RECONSTRUCTION WASTES RATE: SELECTIVE PREDICTION OVER CODE TARGETS FOR BIOSIGNAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete tokenizers for physiological signals are typically trained to reconstruct waveforms or spectral features. Reconstruction rewards encoding any component that lowers squared error, including noise that the surrounding context cannot predict, whereas a context model benefits from codes that the context can predict. We introduce **SPOT** (Selective Prediction Over Targets), which trains a context encoder to identify the quantized teacher code of a masked signal cell among candidates, without reconstructing the waveform or spectrum. On clinical EEG with injected noise, a probe recovers $11.9\times$ more of the noise from representations learned with reconstruction than from those learned with SPOT. We use SPOT to train **CORTEX**, a channel-agnostic encoder for EEG, EOG, EMG, ECG, and respiratory channels, and compare it with eleven baselines on eight clinical tasks spanning sleep and epilepsy. We show that SPOT is robust to increasing input noise and largely retains its clean-data performance under muscle-band contamination. A controlled comparison on an existing backbone, varying only the pretraining objective, attributes this robustness to the objective rather than the architecture. These results suggest that discriminative, prediction-based objectives yield representations that are more robust to noise than reconstructive ones, with direct implications for physiological signal encoders deployed in real-world, low-SNR clinical settings.

1 INTRODUCTION

Foundation models for EEG signals commonly rely on reconstruction-based objectives to learn generalizable representations. Particularly, the model is trained to reconstruct a masked portion of the signal (Wang et al., 2025; Zhou et al., 2025), or a latent representation thereof (Jiang et al., 2024; Barmpas et al., 2025; Cui et al., 2026; Chen et al., 2025), from the surrounding unmasked context. This masked-reconstruction objective encourages the model to infer structure from context, and has proven effective for learning transferable representations in other domains (Hsu et al., 2021; Bao et al., 2022). In this paper, we argue that this approach is not as effective for low-SNR EEG representation learning and propose an alternative recipe to train EEG foundation models.

EEG is a low signal-to-noise-ratio (SNR) signal: much of its variance is dominated by artifacts and noise rather than by event-related neural activity (Chang et al., 2020). Under a reconstruction objective, the model can minimize its loss simply by learning to predict this dominant noise component, without being pressured to capture the sparser event-related features that are of actual scientific interest.

We propose CORTEX-SPOT, an EEG foundation model that trains its tokenizer with a predictive objective instead of a reconstructive one. CORTEX (COntext-Robust Tokenizer for EEG and miXed biosignals) is the encoder, and SPOT (Selective Prediction Over Targets) is the objective used to train it. A context encoder reads the masked recording and selects the correct quantized code for each hidden cell from a set of candidates, rather than reconstructing the patch itself. Because the codebook is never scored on reconstruction, it is not rewarded for encoding the part of the signal that its neighbors cannot predict. This closes the gap described above between what current EEG tokenizers are trained to preserve and what the downstream context model needs.

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.

Existing EEG foundation models are trained with a reconstruction objective. LaBraM (Jiang et al., 2024) trains a vector-quantized tokenizer to reconstruct each patch's Fourier amplitude and phase, and uses the resulting codes as prediction targets for a masked-token model. Subsequent work keeps this reconstruction target and modifies the quantizer, using residual vector quantization (Barmpas et al., 2025), dual-domain residual quantization (Cui et al., 2026), or extending the same recipe across signal families (Chen et al., 2025). Other models skip the codebook and reconstruct masked patches directly in a continuous latent space (Wang et al., 2025; Zhou et al., 2025). A separate group of EEG foundation models use contrastive or hybrid objectives, or focus on extending subject and task coverage (Kostas et al., 2021; Wang et al., 2024; Yang et al., 2023a; El Ouahidi et al., 2025; Jiang et al., 2025); these are independent of the tokenizer objective we study here. In all of these approaches, the tokenizer is trained to reconstruct rather than to predict.

Predicting rather than reconstructing is an established approach outside biosignals. In speech, wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) predict discrete targets from context rather than regressing the waveform. In audio, BEATs (Chen et al., 2023) trains a tokenizer and a predictor together toward the same objective, and data2vec (Baevski et al., 2022) predicts contextualized teacher representations across modalities. In vision, I-JEPA (Assran et al., 2023) and V-JEPA 2 (Assran et al., 2025) predict latent targets for masked regions instead of pixels. To our knowledge, this approach has not been applied to EEG tokenization before. Appendix A discusses this literature in full, including the connection between SPOT and predictive information (Bialek et al., 2001) and the information bottleneck (Tishby et al., 1999).

Contributions. Our contributions are the following. **(i)** SPOT, a selective-prediction objective that trains CORTEX with no reconstruction term (§2.1). **(ii)** A result, under a Gaussian additive-noise model, showing that a reconstruction-optimal code allocates capacity to context-unpredictable directions while a predictive code does not (§3). **(iii)** An empirical measurement of this allocation, in a synthetic setting where the unpredictable subspace is known by construction and on real clinical recordings, showing the predicted separation in both cases (§4.2). **(iv)** Eight clinical evaluations against eleven baselines, showing task-dependent clean performance and improved robustness under contamination, with SPOT the most robust frozen encoder overall (§4.3–§4.4). **(v)** The same effect reproduced on LaBraM's published backbone, pretrained twice with only the objective changed, which rules out an architectural explanation (§4.5).

2 METHOD

2.1 SPOT: SELECTIVE PREDICTION OVER CODE TARGETS

Setup and principle. Let x be a one-second recording from one sensor, and let c be its context: the remaining sensor-time cells from the same recording. A tokenizer maps x to a discrete code z via $q(z \mid x)$. Two quantities describe this code: $\mathrm{I}(Z;X)$, the information it retains about the input cell, and $\mathrm{I}(Z;C)$, the information it retains about the context. Following the information-bottleneck principle (Tishby et al., 1999), SPOT aims to keep z context-predictive under a finite rate budget:

$$\max_{q} \ \mathrm{I}(Z;C) \quad \text{subject to} \quad \mathrm{I}(Z;X) \leq R. \tag{1}$$

We treat equation 1 as a guiding principle rather than a direct training objective: both mutual-information terms require joint distributions that are unavailable for learned representations. In practice, SPOT is motivated by this principle and uses a contrastive objective that rewards context-predictive codes, while a finite code space (produced by finite scalar quantization, FSQ) caps their capacity. §C.1 studies the resulting maximum predictive information, $P(R)$.

Where reconstruction spends its budget. At a fixed rate, an MSE reconstruction objective allocates capacity by input variance: whichever directions vary most get the most rate, regardless of whether that variance is physiological signal or noise. A predictive objective instead allocates capacity by contextual predictability. Under the Gaussian model of §3, reconstruction follows reverse water-filling (Cover & Thomas, 2006), so when movement, muscle, or electrode noise dominates the high-variance directions, reconstruction spends rate there even though prediction does not need to. Theorem 1 formalizes this, and §4.2 examines it empirically.

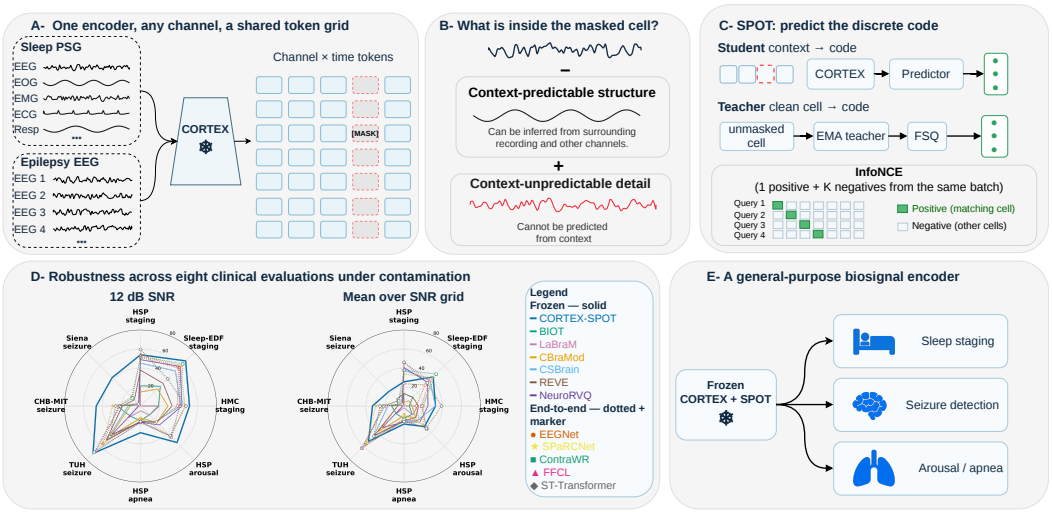


Figure 1: **Overview of CORTEX and SPOT. (A)** One shared encoder reads any sensor set as a channel × time grid of cells. **(B)** A masked cell mixes context-predictable structure with unpredictable detail. **(C)** SPOT trains the encoder to predict the discrete code of the hidden cell. **(D)** Robustness across eight evaluations under contamination. **(E)** The frozen encoder serves sleep, seizure and apnea tasks.

Training objective. SPOT is trained with InfoNCE (van den Oord et al., 2018) over discrete codes. A context encoder sees the masked recording; a separate, exponential-moving-average teacher sees the unmasked recording and encodes each hidden cell into a latent representation. A residual finite-scalar quantizer (FSQ) (Mentzer et al., 2024) then converts that latent into the target code z^+ (quantization happens only in the teacher branch). Given the context, the predictor identifies z^+ among $K+1$ candidates: the positive target and $K = 4096$ negative codes sampled from other masked cells:

$$\mathcal{L}_{\text{SPOT}} = -\,\mathbb{E}\left[\log \frac{\exp(\text{sim}(g(c), z^+)/\tau)}{\sum_{z' \in \{z^+\} \cup \mathcal{N}_K} \exp(\text{sim}(g(c), z')/\tau)}\right]. \tag{2}$$

This trains the encoder to recover the hidden cell's code from its context alone; the model never reconstructs a waveform or spectral value, which is what makes the objective selective rather than reconstructive.

Context and masking. Each recording is represented as a channel-by-time grid of one-second cells. Within each channel, we mask one contiguous, variance-weighted span, sampled independently across channels. Each channel retains visible cells, and prediction uses the unmasked sensor-time cells across the recording.

The $K = 4096$ negative codes are sampled in-batch from other masked cells. Because candidates can come from the same recording or subject as the target, they are not independent draws from the marginal distribution, so we treat InfoNCE as a contrastive surrogate rather than a certified mutual-information bound (Appendix D.3). Masking schedule and other training details are in Appendix F.

