitivity for RASA.** Event columns give cohort median counts and median percentages of all NREM bins. Pooled medians include four BECA readouts for each of 62 held-out units (248 observations) at every threshold. Bold marks the prespecified threshold.

Event percentile	Held-out firing events	Classifier-score events	Embedding-similarity events	Temporal alignment	Specificity
85	5,388 (7.67%)	11,592 (15.00%)	11,600 (15.02%)	1.173	82.73
86	5,236 (7.51%)	10,820 (14.00%)	10,822 (14.02%)	1.136	81.65
87	5,054 (6.78%)	10,047 (13.00%)	10,050 (13.01%)	1.121	80.30
88	4,522 (6.10%)	9,274 (12.00%)	9,277 (12.02%)	1.126	81.35
89	4,055 (5.88%)	8,501 (11.00%)	8,506 (11.02%)	1.124	82.40
90	3,330 (4.88%)	7,728 (10.00%)	7,730 (10.02%)	1.174	82.48
91	2,850 (4.36%)	6,956 (9.00%)	6,963 (9.02%)	1.178	81.20
92	2,312 (3.47%)	6,183 (8.00%)	6,184 (8.01%)	1.202	81.05
93	2,152 (3.03%)	5,410 (7.00%)	5,412 (7.01%)	1.141	81.05
94	1,884 (2.73%)	4,637 (6.00%)	4,642 (6.01%)	1.076	80.85
95	1,676 (2.37%)	3,864 (5.00%)	3,869 (5.01%)	1.059	80.65

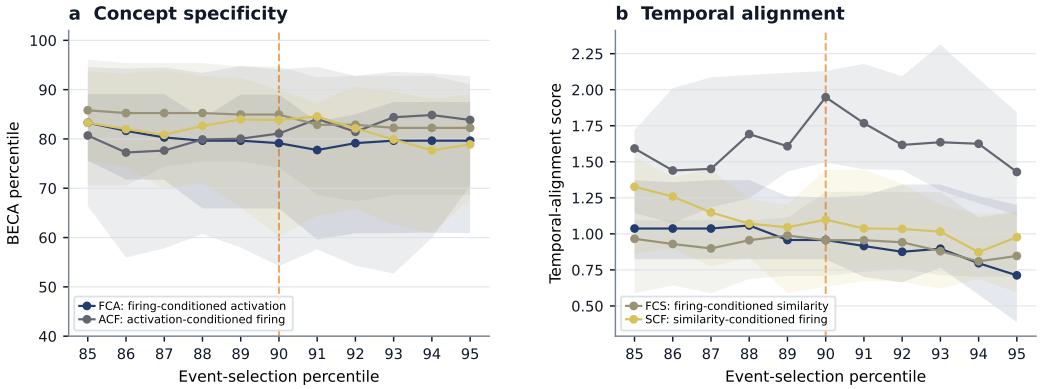


Figure 8: **Event-threshold sensitivity. (a,b)** Directional cohort medians for specificity and temporal alignment from the 85th to the 95th percentile. Shading denotes pointwise participant-clustered 95% confidence intervals from 10,000 bootstrap resamples. Specificity was defined for all four readouts of all 62 held-out units at every threshold. The dashed orange line marks the prespecified 90th percentile.

Iridium microwires (8 high-impedance recording electrodes and 1 low-impedance reference wire) that were threaded through the depth electrodes.

Across the 13 participants, the analysis included microwire recordings from the hippocampus, amygdala, entorhinal cortex, and parahippocampal cortex; frontal, pars, and cingulate regions; fusiform gyrus, insular cortex, lateral temporal cortex, visual cortex, and precuneus; and, in a small number of participants, the thalamus and basal ganglia. Coverage differed across participants because electrode placement was determined solely by clinical criteria. We retained all eligible non-held-out microwire channels and allowed the hierarchical encoder to learn their contributions, rather than imposing an MTL-only anatomical filter.

C.2 Data Preprocessing

Spike sorting, unit curation, time-resolved concept annotation and unit selectivity analyses follow the pipeline of Ding et al. (2025). Briefly, spikes were detected and sorted with WaveClus3 Chaure et al. (2018), manually curated, and units firing at < 0.5 spk/s during video viewing were excluded, yielding 452 units across 13 participants. Concept-presence annotations were manually curated over time and represented as multi-label vectors on the fixed, non-overlapping 250-ms grid aligned with the neural input windows. Source annotations were mapped or, where required, upsampled to this grid. Multiple concepts could co-occur within a window, while windows containing none of the predefined concepts received an all-zero label. The stimuli were 30–45-min naturalistic audiovisual segments

Table 9: **Movie-specific concept vocabularies.** Each participant-specific decoder used the concept vocabulary associated with the viewed stimulus. The four vocabularies comprised 25 distinct concepts in total.

Stimulus	K	Annotated concepts
24	8	White House, CIA, Sacrifice, Handcuff, Jack, Bill, Fayed, Amar
House	6	House, Cuddy, Fran, Wilson, Chase, Cameron
Twilight (2008)	6	Bella Swan, Edward Cullen, Charlie Swan, Jessica Stanley, Angela Weber, Jacob Black
The Mandalorian	5	Cara Dune, Grogu, Mandalorian, Omera, Winta

drawn from 24, House M.D., Twilight (2008), and The Mandalorian. Each participant-specific decoder used the concept vocabulary associated with the viewed stimulus, comprising 5–8 concepts per decoder and 25 distinct annotated concepts across the four stimuli (Table 9).

We restricted selectivity analyses to events in which a single concept appeared onscreen. Unit selectivity was assessed by comparing firing rates in the 1 s pre-onset and 1 s post-onset windows using a paired t-statistic tested against a sign-flip permutation null distribution. Units were classified as concept-selective if they fired at least one spike in $> 30\%$ of post-onset windows, exceeded the permutation threshold of $p < 0.01$ one-tailed, and responded significantly to only one concept. Overall, 62 of the 452 curated units (13.7%) met these criteria and entered the final evaluation cohort.

Sleep scoring follows the automated procedure of Geva-Sagiv et al. (2023). Briefly, NREM epochs were identified in a clean intracranial EEG channel by estimating slow-wave and spindle-band power in 30 s windows and clustering with a two-component Gaussian mixture model to separate NREM from desynchronized REM/wake epochs.

The underlying commercial videos and clinical recordings cannot be publicly redistributed because of copyright and patient-privacy constraints; the annotation and preprocessing procedures are documented here to support reproducibility.

C.3 Architecture Details

Each input is a 250-ms clusterless spike-amplitude window represented as $x \in \mathbb{R}^{2\times W\times T}$, where the two channels correspond to positive- and negative-polarity spike components. We use $T = 50$ temporal samples per 250-ms window. The number of recording wires W varies across patients from 40 to 96. Because both W and anatomical coverage varied by patient, we instantiate one encoder per patient while keeping all architecture hyperparameters fixed.

The encoder uses a hierarchical CSA–CCA design. Wires are partitioned into anatomical 8-wire microwire bundles. Channel-specific attention (CSA) encodes temporal patches within each bundle using convolutional patch embedding, learned ViT-style positional embeddings, and self-attention. Cross-channel attention (CCA) then integrates the resulting bundle-level CLS tokens into a global CLS representation. The classifier head maps this representation to K concept logits, while the embedding head projects it to a normalized d-dimensional space and compares it by cosine similarity to learned concept embeddings and a learned null anchor. Numerical architecture settings are listed in Table 10. Figure 9 illustrates how concept-contrastive learning organizes neural embeddings around learned concept anchors.

C.4 RASA Training Details

The source objective is used throughout training and combines multi-label BCE with contrastive concept alignment using weights $\lambda_{\mathrm{BCE}} = 1.0$ and $\lambda_{\mathrm{ctr}} = 0.5$. Concept embeddings and the null anchor are Xavier-initialized and learned jointly with the encoder. The contrastive temperature is fixed at $\tau = 0.07$.