2.2 The CORTEX encoder

SPOT is not tied to a particular encoder architecture; it only requires a tokenizer mapping each signal cell to a latent representation, and §4.5 applies the same objective to LaBraM's published backbone. For our primary experiments, we pair SPOT with CORTEX (COntext-Robust Tokenizer for EEG and miXed biosignals), the tokenizer and encoder we use throughout this work. CORTEX is not architecturally novel: it combines design choices from prior EEG and biosignal tokenizers into one channel-agnostic encoder suited to recordings with heterogeneous sensor sets and montages.

CORTEX partitions a recording into *(sensor, 1-second)* cells, representing each as one token. A shared convolutional front end processes the raw waveform and its smoothed envelope into a D=192-dimensional embedding, and each token carries a sensor-identity embedding keyed to its

electrode or sensor name rather than its position in a channel list. A montage is thus a set of available sensor-time cells: no parameters are montage-specific, and a missing sensor simply drops its tokens rather than zero-filling a channel. A 6-layer, 6-head transformer integrates information across up to five minutes of context (full architecture in Appendix B). The encoder has $5{,}589{,}440$ parameters, third smallest of the seven pretrained backbones compared (Appendix I.1), letting one frozen model serve every heterogeneous evaluation in §4. CORTEX-SPOT is used throughout our primary experiments; §4.5 isolates the contribution of the SPOT objective from the encoder architecture.

3 Where Reconstruction and Prediction Allocate Rate

Under a finite rate budget, reconstruction and prediction allocate capacity differently: reconstruction can spend rate on variation the context cannot predict, while prediction need not. We capture this with ψ, the fraction of a code's information spent on context-uninformative directions. §4.2 tests this allocation with probe-based diagnostics, in both a controlled synthetic setting and real EEG. The model below is a simplified, Gaussian account of the allocation mechanism, not of the FSQ system itself: it does not claim reconstruction wastes rate at every SNR (that also depends on how nuisance variance is distributed, Remark 2), and it cannot separate noise that lies in the same direction as the signal (Remark 3). Full assumptions and proofs are in Appendices C and D.

Setup. Let $x \in \mathbb{R}^d$ be a signal cell, split into physiological signal s and noise n, with context $c \in \mathbb{R}^m$. We assume x, s, n, c are jointly Gaussian and zero-mean, with

$$x = s + n, \qquad s \perp n, \qquad n \perp c. \tag{3}$$

so only the physiological part of x correlates with the context. We also assume one orthonormal basis simultaneously diagonalizes $\mathrm{Cov}(s)$ and $\mathrm{Cov}(n)$, giving per-coordinate variances $\sigma^2_{s,i}, \sigma^2_{n,i}$. Define the *predictive-null* coordinates and their span:

$$\mathcal{I}_0 := \{i : \mathrm{Cov}(x_i, c) = 0\}, \qquad \mathcal{N} := \mathrm{span}\{e_i : i \in \mathcal{I}_0\}. \tag{4}$$

Writing $X_{\mathcal{N}}, X_{\mathcal{P}}$ for the projections of X onto $\mathcal{N}$ and $\mathcal{P} := \mathcal{N}^{\perp}$, this alignment gives $X_{\mathcal{N}} \perp (X_{\mathcal{P}}, C)$: the predictive-null directions carry no information about the context, by construction. A nontrivial separation into predictive and predictive-null subspaces requires $d > 1$ (Remark 4).

Because the coordinates are aligned, comparing the two objectives reduces to comparing independent scalar directions, exactly the setting where reverse water-filling is easiest to read (Cover & Thomas, 2006). We compare, under the same rate budget $\mathrm{I}(Z; X) \le R$: a reconstruction code Z_{rec}, which minimizes $\mathbb{E}\|x - \hat{x}\|^2$, and a predictive code Z_{pred}, which maximizes $\mathrm{I}(Z; C)$ (the predictive optimum need not use the full budget). For any code with $\mathrm{I}(Z; X) > 0$, define the predictive-null allocation

$$\psi(Z) := \frac{\mathrm{I}(Z; X_{\mathcal{N}} \mid X_{\mathcal{P}})}{\mathrm{I}(Z; X)}.$$

For zero-rate codes, we set $\psi(Z) = 0$ by convention. For the Gaussian reconstruction optimum, this works out to

$$\psi_{\mathrm{rec}}(R) = \frac{\sum_{i \in \mathcal{I}_0} r_i^{\mathrm{rec}}}{R}, \qquad r_i^{\mathrm{rec}} = \tfrac{1}{2} \log^+(\sigma^2_{x,i}/\theta),$$

with θ set by the rate budget (Cover & Thomas, 2006, Thm. 10.3.3): reconstruction allocates rate by total variance $\sigma^2_{x,i} = \sigma^2_{s,i} + \sigma^2_{n,i}$, with no way to tell signal variance from nuisance variance.

Theorem 1 (Reconstruction funds predictive-null directions; prediction need not)**.** *Under Assumption 1, fix a rate budget $R > 0$ and suppose $\dim \mathcal{N} > 0$. Then: **(i)** there is a predictive-optimal code with $q^*(z \mid x_{\mathcal{P}}, x_{\mathcal{N}}) = q^*(z \mid x_{\mathcal{P}})$ (equivalently $\mathrm{I}(Z_{\mathrm{pred}}; X_{\mathcal{N}} \mid X_{\mathcal{P}}) = 0$), so $\psi_{\mathrm{pred}}(R) = 0$; and **(ii)** the reconstruction optimum assigns positive rate to $\mathcal{N}$ whenever some predictive-null coordinate has $\sigma^2_{x,i} > \theta$. Under the condition in part (ii),*

$$\psi_{\mathrm{pred}}(R) = 0, \qquad \psi_{\mathrm{rec}}(R) > 0. \tag{5}$$

Part (ii) is reverse water-filling directly. Part (i) follows because $X_{\mathcal{N}} \perp (X_{\mathcal{P}}, C)$ means an encoder can be marginalized over $X_{\mathcal{N}}$ without changing its joint distribution with C: this removes any dependence on $X_{\mathcal{N}}$ and cannot increase $\mathrm{I}(Z; X)$. Corollary 1 shows $\psi_{\mathrm{rec}}(R) \to 1$ as nuisance variance comes to dominate. Note that optimality also gives $\mathrm{I}(Z_{\mathrm{pred}}; C) \ge \mathrm{I}(Z_{\mathrm{rec}}; C)$, but the point of this result is where each objective spends its rate, not just how much context information it recovers.

Remark 1 (Scope of the result)**.** Both halves of Theorem 1 build on standard results: part (ii) on reverse water-filling (Cover & Thomas, 2006), part (i) on the Gaussian information bottleneck and its canonical-correlation solution (Chechik et al., 2005). What the theorem adds is the comparison between the two allocation rules, which motivates the probe-based analysis in §4.2.

The theorem is an existence statement about the predictive *optimum*; it does not guarantee a trained encoder reaches $\psi \approx 0$ in practice. §4.2 checks whether learned representations show this pattern in probe-recoverable nuisance. Likewise, the theorem predicts a difference in capacity allocation, not downstream robustness directly; we test the resulting hypothesis (that allocating less capacity to context-unpredictable nuisance means less sensitivity to test-time contamination) in §4.4, over the SNR range where the effect appears.

4 EXPERIMENTS

4.1 SETUP

Pretraining. Our primary model, CORTEX-SPOT, is pretrained on a pooled unlabeled corpus combining Temple University Hospital EEG, including the seizure corpus (Obeid & Picone, 2016; Shah et al., 2018), with full polysomnography from the Human Sleep Project (HSP) (Brain Data Science Platform, 2020). For HSP and TUH/TUSZ, which also contribute downstream evaluations, all downstream test subjects are excluded from pretraining, ensuring that no downstream test subject is represented in the pretraining data. The resulting corpus includes EEG, EOG, EMG, ECG, and respiratory channels. For each masked cell, an EMA teacher quantizes the unmasked signal through a depth-stacked FSQ codebook, and the context encoder selects the corresponding discrete code among one positive target and 4096 negative codes. Table 5 in Appendix F provides the complete pretraining configuration. The objective-transfer experiment in §4.5 isolates the effect of SPOT by pretraining matched masked-reconstruction and SPOT models on both the CORTEX and LaBraM backbones.

Downstream tasks and baselines. We evaluate the learned representations on eight downstream evaluations spanning sleep medicine and epilepsy. The sleep evaluations comprise sleep staging, arousal detection, and apnea detection on HSP (Brain Data Science Platform, 2020), together with sleep staging on Sleep-EDF (Kemp et al., 2000) and HMC (Alvarez-Estevez & Rijsman, 2021). The epilepsy evaluations comprise seizure detection on TUH (Shah et al., 2018), CHB-MIT (Shoeb, 2009), and Siena (Detti et al., 2020). Positive-class prevalence in the seizure datasets ranges from 0.27% to 6.89%; we therefore report the corresponding trivial-performance floor alongside every score. Table 4 summarizes the evaluation units, metrics, trivial-performance floors, and evaluation-set sizes.

We compare against eleven baselines using identical subject-level splits. The first group comprises six published EEG foundation models evaluated with their released checkpoints and frozen probes: BIOT (Yang et al., 2023a), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025), REVE (El Ouahidi et al., 2025), and NeuroRVQ (Barmpas et al., 2025). The second group comprises five task-specific supervised models from the benchmark suite of Wang et al. (2025): EEGNet (Lawhern et al., 2018), SPaRCNet (Jing et al., 2023), ContraWR (Yang et al., 2023b), FFCL, and ST-Transformer (Yang et al., 2023a). Table 11 reports model parameter counts. Because these models are trained end-to-end for each downstream task, their clean and contaminated performance is not directly comparable to that of a frozen representation. Models that key on named electrodes read the bipolar corpora (Sleep-EDF, CHB-MIT) through a documented mapped-montage assignment (Appendix H.3), and Appendix E documents the input units, metrics, trivial floors, and any evaluation a released checkpoint cannot accommodate.

Evaluation under contamination. Throughout, *clean* denotes downstream performance on uncontaminated recordings. Table 1 reports clean performance, whereas Table 2 reports absolute performance under muscle-band contamination in each task's native metric. Each contaminated-result entry is reported as 12 **dB / grid mean**: the score at 12 dB and the mean score over the full -3 to 12 dB SNR grid. These are absolute scores, not values normalized by clean performance.