Source-domain batches are constructed with concept-balanced sampling. For each mini-batch, samples are drawn with replacement from each concept's positive pool, and the remaining samples are drawn from time bins with no active annotated concept. For sessions with a strongly dominant

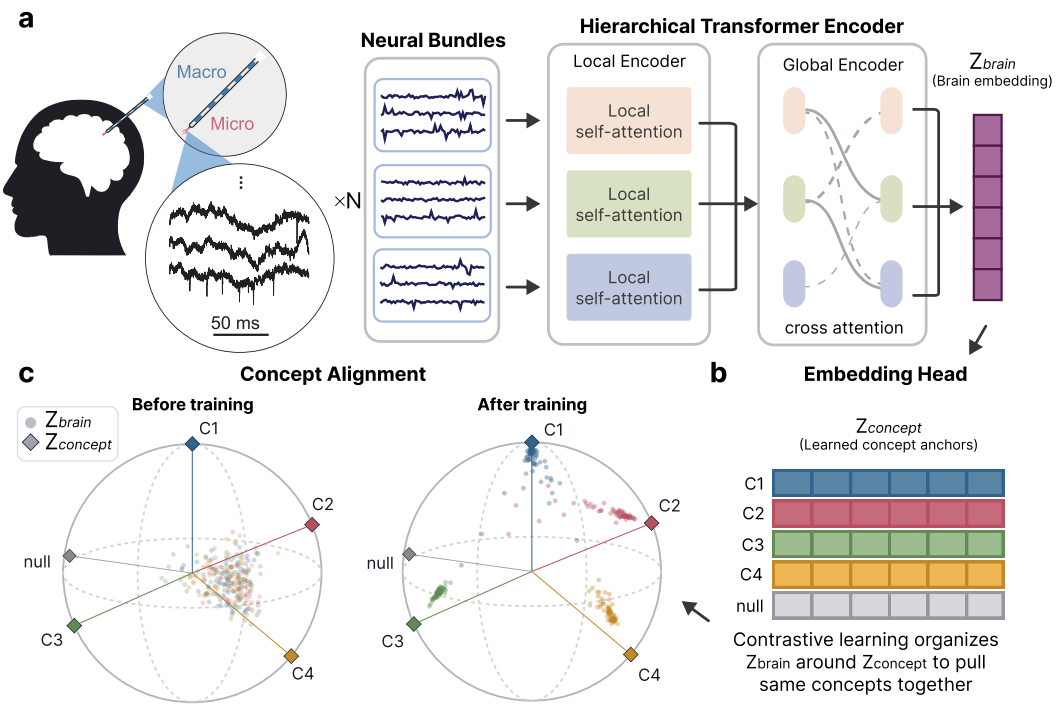


Figure 9: **Concept-contrastive learning organizes neural embeddings around learned concept anchors. (a)** Neural population signals are encoded by the hierarchical transformer, following the input and encoder pathway illustrated in Figure 1a–b. **(b)** The shared encoder produces Z_{brain}, a neural embedding that summarizes population activity within a time window. A concept-contrastive embedding head jointly learns concept anchors Z_{concept} in the neural embedding space. The null anchor is a negative target for windows with annotated concepts and the positive target for windows with no annotated concept. **(c)** Each annotated movie concept is represented by a learned concept anchor, denoted z_{concept}. The contrastive learning objective pulls z_{brain} closer to the z_{concept} anchors for concepts present in the current movie window and pushes it away from anchors for absent concepts. After training, the embedding similarity between z_{brain} and each z_{concept} provides a model-derived, concept-specific readout of neural activity. The before- and after-training arrangements are schematic.

Table 10: Architecture hyperparameters used in the main experiments.