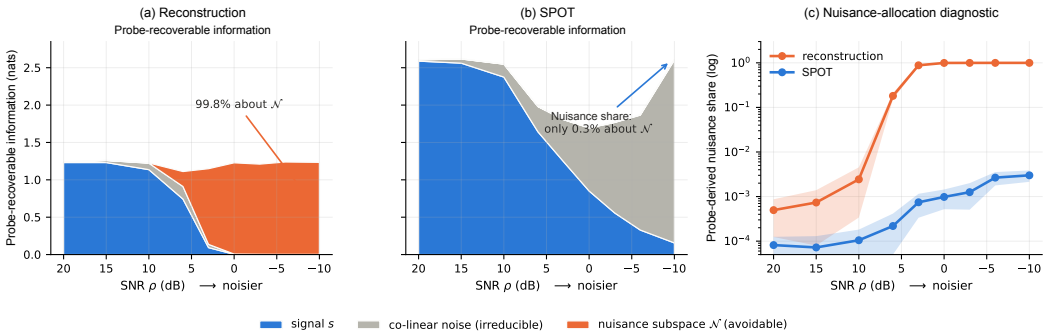


Figure 2: **Synthetic allocation test** (three seeds). **(a,b)** Probe-derived diagnostics for signal, co-linear noise, and the predictive-null subspace $\mathcal{N}$. **(c)** Nuisance-allocation diagnostic $\widehat{\psi}_{\text{probe}}$ across SNR.

We add noise only at test time and scale it to the power of each recording so that all models are evaluated at the same measured signal-to-noise ratio (SNR). Results are aggregated over three training seeds under a fixed noise schedule. We evaluate each baseline using its released input pipeline and evaluation unit; details of the input units, metrics, and unavailable evaluations are provided in Appendix E. Results for individual corpora and additional contaminant types are reported in Appendix H.

4.2 The mechanism, measured twice

Theorem 1 predicts that a reconstruction objective can allocate representation capacity to directions that are uninformative about the context, whereas a predictive objective need not do so. Because the predictive-null subspace is not identifiable in real recordings, we first test this mechanism in a synthetic setting where it is known by construction. The input has sixteen independent directions. Eight directions contain signal predictable from the context, each corrupted by co-linear noise at SNR ρ. The remaining eight directions contain no signal, are independent of the context, and have four times the power of the co-linear noise; these directions constitute $\mathcal{N}$.

Both models receive the same noisy input x through the same 12-bit bottleneck and differ only in their training objective: reconstruction minimizes MSE to x, whereas SPOT performs InfoNCE-based selection from the context. We train each model across three seeds and nine SNR levels. The theoretical quantity $\psi(Z)$ is not directly observable for a learned representation. Therefore, we measure the fraction of information about each known source group that is linearly recoverable from the frozen code. Specifically, we fit a ridge probe on one half of a newly generated sample and evaluate it on the other half, yielding $\widehat{\psi}_{\text{probe}}$. This quantity is an empirical diagnostic of nuisance allocation, not an estimate of $\psi(Z)$. The probe recovers $\mathcal{N}$ from the reconstruction code with a share of 0.998, showing that the probe detects substantial nuisance information in the reconstruction representation.

The probe diagnostics reveal different allocation patterns (Figure 2). At $-10\,\text{dB}$, the nuisance-allocation share $\widehat{\psi}_{\text{probe}}$ is 99.8% for reconstruction and 0.3% for SPOT. Under the same 12-bit budget, the signal-related probe diagnostic $\widehat{I}_{\text{probe}}(Z; s)$ is 0.001 nats for reconstruction and 0.156 nats for SPOT. Across the nine SNR levels, the nuisance-allocation diagnostic remains approximately 0.003 or lower for SPOT and approaches one for reconstruction at low SNR, with a seed spread below 0.001.

We report the two allocation shares rather than their ratio, since a ratio such as $333\times$ is dominated by the near-zero denominator. Importantly, SPOT is not prevented from encoding the high-variance directions in $\mathcal{N}$: they have four times the power of the co-linear noise and are available under the same 12-bit budget. Instead, the predictive objective learns to deprioritize them because they provide no information about the context. At $-10\,\text{dB}$, the co-linear-noise component remains the largest contribution to the SPOT code, as expected: because it shares the signal direction, it is irreducible by either objective (Remark 3).

Table 1: **Clean performance** (mean±sd over three seeds). **Bold** indicates the best result in each column among the frozen-probe models. The lower block contains end-to-end supervised models, whose clean performance is not directly comparable with that of frozen-probe models (§4.1).

	Staging			HSP		Seizure		
Model	HSP	Sleep-EDF	HMC	Arousal	Apnea	TUH	CHB-MIT	Siena
metric	κ	κ	κ	AUPRC	macro-F1	AUPRC	AUPRC	AUPRC
floor	0.0	0.0	0.0	11.1	15.7	6.89	0.27	0.82
CORTEX-SPOT	**55.3±1.1**	**68.6±0.9**	52.8±1.1	**57.6±0.5**	**41.8±0.3**	**71.4±1.2**	49.7±1.0	**46.3±3.8**
BIOT	49.0±1.1	62.1±0.5	36.4±2.0	29.5±0.4	16.1±0.4	26.0±0.5	**61.6±4.7**	9.0±1.1
LaBraM	52.1±0.7	61.7±1.0	52.8±0.4	27.1±0.0	15.4±0.9	53.7±0.3	7.3±0.1	11.2±1.8
CBraMod	51.2±0.4	60.8±1.3	52.1±1.7	32.6±0.1	16.2±0.3	47.2±0.9	10.6±3.2	3.5±1.3
CSBrain	46.7±1.1	59.1±1.9	45.1±0.5	24.5±0.1	15.6±0.4	51.8±0.3	14.0±0.2	3.3±0.3
REVE	47.2±0.5	67.2±0.8	**59.3±0.1**	31.8±0.1	17.6±0.4	55.7±0.1	28.0±0.6	9.2±0.5
NeuroRVQ	48.4±1.0	61.4±0.2	51.0±0.1	30.7±0.1	16.2±0.4	46.6±0.4	1.9±0.1	3.4±1.6
EEGNet	54.7±1.8	63.2±0.9	59.0±2.4	23.5±0.8	15.9±0.9	55.9±2.4	23.0±3.2	14.8±4.2
SPaRCNet	55.7±2.4	66.6±2.2	63.0±2.4	55.3±1.1	15.5±1.8	53.9±3.9	28.7±13.6	4.2±2.2
ContraWR	64.2±1.8	70.5±0.8	66.2±0.6	60.0±0.2	19.8±1.2	68.3±4.9	21.7±8.7	12.8±7.9
FFCL	63.0±0.7	68.9±2.1	64.5±1.7	58.6±0.6	19.6±0.4	66.3±2.2	23.7±5.5	1.5±0.7
ST-Transformer	60.8±1.9	63.1±1.2	62.9±0.4	46.1±4.5	16.2±0.4	46.1±5.0	25.2±5.5	3.6±0.8

The same signature on real EEG. We next examine whether the same allocation pattern appears in real EEG, where the predictive-null directions are not directly observed. For each TUH cell x, we obtain $x = \hat{s}+\hat{n}$ with a denoiser, encode it with both frozen encoders, and probe the nuisance-related information recoverable from each code, $\widehat{I}_{\mathrm{probe}}(Z;\hat{n})$. We treat this quantity as a comparative diagnostic rather than a calibrated estimate, since $\hat{n}$ is itself estimated. We therefore also inject known nuisance: wavelet-derived residuals at a controlled SNR and expert-labelled muscle artifact from the TUH Artifact Corpus (Buckwalter et al., 2021) at 10, 6, and 3 dB (Appendix H.1).

The recording's own residual, re-injected into the cell, gives the largest separation: 7.27 nats for reconstruction against 0.61 for SPOT, an $11.9\times$ difference in the diagnostic and $5.9\times$ in within-model nuisance share. Expert-labeled muscle artifact gives the same ordering at $10\,\mathrm{dB}$ (10.56 against 3.06 nats), and the denoiser residual gives a $2.25\times$ separation. Because the probe target is a power spectrum, which is also the reconstruction model's pretraining target, absolute nats can favor reconstruction. We therefore also report the within-model nuisance share. Both diagnostics favor SPOT except for the nuisance share at $3\,\mathrm{dB}$, where the reconstruction-to-SPOT ratio is 0.97. The ratio of absolute probe diagnostics narrows from 3.45 at $10\,\mathrm{dB}$ to 1.62 at $3\,\mathrm{dB}$, consistent with the reduced downstream advantage under severe contamination (§4.4). This is an empirical property of the learned encoders, not a consequence of Theorem 1, whose predictive-null subspace is not identified with these nuisances. Table 6 and Figure 4 in Appendix H report the full probe results. Figure 5 additionally reports code utilization and overlap between the marginal code distributions of sleep and epilepsy recordings.

4.3 Clean performance

Table 1 compares CORTEX-SPOT with published EEG foundation models and task-specific supervised baselines on uncontaminated recordings across eight downstream evaluations.

CORTEX-SPOT remains competitive on clean recordings. Among the frozen-probe models, CORTEX-SPOT achieves the highest clean score on six of the eight evaluations. Its two deficits are HMC staging, where REVE leads by 6.5 points, and CHB-MIT seizure detection, where BIOT leads by 11.9 points. The apnea result reflects channel coverage rather than the SPOT objective: CORTEX can process respiratory and saturation channels, whereas fixed-montage EEG baselines cannot. The CHB-MIT comparison is affected by differences in input-window duration, as discussed in Appendix H.3. NeuroRVQ's low CHB-MIT score reflects an out-of-domain, bipolar corpus read through a frozen probe, not a change of protocol: its pretraining covers TUSZ and Siena but not CHB-MIT, and none of the input-handling variants we tried on the development split changed it (Appendix H.3).

4.4 Performance under contamination

We evaluate robustness by adding muscle-band contamination at test time and measuring absolute downstream performance. Muscle artifact is a clinically common EEG nuisance and is difficult to

Table 2: **Absolute performance under muscle-band contamination**, reported as 12 **dB / whole-grid mean** in each task's native metric. **Bold** indicates the best result in each column among frozen-probe models, shown above the second rule; end-to-end supervised baselines, trained on each target task rather than read by a frozen probe, are shown below it and excluded from the bolding.

	Staging			HSP		Seizure		
Model	HSP	Sleep-EDF	HMC	Arousal	Apnea	TUH	CHB-MIT	Siena
CORTEX-SPOT	**53.9**/25.5	**67.2/43.2**	**51.6/33.0**	**54.4/30.5**	**28.1/19.0**	**70.6/52.6**	**46.6/33.0**	**42.2/22.0**
BIOT	21.2/13.2	29.4/12.7	19.0/ 5.5	13.2/12.2	15.6/13.2	15.1/11.6	24.5/13.1	2.2/ 1.2
LaBraM	20.5/ 8.1	0.0/ 0.0	19.9/13.8	13.5/13.3	5.1/ 1.8	17.4/11.6	2.7/ 1.2	2.8/ 1.8
CBraMod	14.7/ 4.1	25.3/ 7.5	28.8/ 8.4	24.4/17.5	12.6/10.5	45.3/37.8	10.8/ 6.7	2.7/ 2.0
CSBrain	45.1/37.9	52.1/38.4	42.1/33.0	23.0/20.8	16.0/15.3	49.9/45.1	12.9/16.8	1.7/ 3.5
REVE	37.7/13.0	33.7/13.1	33.3/15.2	24.4/17.4	16.6/15.2	51.2/44.0	19.0/ 6.6	9.2/10.5
NeuroRVQ	48.8/**40.5**	58.9/35.1	43.7/25.1	28.2/23.5	17.1/17.1	41.9/32.0	1.8/ 1.3	2.7/ 2.0
EEGNet	51.1/23.5	56.5/31.2	38.7/12.9	23.6/22.3	15.6/13.5	56.1/56.7	23.2/22.6	13.1/ 7.8
SPaRCNet	54.3/33.8	59.9/41.9	46.5/20.7	45.6/28.2	12.5/ 8.6	55.9/56.3	29.4/29.6	3.6/ 3.1
ContraWR	52.0/37.5	62.8/47.6	41.8/18.9	45.0/33.8	17.2/14.7	68.9/62.6	21.7/21.2	12.1/ 7.9
FFCL	53.6/46.0	57.6/37.8	41.4/18.1	46.0/33.2	17.3/15.6	65.9/62.7	21.3/19.9	1.6/ 1.7
ST-Transformer	59.6/46.0	40.5/34.4	47.8/39.3	44.0/33.2	16.3/16.3	45.8/44.8	26.2/25.2	3.2/ 3.0

predict from the surrounding context, making it the primary contaminant in the main analysis and the recorded artifact evaluated in §4.2. Theorem 1 does not rank different contaminants. Nevertheless, the mechanism suggests that the advantage of selective prediction should be smaller when the contaminant is itself more predictable. Appendix H provides full SNR curves (Figures 6, 7, and 8), pink- and white-noise results (Tables 8 and 9), retention comparisons across noise types (Table 7), and additional TUH results with recorded muscle artifacts (Table 10). Table 2 reports absolute performance under muscle-band contamination at $12\,\mathrm{dB}$ and averaged over the SNR grid.

CORTEX-SPOT achieves the strongest absolute performance under contamination among the frozen-probe models. At $12\,\mathrm{dB}$, it is the highest scoring frozen encoder on all eight evaluations. Over the full SNR grid, it remains highest in six evaluations and is tied on HMC staging at the reported precision. The exception is HSP staging, where NeuroRVQ achieves the highest grid mean (40.5 against 25.5 κ), despite CORTEX-SPOT leading by 5.1 κ at $12\,\mathrm{dB}$.

Some end-to-end supervised baselines achieve higher scores than CORTEX-SPOT on individual evaluations (Table 2). These models optimize all parameters for the target task, whereas CORTEX-SPOT trains only a $1{,}005$-parameter probe on a frozen encoder. Thus, CORTEX-SPOT is competitive with, rather than uniformly superior to, task-specific supervised training under contamination.

4.5 Objective transfer across two encoder backbones

The main experiments establish the performance of CORTEX-SPOT, but they do not isolate the contribution of the SPOT objective from that of the CORTEX architecture. To address this, we compare masked reconstruction and SPOT on CORTEX and LaBraM, denoting the reconstruction variants as CORTEX-recon and LaBraM-recon. Within each backbone, we hold the architecture, pretraining corpus and sampling order, training budget, and downstream probe protocol fixed, changing only the pretraining objective. CORTEX tests this change within the proposed channel-agnostic encoder, while LaBraM tests whether the effect extends to a published EEG backbone. Appendix I provides configuration checks and full SNR curves under muscle-band, pink, and white noise (Figures 9, 10, and 11).

SPOT improves robustness across both backbones. On CORTEX, SPOT achieves higher absolute performance under contamination in every task in Table 3. On LaBraM, it improves absolute performance in every task with performance above the trivial floor. Clean performance remains task-dependent: reconstruction is higher on some LaBraM tasks, while SPOT is higher on others. Under contamination, however, the ordering is consistent across both architectures.

Statistical support. We evaluate five probe seeds per arm using a single pretrained checkpoint for each. Welch's t-tests across the six tasks evaluated on each of the two backbones use a Bonferroni-corrected threshold of $0.05/12$, separately for clean scores, $12\,\mathrm{dB}$, and grid means. SPOT achieves significant gains in eight comparisons under each contamination condition; on clean data, each objective is significantly better in four comparisons.

Table 3: **Objective transfer across CORTEX and LaBraM.** Clean scores are mean±sd over five probe seeds; contaminated scores are 12 dB/grid mean. Bold marks the higher score within each backbone. p-values use Welch's t-test; * denotes $p < 0.05/12$ within each condition.

		Staging		HSP		Seizure	
Model	**Condition**	HSP	HMC	Arousal	Apnea	TUH	Siena
CORTEX-recon	clean	54.8±0.4	**55.3±0.9**	39.0±1.9	25.3±1.2	69.1±2.0	11.3±6.6
	12 dB/grid	42.5/24.2	44.3/16.2	24.0/18.1	16.1/15.8	58.4/39.1	10.0/ 6.9
CORTEX-SPOT	clean	**55.3±0.8**	52.9±0.7	**57.3±0.6**	**42.2±0.7**	**70.7±1.3**	**33.5±17.4**
	12 dB/grid	**53.8/25.2**	**51.7/32.7**	**53.9/30.0**	**29.3/19.5**	**70.1/51.8**	**30.2/16.0**
	p, clean	0.275	0.003*	$<$0.001*	$<$0.001*	0.222	0.061
	p, 12 dB/grid	$<$0.001*/0.547	0.012/$<$0.001*	$<$0.001*/$<$0.001*	$<$0.001*/0.002*	$<$0.001*/0.001*	0.077/0.140
LaBraM-recon	clean	**54.1±0.8**	**51.9±0.6**	**24.4±0.1**	15.6±0.3	50.4±0.3	2.4±0.1
	12 dB/grid	27.8/12.6	0.4/ 0.1	17.1/15.0	**15.1/** 6.9	22.4/22.9	2.8/ 4.1
LaBraM-SPOT	clean	46.8±2.6	47.4±1.1	21.4±0.1	**15.8±0.7**	**54.3±0.7**	**27.7±0.5**
	12 dB/grid	**33.9/14.6**	**27.4/ 6.9**	**22.6/17.3**	14.0/ **7.7**	**55.8/45.6**	**27.3/25.8**
	p, clean	0.003*	$<$0.001*	$<$0.001*	0.662	$<$0.001*	$<$0.001*
	p, 12 dB/grid	0.018/0.347	$<$0.001*/$<$0.001*	$<$0.001*/$<$0.001*	0.055/0.171	$<$0.001*/$<$0.001*	$<$0.001*/$<$0.001*

The CORTEX comparison isolates the objective's contribution within the proposed encoder, while the LaBraM comparison shows that the robustness advantage extends beyond the CORTEX architecture. Together, these controlled comparisons support an objective-level explanation for the observed gains across the two tested backbones.

5 Limitations

Our theoretical analysis assumes a jointly Gaussian additive-noise model with aligned signal and noise covariances, a simplification of real physiological recordings. Theorem 1 establishes the existence of a predictive-optimal code excluding predictive-null components; it does not guarantee a trained encoder attains this allocation, or that allocation alone determines downstream robustness. Objective-transfer experiments cover two architectures, CORTEX and the published LaBraM backbone; broader architectural coverage is ongoing and was not completed in time for submission. The five-seed comparisons vary the downstream probe with a single pretrained checkpoint per arm, so they do not capture variability across pretraining runs, and we do not separately ablate candidate count, FSQ depth, EMA decay, context length, or the contribution of tube masking versus the selective objective itself. The robustness advantage also depends on the contamination regime. On HSP staging, CORTEX-SPOT leads at 12 dB, whereas NeuroRVQ achieves the highest grid-mean score (Table 2). Finally, our evaluations use retrospective datasets with synthetic or injected artifacts, which do not establish clinical validity or reliability in prospective deployment. Appendix J provides further discussion.

6 Conclusion

We introduced SPOT, a selective-prediction objective for biosignal tokenization that learns discrete targets without waveform or spectral reconstruction. Gaussian analysis characterizes conditions under which reconstruction allocates rate to context-uninformative directions, while probe-based experiments reveal corresponding differences in learned representations. Across eight clinical evaluations, CORTEX-SPOT combines competitive clean performance with strong robustness under contamination. Matched comparisons on CORTEX and LaBraM support the contribution of the pretraining objective across both backbones. These findings motivate context-based prediction for biosignal tokenization, while leaving prospective clinical validation and broader architectural evaluation to future work.

Reproducibility statement

All corpora are existing public or credentialed-access datasets; we introduce no new data: TUH EEG and TUSZ (Obeid & Picone, 2016; Shah et al., 2018), the TUH Artifact Corpus (Buckwalter et al., 2021), Sleep-EDF (Kemp et al., 2000), HMC (Alvarez-Estevez & Rijsman, 2021), CHB-MIT (Shoeb, 2009), Siena (Detti et al., 2020), and the Human Sleep Project (Brain Data Science Platform, 2020), the last of which requires a data use agreement. Appendix F gives the complete pretraining configurations for both objective arms, read directly from the released checkpoints' stored configs

and launch scripts rather than from our notes, including the fields confirming that neither encoder was trained with augmented noise. Appendix G records the implementation details needed to reproduce the experiments, including the precision requirement for LaBraM's tokenizer target, without which the objective swap in §4.5 cannot be run, and the compute resources used. Appendix I reports configuration checks confirming that, within each backbone, the two objective arms share the same encoder architecture, pretraining corpus, sampling order, training budget, and downstream probe protocol, with the pretraining objective as the sole experimental variable. Section 4.1 and Appendix E describe the evaluation protocol. Appendix H.4 defines retention and documents the exclusions in Table 7. Code, pretrained CORTEX checkpoints for both arms, the probe definitions used in §4.2, and scripts that regenerate every table and figure from a single results file will be released publicly on acceptance; an anonymized copy is available to reviewers as supplementary material.