Hyperparameter	Value
Input window (ms)	250
Temporal samples per window (T)	50
Effective sampling rate (Hz)	200
Recording wires per patient (W)	40–96
Wires per CSA bundle	8
CSA temporal patch	5 samples (25 ms)
CSA layers	5
CCA layers	5
Hidden size	384
Attention heads	6
Embedding dimension (d)	128

concept, the dominant concept label is down-weighted from 1.0 to 0.1 in co-occurring time bins, preventing the model from learning a constant-on prior.

For RASA, the source objective remains active while recall-guided embedding distillation and sleep-domain CORAL alignment are added after 30 source-only epochs. The recall teacher is an independently trained movie-only baseline model at epoch 50. Recall distillation matches student and teacher recall embeddings using a cosine-distance loss with weight $\lambda_{\mathrm{distill}} = 0.35$. Sleep CORAL uses an unlabeled 30-min NREM sleep segment and is linearly ramped over 20 epochs to $\lambda_{\mathrm{sCoral}} = 1.0$. The final RASA checkpoint is taken at epoch 70.

Table 11: Training hyperparameters used in the main experiments.

Hyperparameter	Value
Optimizer	Adam
Learning rate	10^{-4}
Weight decay	10^{-4}
Batch size	128
Samples per epoch	2048
Random seed	42
λ_{BCE}	1.0
λ_{ctr}	0.5
Contrastive temperature (τ)	0.07
Dominant-concept co-occurrence label	0.1
Movie-only teacher checkpoint	Epoch 50
RASA final checkpoint	Epoch 70
RASA adaptation start	After 30 source-only epochs
Recall distillation weight $\lambda_{\mathrm{distill}}$	0.35
Sleep CORAL segment	30 min NREM
Sleep CORAL final weight $\lambda_{\mathrm{sCoral}}$	1.0
Sleep CORAL ramp-up	20 epochs

C.5 BECA Evaluation Details

All BECA analyses are performed on NREM sleep time bins at 250-ms resolution. For each held-out concept-selective neuron and its associated concept k, we evaluate the corresponding classifier activation $\ell_k(t)$ and embedding similarity $s_k(t)$. Classifier activations are smoothed with a 1-s moving average before event detection and window extraction. Unit-conditioned events are defined as time bins in the top 10% of nonzero held-out unit firing within NREM sleep. Model-conditioned events are defined as time bins in the top 10% of the corresponding model-derived signal within NREM sleep. FCA and FCS use unit-firing events as the conditioning signal, while ACF uses classifier-activation events and SCF uses embedding-similarity events.

The 90th percentile identifies high-activity candidate events rather than imposing a biologically special cutoff. This unit- and readout-specific rule accommodates differences in firing rates and score scales while focusing on the strongest candidate events and retaining enough events for stable estimation. Because held-out-unit activity is discrete-valued spike count, nearby percentile cutoffs can yield the same selected event set. Unless otherwise specified, the event-selection threshold is the 90th percentile.

For each event, we extract a 2-s center window and two 2-s flanking windows located from 3 to 1 s before and from 1 to 3 s after the event. The evaluated signal $q(t)$ is $\ell_k(t)$ for FCA, $s_k(t)$ for FCS, and the held-out-unit firing activity $a_k(t)$ for ACF and SCF. We use $\mathrm{mean}_{\mathcal{W}}[q(t)]$ to denote the temporal mean of $q(t)$ within window $\mathcal{W}$; thus, $\mathrm{mean}_{\mathcal{W}_c(t_0)}[q(t)]$ and $\mathrm{mean}_{\mathcal{W}_f(t_0)}[q(t)]$ denote the center-window and combined-flank responses, respectively. The temporal-alignment statistic z_{peak} is computed as the paired center-versus-flank t-statistic across events, where each event contributes one center value and one flank value.

The specificity score is computed by comparing the real center-window response against a circular-shift surrogate distribution. In each surrogate, the event times are shifted along the NREM time axis by a random offset of at least 1 min, with shifted positions constrained not to coincide with any original event, preserving the number and temporal structure of events while disrupting alignment

between the held-out unit and model signal. The reported specificity score is the percentile rank of the real center-window response within this surrogate distribution; values above 95 correspond to stronger-than-surrogate event locking. Evaluation parameters are summarized in Table 12.