Ethics statement

All recordings come from existing de-identified corpora used under their published terms; we collected no new data and had no contact with participants. This work reports frozen-probe accuracy under synthetic and injected contamination and is not a clinical validation: the seizure and sleep scores here are benchmark measurements, not evidence of diagnostic fitness, and a robustness improvement measured this way should not be read as a safety argument for deployment. Automated sleep and seizure scoring carries the usual risks of unequal performance across populations and recording hardware; our evaluations span three seizure corpora and three sleep corpora with different hardware and populations, but we do not report performance stratified by demographic group, and the corpora themselves are not demographically representative.

AI-use statement

We used generative AI tools to assist with coding, the preparation of tables and figures from saved experimental outputs, polishing the writing, and drafting sections of the paper. All AI-drafted or AI-polished text was reviewed and edited by the authors, and all AI-assisted code was reviewed by the authors before use; for example, the LaBraM objective-swap implementation was checked term by term against its published description. All research questions, experimental design, model implementation, training runs, and measurements are the authors' own. Every number reported in this paper was produced by the authors' code and verified by the authors against the corresponding result files on disk. We did not use generative AI tools to generate research ideas or to develop theoretical models or conceptual frameworks, and no AI system generated experimental results, citations, or claims that the authors did not independently verify. No AI system is an author of this work, and the authors take full responsibility for the final content of this submission, including any text that originated as an AI-assisted suggestion.

References

Diego Alvarez-Estevez and Roselyne M. Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):e0256111, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Mido Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Konstantinos Barmpas, Na Lee, Dimitrios Chalatsis, William Raftery, Yannis Panagakis, Dimitrios A. Adamos, et al. Neurorvq: Multi-scale biosignal tokenization for generative foundation models. *arXiv preprint arXiv:2510.13068*, 2025.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

Brain Data Science Platform. The brain data science platform and the human sleep project, 2020. Massachusetts General Hospital; credentialed access, `https://bdsp.io`.

G. Buckwalter, S. Chhin, S. Rahman, I. Obeid, and J. Picone. Recent advances in the TUH EEG corpus: Improving the interrater agreement for artifacts and epileptiform events. In *IEEE Signal Processing in Medicine and Biology Symposium*, 2021.

Chi-Yuan Chang, Sheng-Hsiou Hsu, Luca Pion-Tonachini, and Tzyy-Ping Jung. Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel EEG recordings. *IEEE Transactions on Biomedical Engineering*, 67(4):1114–1121, 2020.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, et al. Uni-ntfm: A unified foundation model for eeg signal representation learning. *arXiv preprint arXiv:2509.24222*, 2025.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Felix Creutzig, Amir Globerson, and Naftali Tishby. Past-future information bottleneck in dynamical systems. *Physical Review E*, 79(4):041925, 2009.

Mingzhe Cui, Tao Chen, Yang Jiao, Yiqin Wang, Lei Xie, Yi Pan, and Luca Mainardi. Brainrvq: A high-fidelity eeg foundation model via dual-domain residual quantization and hierarchical autoregression. *arXiv preprint arXiv:2602.16951*, 2026.

Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7):846, 2020.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG. adapting to any setup with large-scale pretraining on 25,000 subjects. *Advances in Neural Information Processing Systems*, 2025.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=QzTpTRVtrP`.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multitask foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17):e1750–e1762, 2023.

Bob Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A. C. Kamphuisen, and Jos J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

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, 2018.

Sarah E. Marzen and James P. Crutchfield. Predictive rate-distortion for infinite-order Markov processes. *Journal of Statistical Physics*, 163(6):1312–1338, 2016.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, 2020. arXiv:1811.04251.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *The Twelfth International Conference on Learning Representations*, 2024. arXiv:2309.15505.

Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5171–5180, 2019.

Vinit Shah, Eric von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018.

Ali Hossam Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Susanne Still. Information bottleneck approach to predictive inference. *Entropy*, 16(2):968–989, 2014.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999. arXiv:physics/0004057.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. *Advances in Neural Information Processing Systems*, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBramod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, 2023a.

Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised EEG representation learning for automatic sleep staging. In *JMIR AI*, 2023b. arXiv:2110.15278.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *arXiv preprint arXiv:2506.23075*, 2025.

A EXTENDED RELATED WORK

Biosignal tokenization and pretraining. LaBraM (Jiang et al., 2024) adapts BEiT (Bao et al., 2022) to EEG by first training a discrete tokenizer to reconstruct Fourier amplitude and phase, then using its codes as masked-prediction targets. NeuroRVQ (Barmpas et al., 2025) improves reconstruction through multi-scale features and residual vector quantization, while BrainRVQ (Cui et al., 2026) combines dual-domain residual quantization with hierarchical code prediction. These methods distinguish tokenizer training from subsequent prediction of the learned codes. SPOT changes the former: it trains discrete representations through contextual prediction without a reconstruction objective.

Other EEG foundation models use reconstruction without a discrete codebook, including CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025), and the dual-domain reconstruction framework Uni-NTFM (Chen et al., 2025). Contrastive and hybrid approaches offer alternative pretraining objectives (Kostas et al., 2021; Wang et al., 2024; Yang et al., 2023a). Our focus is specifically on how the objective used to learn discrete targets affects the information they retain. Matched comparisons on CORTEX and LaBraM examine this choice while holding the encoder architecture fixed within each comparison (§4.5).

Prediction in representation space. Contextual prediction is established in speech, audio, and vision. Contrastive predictive coding (van den Oord et al., 2018) learns representations by identifying future latent targets from context. wav2vec 2.0 (Baevski et al., 2020) applies contrastive prediction to quantized speech targets, HuBERT (Hsu et al., 2021) predicts cluster assignments, and BEATs (Chen et al., 2023) iteratively refines an acoustic tokenizer and a discrete-target predictor. data2vec (Baevski et al., 2022), I-JEPA (Assran et al., 2023), and V-JEPA 2 (Assran et al., 2025) instead predict teacher or latent representations without reconstructing the input. SPOT builds on these principles for biosignal tokenization. Its contribution is to connect the choice between reconstruction and contextual prediction to nuisance allocation and to evaluate the resulting representations under controlled contamination.

Predictive information and rate allocation. Predictive information quantifies the information that past observations contain about future observations (Bialek et al., 2001). Our objective in equation 1 adopts the information-bottleneck formulation (Tishby et al., 1999), with the surrounding recording as the relevance variable. Related formulations study compression for future prediction and predictive rate-distortion (Creutzig et al., 2009; Still, 2014; Marzen & Crutchfield, 2016). For jointly Gaussian variables, the information-bottleneck solution allocates rate along canonical-correlation directions (Chechik et al., 2005), whereas reconstruction under mean-squared error follows reverse water-filling (Cover & Thomas, 2006).

Our analysis compares these allocation rules to establish conditions under which reconstruction assigns rate to context-uninformative directions, while a predictive-optimal code excludes them. We investigate the corresponding behavior in learned representations through probe-based diagnostics. These diagnostics measure recoverable information for comparison across encoders, rather than directly estimating mutual information, whose estimation from finite samples presents substantial challenges (McAllester & Stratos, 2020; Poole et al., 2019).

Robustness and downstream evaluation. Artifact-removal methods, including artifact subspace reconstruction (Chang et al., 2020), address contamination through preprocessing; noise augmentation modifies the training distribution. Our experiments examine a different factor: the pretraining objective, with contamination introduced only at evaluation. Synthetic noise tests are complemented by injections of expert-labelled muscle artifacts from TUAR (Buckwalter et al., 2021). Task-specific supervised models provide an additional reference (Lawhern et al., 2018; Jing et al., 2023; Yang et al., 2023b). These models adapt all parameters to the downstream task, whereas our frozen-encoder evaluations train only a prediction head. We therefore report the two groups separately when assessing robustness.

B CORTEX ARCHITECTURE DETAILS

Figure 3 shows the CORTEX encoder from the raw recording to its contextualized cell representations. Each token corresponds to one *(sensor, 1-second)* cell. Sensor identity is indexed by electrode or sensor name rather than channel-list position, allowing the same encoder to process recordings with different montages and sensor sets.

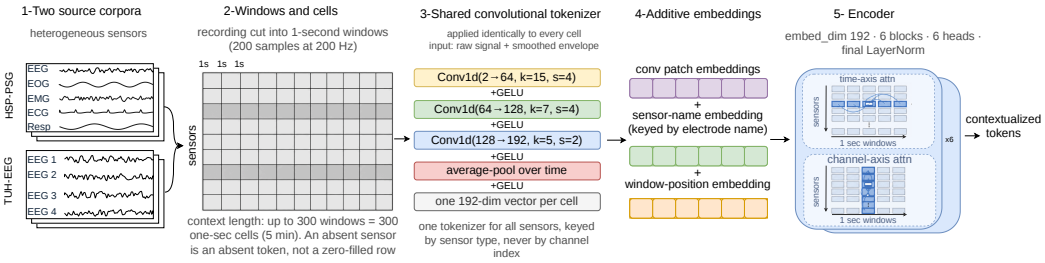


Figure 3: **The CORTEX encoder** (§2.2). Each token represents a one-second sensor cell. Sensor identity is indexed by name rather than channel order, allowing the same encoder to process different montages. Missing sensors are represented by omitted tokens.

The CORTEX encoder shown in Figure 3 is the encoder used in the main-text overview of SPOT (Figure 1). In that figure, CORTEX encodes the visible sensor-time cells into contextual representations, which are used to predict the discrete code of a hidden cell.

C ADDITIONAL THEORETICAL ANALYSIS

This appendix states the assumptions underlying Theorem 1, clarifies its scope, and relates predictive rate allocation to the Gaussian information bottleneck. Full derivations are provided in Appendix D. All logarithms are base two, and rates are measured in bits.

Assumption 1 (Jointly Gaussian aligned additive-noise model)**.** Let $X \in \mathbb{R}^d$ denote a signal patch, S its physiological component, N its noise component, and $C \in \mathbb{R}^m$ its context. These variables are jointly Gaussian and zero-mean, with

$$X = S + N, \qquad S \perp N, \qquad N \perp C.$$

Thus, any dependence between the patch and its context arises through S. Assume that an orthonormal basis simultaneously diagonalizes $\mathrm{Cov}(S)$ and $\mathrm{Cov}(N)$. In this basis, let $\sigma^2_{s,i}$ and $\sigma^2_{n,i}$ denote the component variances, and define

$$\mathcal{I}_0 := \{i : \mathrm{Cov}(X_i, C) = 0\}, \qquad \mathcal{N} := \mathrm{span}\{e_i : i \in \mathcal{I}_0\}, \qquad \mathcal{P} := \mathcal{N}^{\perp}.$$

The set $\mathcal{I}_0$ indexes the *predictive-null coordinates*, which carry no information about the specified context. Let $X_{\mathcal{N}}$ and $X_{\mathcal{P}}$ denote the corresponding projections. The diagonal covariance structure and joint Gaussianity imply

$$X_{\mathcal{N}} \perp (X_{\mathcal{P}}, C).$$

The alignment assumption holds, for example, when the signal and noise covariances are diagonal in a common real Fourier basis. The noise term collects components assumed independent of the context, such as sensor noise or unpredictable artifacts. Predictive-null coordinates are defined relative to C and can contain both noise and physiological variation that is independent of this context.

Corollary 1 (Artifact dominance)**.** *Under Assumption 1, fix $R > 0$ and a finite dimension d, and suppose that both $\mathcal{I}_0$ and its complement are nonempty. If*

$$\frac{\min_{i \in \mathcal{I}_0} \sigma^2_{n,i}}{\max_{i \notin \mathcal{I}_0} \sigma^2_{x,i}} \longrightarrow \infty, \qquad \sigma^2_{x,i} = \sigma^2_{s,i} + \sigma^2_{n,i},$$

then $\psi_{\mathrm{rec}}(R) \to 1$, whereas a predictive-optimal code can be chosen with $\psi_{\mathrm{pred}}(R) = 0$.

Remark 2 (Low SNR is insufficient for artifact dominance)**.** Corollary 1 requires a particular distribution of noise variance across coordinates; low SNR alone does not imply its conclusion. For

example, consider $X_1 = S_1 + N_1$ and $X_2 = N_2$, with mutually independent S_1, N_1, N_2, noise independent of C, $\mathrm{Var}(N_1) = \mathrm{Var}(N_2) = 1$, and $\mathrm{Var}(S_1) = \rho$. Suppose $\mathrm{Cov}(S_1, C) \neq 0$ for each $\rho > 0$, so that only the second coordinate is predictive-null. As $\rho \to 0$, the two input variances converge to one. At any fixed $R > 0$, reverse water-filling therefore allocates asymptotically equal rates to the two coordinates, giving $\psi_{\mathrm{rec}}(R) \to \frac{1}{2}$. The synthetic experiment in Section 4.2 instead assigns the nuisance directions four times the noise power of the signal-bearing directions. This finite variance imbalance illustrates the allocation mechanism; it is not itself the asymptotic condition in Corollary 1.

Remark 3 (Noise within predictive directions)**.** Theorem 1 concerns the allocation of rate to a predictive-null subspace that is independent of the remaining coordinates and the context. It does not establish that predictive coding eliminates noise superimposed on context-predictive directions. Reallocating rate across coordinates cannot separate signal and noise within the same coordinate.

Remark 4 (The scalar case)**.** When $d = 1$ and $\mathrm{Cov}(X, C) \neq 0$, the predictive-null subspace is $\mathcal{N} = \{0\}$, leaving no separate null direction to discard. If $\mathrm{Cov}(X, C) = 0$, the entire scalar input is predictive-null and $\mathrm{I}(X; C) = 0$. Thus, a nontrivial separation between predictive and predictive-null subspaces requires $d > 1$.

For a nondegenerate scalar input, let r_{xc} denote its canonical correlation with C (the absolute Pearson correlation when C is scalar), and define r_{sc} analogously. With $\rho := \mathrm{Var}(S)/\mathrm{Var}(N)$, the additive-noise assumptions give

$$r_{xc}^2 = \frac{\rho}{1+\rho} r_{sc}^2.$$

Every encoder satisfies $C \to X \to Z$, so data processing yields

$$\mathrm{I}(Z; C) \leq \mathrm{I}(X; C) = -\tfrac{1}{2}\log(1 - r_{xc}^2) \longrightarrow 0 \qquad \text{as } \rho \to 0.$$

Moreover, the scalar Gaussian MSE-optimal reconstruction channel attains the predictive frontier at the same rate:

$$P(R) = -\tfrac{1}{2}\log\big(1 - (1 - 2^{-2R}) r_{xc}^2\big), \qquad \Delta_{\text{pred}-\text{rec}}(R) = 0, \tag{6}$$

where $\Delta_{\text{pred}-\text{rec}}(R)$ denotes the difference in retained predictive information between the two optimal codes.

C.1 Gaussian information-bottleneck interpretation

The constrained objective in equation 1 defines the rate–predictiveness frontier

$$P(R) := \sup_{q(z|x):\, \mathrm{I}(Z;X) \leq R} \mathrm{I}(Z; C).$$

This is the information-bottleneck value function with the context as the relevance variable (Tishby et al., 1999; Creutzig et al., 2009). For jointly Gaussian (X, C), the Gaussian information-bottleneck solution decomposes the allocation problem along canonical-correlation directions (Chechik et al., 2005). Let $1 \geq \lambda_1 \geq \cdots \geq \lambda_K \geq 0$ denote the canonical correlations on the nondegenerate supports of X and C. Then

$$P(R) = \max_{\substack{\sum_{k=1}^K r_k \leq R \\ r_k \geq 0}} \sum_{k=1}^{K} -\tfrac{1}{2}\log\big(1 - (1 - 2^{-2r_k})\lambda_k^2\big)\,. \tag{7}$$

Here, r_k is the rate allocated to the kth canonical direction. Directions with stronger canonical correlations receive priority, whereas those with zero canonical correlation provide no predictive information at any rate. Under Assumption 1, every direction in $\mathcal{N}$ is uncorrelated with C, consistent with Theorem 1. This coordinate-defined subspace need not exhaust all context-uninformative linear combinations of the input.

The two objectives therefore induce different allocation criteria: reconstruction reverse water-filling depends on input variance, whereas predictive allocation depends on canonical correlation with the context. This established Gaussian result provides an interpretation of the allocation mechanism rather than an additional theoretical contribution.

D PROOFS OF THEORETICAL RESULTS.

Proofs of every result stated in §3 and Appendix C, with the lemmas they rest on.

D.1 PRELIMINARIES

For a scalar Gaussian source of variance v under mean-squared error, the rate-distortion function is $R(D) = \frac{1}{2}\log^+(v/D)$ with $\log^+ t := \max\{\log t, 0\}$, and a Gaussian test channel at rate r leaves distortion $D = v\,2^{-2r}$; equivalently the backward channel $x = \hat{x} + e$ has $\mathrm{Var}(\hat{x}) = v(1 - 2^{-2r})$, so the *represented-variance fraction* is $1 - 2^{-2r}$. For jointly Gaussian (a, b) with correlation r_{ab}, $\mathrm{I}(a; b) = -\frac{1}{2}\log(1 - r_{ab}^2)$. If $\hat{u}$ is the rate-r representation of u and (u, c) has correlation r_{uc}, then $r_{\hat{u}c}^2 = (1 - 2^{-2r})r_{uc}^2$, hence

$$\mathrm{I}(\hat{u}; c) = -\tfrac{1}{2}\log\big(1 - (1 - 2^{-2r})\, r_{uc}^2\big), \tag{8}$$

the identity used throughout. Logs are base 2 and rates are in bits.

D.2 PROOF OF THEOREM 1

Lemma 1 (Reverse water-filling)**.** *Under Assumption 1, the minimum-distortion code at rate $\mathrm{I}(Z; X) \le R$ assigns coordinate i the rate $r_i^{\mathrm{rec}} = \frac{1}{2}\log^+(\sigma_{x,i}^2/\theta)$, where θ is the unique solution of $\sum_i \frac{1}{2}\log^+(\sigma_{x,i}^2/\theta) = R$; that is, coordinate i is coded to residual distortion θ if and only if $\sigma_{x,i}^2 > \theta$, and dropped otherwise.*

Proof. The aligned coordinates are independent Gaussians, and both $\mathrm{I}(Z; X) = \sum_i \mathrm{I}(Z_i; X_i)$ and the distortion $\sum_i \mathbb{E}(X_i - \hat{X}_i)^2$ are additive, so the problem is the allocation $\min \sum_i \sigma_{x,i}^2 2^{-2r_i}$ subject to $\sum_i r_i = R$, $r_i \ge 0$. The Lagrangian $\sum_i \sigma_{x,i}^2 2^{-2r_i} + \lambda(\sum_i r_i - R)$ has stationarity $-2\ln 2\, \sigma_{x,i}^2 2^{-2r_i} + \lambda = 0$ on active coordinates, that is a common residual distortion $D_i = \sigma_{x,i}^2 2^{-2r_i} = \theta := \lambda/(2\ln 2)$ wherever $r_i > 0$, and $r_i = 0$ where the unconstrained optimum is negative, that is $\sigma_{x,i}^2 \le \theta$. Solving $D_i = \theta$ gives the stated rate; the map $1/\theta \mapsto \sum_i \frac{1}{2}\log^+(\sigma_{x,i}^2/\theta)$ is strictly increasing wherever it is positive, so the θ meeting the budget is unique. □

Lemma 2 (Predictive optimum ignores the predictive-null subspace)**.** *Under Assumption 1, let $X = (X_{\mathcal{P}}, X_{\mathcal{N}})$ be the decomposition induced by $\mathcal{P} = \mathcal{N}^{\perp}$ and $\mathcal{N}$. Among the maximisers of*

$$\max_{q(z|x):\, \mathrm{I}(Z;X) \le R} \mathrm{I}(Z; C),$$

there exists one satisfying $q^*(z \mid x_{\mathcal{P}}, x_{\mathcal{N}}) = q^*(z \mid x_{\mathcal{P}})$, and therefore $\mathrm{I}(Z^*; X_{\mathcal{N}} \mid X_{\mathcal{P}}) = 0$, equivalently $\psi(Z^*) = 0$.

Proof. By construction $\mathrm{Cov}(X_{\mathcal{N}}, C) = 0$, and the aligned basis also gives $\mathrm{Cov}(X_{\mathcal{N}}, X_{\mathcal{P}}) = 0$. Since $(X_{\mathcal{N}}, X_{\mathcal{P}}, C)$ is jointly Gaussian, $X_{\mathcal{N}} \perp (X_{\mathcal{P}}, C)$.