Table 12: BECA evaluation parameters.

Parameter	Value
Evaluation bin size	250 ms
Evaluation sampling rate	4 Hz
Unit-event threshold	90th percentile of nonzero NREM firing
Model-event threshold	90th percentile of NREM model signal
Center window	2 s
Left flank window	$[-3, -1]$ s
Right flank window	$[1, 3]$ s
Display window for plots	6 s
Surrogate scheme	Circular shift within NREM
Number of surrogates	1000
Minimum circular shift	1 min

C.6 Compute Resources

All experiments are conducted on a server equipped with 4 NVIDIA RTX A6000 GPUs, each with 48 GB of memory. Training time varies by subject recording length, typically about 5 minutes for the full pipeline for a single subject.

C.7 Implementation Details for Architecture Baselines

We compared the proposed dual-head encoder against external and internal baselines under a shared preprocessing and BECA evaluation protocol. All models used the same 250-ms neural windows, movie-domain concept annotations, participant-specific datasets, and held-out-cell exclusion procedure. For each evaluated concept-selective neuron, a separate participant-specific model was trained with its recording wire excluded from all inputs throughout training, adaptation where applicable, and sleep evaluation.

Because the baselines were designed for different output spaces, we used output-matched BECA readouts. Classifier-output models were evaluated with FCA and ACF, while embedding-output models were evaluated with FCS and SCF. Values in Tables 2 and 5 are median $\pm$ SEM across pooled unit-readout observations within each benchmark.

EEG Conformer. EEG Conformer was adapted as a classifier-style neural decoding baseline by replacing the final task head with a K-dimensional multi-label concept-logit head and training on movie-domain labels with binary cross-entropy.

EEGNet. EEGNet was adapted as a classifier-style baseline that operates directly on the structured clusterless spike-window tensor. Its final task head was replaced with a K-dimensional multi-label concept-logit head and trained with binary cross-entropy.

SPaRCNet. We implemented SPaRCNet following the published IIIC-SPaRCNet DenseNet-style temporal convolutional architecture, including its stem, dense blocks, transition layers, and classifier design. The output layer was adapted to K-dimensional multi-label concept prediction and trained with binary cross-entropy.

CEBRA. CEBRA was adapted as an embedding-style baseline using the same movie-domain neural windows and concept annotations. Its learned neural embeddings were used to produce concept-similarity traces for embedding-space BECA evaluation.

SIMPL. SIMPL was evaluated as an embedding-style baseline. Because SIMPL expects a time-by-neuron spike-count matrix, each 250-ms clusterless spike window was converted to one event-count

observation per retained wire, yielding a time-by-wire input matrix. The resulting latent trajectory was used to produce concept-similarity traces for embedding-space BECA evaluation.

Internal ablations. The classifier-head-only ablation retained the proposed encoder but used only the BCE classifier objective. The embedding-head-only ablation retained the proposed encoder but used only the contrastive embedding objective. The full dual-head baseline used both objectives jointly on movie data, without RASA adaptation.

D BROADER IMPACT

This work develops methods for transferring and evaluating concept-aligned neural representations across brain states when dense behavioral labels are unavailable. The primary potential benefit is scientific: such methods may help researchers study human memory consolidation, internally generated cognition, and sleep-related neural dynamics using independent neural reference signals. In the longer term, related approaches could inform clinical neurotechnology and biomedical sensing settings where dense labels are difficult to obtain.

The present study is foundational and uses invasive recordings collected from neurosurgical patients in controlled clinical research settings. Our analyses depend on patient-specific electrode coverage, annotated stimulus concepts, and rare held-out concept-selective neurons, so the framework is not a deployable system for general-purpose neural decoding or real-world monitoring of mental content.