Let $q(z \mid x_{\mathcal{P}}, x_{\mathcal{N}})$ be any feasible encoder and define an encoder that ignores the predictive-null component by marginalising it out:

$$\widetilde{q}(z \mid x_{\mathcal{P}}) := \int q(z \mid x_{\mathcal{P}}, x_{\mathcal{N}}) p(x_{\mathcal{N}})\, dx_{\mathcal{N}}.$$

Using $X_{\mathcal{N}} \perp (X_{\mathcal{P}}, C)$,

$$\begin{aligned} p_{\widetilde{q}}(z, c) &= \int p(x_{\mathcal{P}}, c)\, \widetilde{q}(z \mid x_{\mathcal{P}})\, dx_{\mathcal{P}} \\ &= \iint p(x_{\mathcal{P}}, c) p(x_{\mathcal{N}}) q(z \mid x_{\mathcal{P}}, x_{\mathcal{N}})\, dx_{\mathcal{N}}\, dx_{\mathcal{P}} \\ &= p_q(z, c), \end{aligned}$$

so $\mathrm{I}_{\widetilde{q}}(Z;C) = \mathrm{I}_q(Z;C)$. The same marginalisation preserves the joint law of $(Z, X_{\mathcal{P}})$, so $\mathrm{I}_{\widetilde{q}}(Z;X_{\mathcal{P}}) = \mathrm{I}_q(Z;X_{\mathcal{P}})$. Because $\widetilde{q}$ does not depend on $X_{\mathcal{N}}$, $\mathrm{I}_{\widetilde{q}}(Z;X) = \mathrm{I}_{\widetilde{q}}(Z;X_{\mathcal{P}})$, whereas the chain rule gives

$$\mathrm{I}_q(Z;X) = \mathrm{I}_q(Z;X_{\mathcal{P}}) + \mathrm{I}_q(Z;X_{\mathcal{N}} \mid X_{\mathcal{P}}) \geq \mathrm{I}_q(Z;X_{\mathcal{P}}).$$

Hence $\mathrm{I}_{\widetilde{q}}(Z;X) \leq \mathrm{I}_q(Z;X)$, so $\widetilde{q}$ remains feasible and has exactly the same predictive information. Applying the construction to a predictive-information maximiser therefore produces another maximiser that ignores $X_{\mathcal{N}}$, proving $\mathrm{I}(Z^*;X_{\mathcal{N}} \mid X_{\mathcal{P}}) = 0$. □

The statement is an existence claim because predictive-information maximisers need not be unique: once predictive capacity is saturated, an encoder may carry additional information about $X_{\mathcal{N}}$ without increasing $\mathrm{I}(Z;C)$. The lemma shows that such dependence is never *required*. It does not show that a particular training procedure will find the optimum that avoids it, which is why §4.2 measures rather than assumes it.

Proof of Theorem 1. Lemma 1 gives the reconstruction-optimal allocation $r_i^{\mathrm{rec}} = \frac{1}{2}\log^+(\sigma_{x,i}^2/\theta)$, so every $i \in \mathcal{I}_0$ with $\sigma_{x,i}^2 > \theta$ receives positive reconstruction rate and consequently $\psi_{\mathrm{rec}}(R) > 0$ whenever at least one predictive-null coordinate clears the water level. For the predictive objective, Lemma 2 shows an optimum exists with $\mathrm{I}(Z_{\mathrm{pred}};X_{\mathcal{N}} \mid X_{\mathcal{P}}) = 0$, and therefore $\psi_{\mathrm{pred}}(R) = 0$. Together these establish equation 5. □

Proof of Corollary 1. Write $\mathcal{A}(\theta) = \{i : \sigma_{x,i}^2 > \theta\}$ for the coded set of Lemma 1 and let $\kappa := \min_{i\in\mathcal{I}_0}\sigma_{n,i}^2 / \max_{i\notin\mathcal{I}_0}\sigma_{x,i}^2$. Along a family with $\kappa \to \infty$, normalise so that $\max_{i\notin\mathcal{I}_0}\sigma_{x,i}^2 = 1$; every nuisance coordinate then has variance at least κ. Fix the budget R. Water-filling gives coordinate i the rate $\frac{1}{2}\log^+(\sigma_{x,i}^2/\theta)$, so the rate on $\mathcal{N}$ is at least $\frac{|\mathcal{I}_0|}{2}\log^+(\kappa/\theta)$ while every signal-bearing coordinate receives at most $\frac{1}{2}\log^+(1/\theta)$. Since the total is held at R, the common water level satisfies $\theta \geq \kappa\, 2^{-2R/|\mathcal{I}_0|}$ for κ large enough that the nuisance coordinates alone can absorb R; from that point on $\sigma_{x,i}^2 = 1 \leq \theta$ for every $i \notin \mathcal{I}_0$, so the signal-bearing coordinates drop out of $\mathcal{A}(\theta)$ entirely and receive zero rate. Hence $\psi_{\mathrm{rec}}(R) \to 1$. That $\psi_{\mathrm{pred}}(R) = 0$ is Lemma 2, which is unconditional on the variances and so is unaffected by κ. □

D.3 InfoNCE as a contrastive surrogate

Under standard sampling assumptions, InfoNCE provides a lower bound on mutual information: positive pairs are drawn from the joint distribution, and negatives are sampled independently from the appropriate marginal (van den Oord et al., 2018; Poole et al., 2019). Our in-batch candidates include cells from the same recording and subject (§2.1), so these assumptions are not guaranteed. We therefore use InfoNCE as a training surrogate and do not interpret its value as a mutual-information estimate or certified lower bound.

Separately, the empirical analysis in §4.2 uses probe-derived diagnostics ($\widehat{\psi}_{\mathrm{probe}}$ and $\widehat{I}_{\mathrm{probe}}$) to assess nuisance allocation, while §H.2 examines code utilization and cross-domain marginal distributions.

E Evaluation details

Table 4 lists the eight downstream evaluations, their input units, metrics, trivial-performance floors, evaluation-set sizes, and coverage across published foundation models. The floor denotes the score of a trivial predictor and provides a reference for interpreting performance under contamination.

F Pretraining configuration

Table 5 reports the configurations stored in the CORTEX checkpoints and launch scripts. For SPOT, tube masking hides a contiguous span of $L = 225$ one-second cells within each 300-second channel window. The start position t is sampled independently per channel with probability proportional to

Table 4: **Downstream evaluations.**

Task	Corpus	Unit	Metric	Floor	eval n
Staging	HSP (Brain Data Science Platform, 2020)	30 s	κ	0.0	1,468,125
Arousal	HSP (Brain Data Science Platform, 2020)	$\geq$3 s	AUPRC	11.1	1,468,125
Apnea	HSP (Brain Data Science Platform, 2020)	$\geq$10 s	macro-F1	15.7	1,468,125
Staging	Sleep-EDF (Kemp et al., 2000)	30 s	κ	0.0	514,500
Staging	HMC (Alvarez-Estevez & Rijsman, 2021)	30 s	κ	0.0	274,920
Seizure	TUH (TUSZ) (Shah et al., 2018)	3 s	AUPRC	6.89	135,537
Seizure	CHB-MIT (Shoeb, 2009)	3 s	AUPRC	0.27	175,239
Seizure	Siena (Detti et al., 2020)	3 s	AUPRC	0.82	32,201

Table 5: **CORTEX pretraining configurations.** The upper block lists shared settings; the lower block lists objective-specific components.

Setting	CORTEX-recon	CORTEX-SPOT (B70)
Shared		
pretraining corpus	HSP per-second cache + TUH, same `splits.json`	
subjects	5,000	
token resolution	`subdiv`= 3 (1-s tokens from a 600-sample window)	
context	`max_windows`= 100 ($\approx$ 5 min)	
encoder body	`embed_dim` 192, depth 6, heads 6	
tokenizer front end	shared convolution (`tokenizer_kind=conv`)	
encoder parameters	5,589,440	
montage	bipolar	
mask ratio	0.75	
epoch budget	70 (best at epoch 65 / 67)	
batch size	32	
learning rate	2.2e$-$4	
early-stop patience	999 (runs to budget)	
weight decay	1e$-$4	
pretrain seed	0	
noise augmentation	`noise_snr_db`: `None`, so **no synthetic noise augmentation.**	
Differing		
objective head	linear log-power spectrum	depth-stacked FSQ + InfoNCE
quantizer	—	`fsq_levels` $[8]^8$, depth 5 (40 code dims)
negative candidates	—	4096
target normalisation	—	`target_norm`= 1
EMA decay	—	0.9999

$\sum_{i=t}^{t+L-1} \mathrm{sd}(x_i)$, favoring spans with higher variability. The context consists of the unmasked cells across the recording.

Neither arm uses noise augmentation during pretraining. Controlled contamination is introduced only during downstream evaluation (§4.1).

G Reproducibility notes

The following implementation details complement the configurations in Appendix F.

Target normalization. LaBraM's VQ-NSP target normalization can overflow in fp16 on our recordings. The reduction involves approximately 24,000 elements, with the patch FFT's DC bin reaching roughly 200$\times$ the mean bin magnitude. We therefore compute target normalization in fp32 while retaining fp16 for the remaining forward pass.

SNR sweeps. We evaluate one saved checkpoint across all noise conditions before loading the next. This keeps one model resident at a time and avoids the memory requirements of processing the 282 GB evaluation array with multiple models in memory.

Batch composition. In our LaBraM pipeline, PSG and TUH windows are batched separately because their montages differ. CORTEX supports mixed-montage batches through sensor-identity embeddings. Within each backbone, the reconstruction and SPOT arms use identical batching procedures.

Input scaling. Cached signals are stored in μV and rescaled according to each backbone's input convention. CBraMod uses a factor of 0.01, matching the $\mu\mathrm{V}/100$ input supplied through CSBrain's `scale=1.0` path. We checked activation ranges across three orders of magnitude in input scale to identify inactive or saturated responses.

Compute. All pretraining and downstream training used a shared institutional cluster with NVIDIA H100, L40S, RTX 6000 Ada, and A100-class GPUs, with one GPU per job. For the matched objective-transfer experiments in §4.5, pretraining wall-clock times were 11.1 and 6.1 hours for CORTEX-recon and CORTEX-SPOT, respectively, and 21.1 and 22.1 hours for LaBraM-recon and LaBraM-SPOT, respectively.

H Additional experimental results

This appendix presents nuisance-probe results, codebook diagnostics, dataset-specific evaluation details, and additional contamination experiments complementing §4.

H.1 Probe-derived nuisance information on real EEG

Probe definition. For each frozen encoder, we probe the continuous, 192-dimensional contextualized representation of each sensor-time cell, before quantization, so that both arms are evaluated using the same representation type. A linear ridge probe predicts a 32-bin log-power spectral summary of the target signal, with logarithmically spaced bins over 0.5–$100\,\mathrm{Hz}$. We fit separate probes for the denoiser-derived signal estimate $\hat{s}$ and residual $\hat{n} = x - \hat{s}$. Each probe is fit on half of the pooled $40{,}000$ cells and evaluated on the remaining half using a cell-level split. Input features are standardized using statistics from the fitting half. The ridge penalty is fixed at $\lambda = 10^{-2}$, with an unregularized intercept.

For target $y \in \{\hat{s}, \hat{n}\}$, we define

$$\widehat{I}_{\mathrm{probe}}(Z; y) = -\frac{1}{2}\sum_{k=1}^{32} \ln\Big(1 - \widetilde{R}_k^2\Big), \qquad \widetilde{R}_k^2 = \min\big\{0.999, \max\{0, R_k^2\}\big\},$$

where R_k^2 is the held-out coefficient of determination for spectral bin k. This transformation yields a probe-derived diagnostic in nats; it is not a calibrated estimate of joint mutual information across spectral bins. The within-model nuisance share is

$$\widehat{s}_{\mathrm{probe}} = \frac{\widehat{I}_{\mathrm{probe}}(Z; \hat{n})}{\widehat{I}_{\mathrm{probe}}(Z; \hat{n}) + \widehat{I}_{\mathrm{probe}}(Z; \hat{s})}.$$

Table 6 and Figure 4 report the nuisance-probe results on $40{,}000$ EEG cells. The absolute probe diagnostic is lower for SPOT in every condition. The difference in within-model nuisance share narrows with increasing contamination and slightly reverses at $3\,\mathrm{dB}$, where the reconstruction-to-SPOT ratio is 0.97.

H.2 Codebook utilization across domains

We assess code utilization using the frozen SPOT encoder on HSP sleep cells from the pretraining pool and held-out TUH recordings. Using the quantizer configuration stored in the checkpoint, we measure per-dimension perplexity, inactive dimensions, and cosine similarity between the two domains' marginal code distributions.

The quantizer has eight scalar levels, five residual stages, and forty code dimensions. No dimension is inactive: the minimum perplexity is 3.49, and mean per-dimension perplexities across the five stages are 5.68, 5.57, 6.21, 5.98, and 6.15, respectively, against a maximum of 8.

Table 6: Probe-derived nuisance information on real EEG (40,000 cells; both encoders frozen, identical inputs and identical probe). Figure 4 plots the same numbers.

		$\widehat{I}_{\text{probe}}(Z;\hat{n})$ (nats)		recon ÷ SPOT	
Nuisance	**Origin**	recon	SPOT	nats	share
Wavelet residual	denoiser	17.45	7.74	2.25×	1.25×
Wavelet-derived residual, re-injected	injected	7.27	0.61	11.89×	**5.86**×
EMG, 10 dB	injected	10.56	3.06	3.45×	1.46×
EMG, 6 dB	injected	11.46	5.59	2.05×	1.08×
EMG, 3 dB	injected	12.75	7.86	1.62×	0.97×

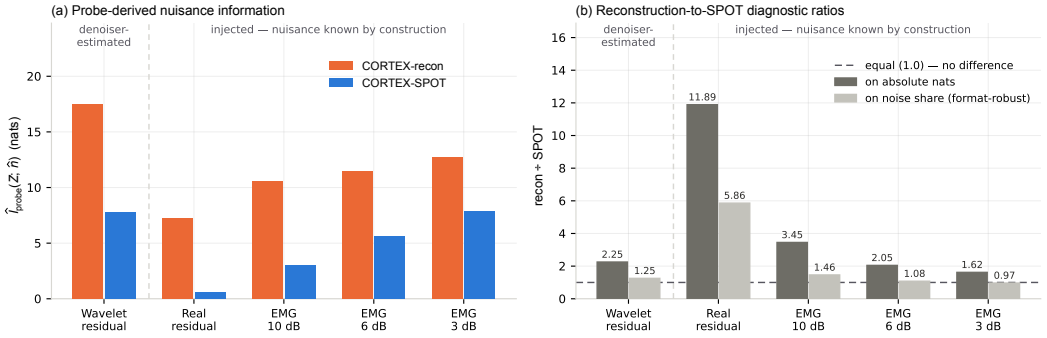


Figure 4: **The mechanism's signature on real recordings** (§4.2). The dashed rule separates denoiser-estimated from injected conditions. **(a)** Probe-derived nuisance information $\widehat{I}_{\text{probe}}(Z;\hat{n})$. **(b)** The recon ÷ SPOT ratio, in absolute nats and as a share.

The mean cosine similarity between domain-specific marginal distributions is 0.953, with 37 of 40 dimensions above 0.90 and a minimum of 0.717 (Figure 5). These results indicate active code utilization and similar marginal distributions across domains; they do not establish identical joint code usage or generalization to unseen sleep recordings.

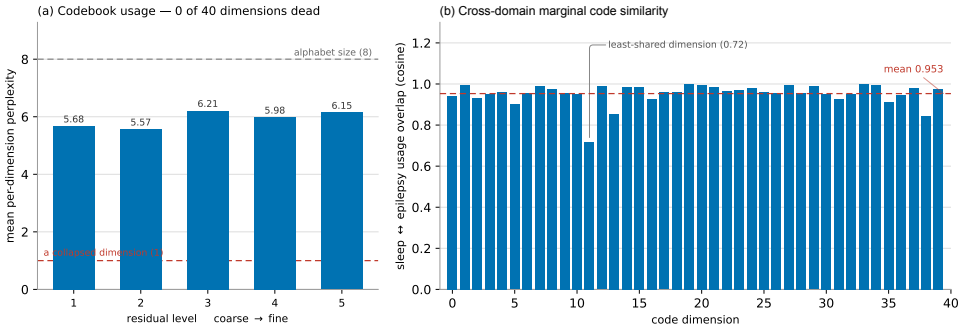


Figure 5: **Code utilization and cross-domain marginal similarity.** **(a)** Mean per-dimension perplexity at each of the five residual levels, bracketed by the alphabet size 8 above and the collapsed value 1 below. **(b)** Per-dimension cosine overlap between the code distributions the frozen encoder produces on sleep and on epilepsy recordings, mean dashed.

H.3 Dataset-specific evaluation details

The following differences in channel coverage, input duration, and montage mapping inform the comparisons in Tables 1 and 2.

Apnea and channel coverage. CORTEX processes respiratory and oxygen-saturation channels used for apnea detection. Comparisons with EEG-only baselines therefore reflect differences in available inputs as well as learned representations.

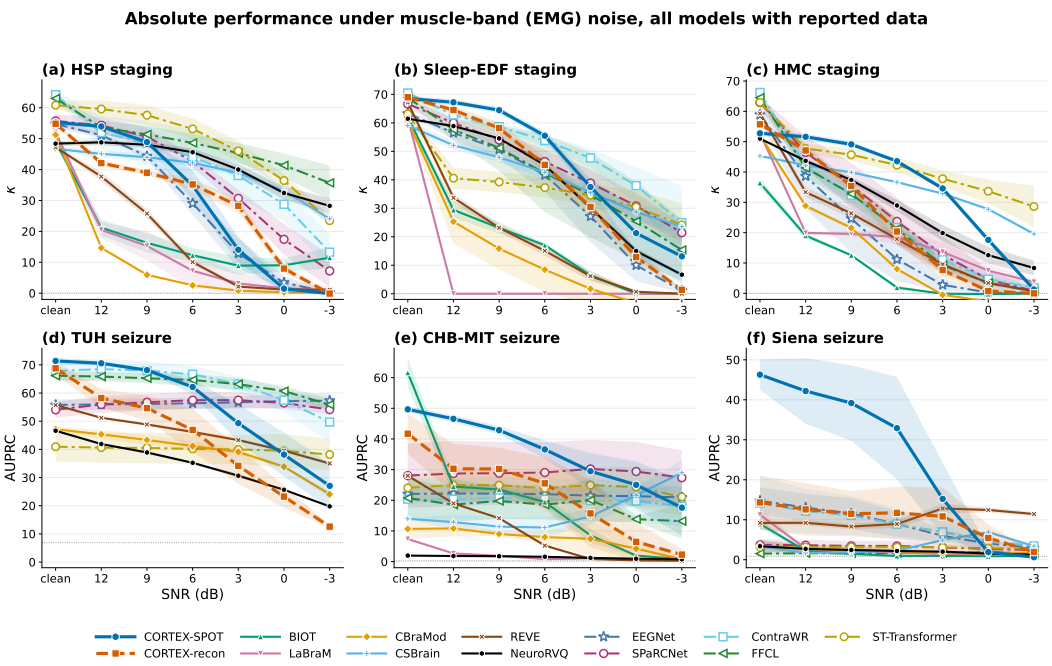


Figure 6: **Absolute performance under muscle-band contamination, sleep-staging and seizure-detection evaluations, every model with reported data** (13 models; the source of Table 2). **(a)** HSP staging. **(b)** Sleep-EDF staging. **(c)** HMC staging. **(d)** TUH seizure. **(e)** CHB-MIT seizure. **(f)** Siena seizure. The two CORTEX arms are heavy and solid; published foundation models read by a frozen probe are solid with thinner lines; end-to-end supervised baselines, trained on each target task rather than read by a frozen probe, are dotted. Colour and marker identify a model consistently across this and the two figures in §H.4. Shading is ± 1 seed sd. Dotted horizontal rule marks each panel's chance floor.

CHB-MIT input duration. BIOT uses 10-second windows, whereas CORTEX uses 3-second windows. This difference may contribute to the clean-performance gap. We retain each model's native input configuration, as described in §4.1.

NeuroRVQ on bipolar corpora. NeuroRVQ (Barmpas et al., 2025) uses spatial embeddings indexed by monopolar electrode names. We use the recorded electrode labels for HSP, HMC, TUH, and Siena, and map bipolar leads to supported labels for Sleep-EDF and CHB-MIT. Sleep-EDF uses FZ and POZ for EEG and FP1 for EOG; CHB-MIT uses an assignment to sixteen standard electrode names, selected over midpoint mapping on the development split. All evaluations retain the same frozen-probe protocol.

NeuroRVQ remains competitive on Sleep-EDF but achieves a clean AUPRC of 1.9 on CHB-MIT, whose positive prevalence is approximately 0.3%. This comparison involves both a corpus absent from its pretraining data and frozen-probe evaluation, whereas the original study reports full fine-tuning. Development-set checks of alternative scaling, the authors' preprocessing, additional filtering, and reconstruction of fourteen scalp potentials from bipolar chains produced no clear improvement. We therefore report the observed score while noting these differences in montage and evaluation setting.

H.4 Performance under pink and white noise

Tables 8 and 9 report absolute performance at 12 dB and averaged over the SNR grid. Figures 7 and 8 show the corresponding curves, using the same layout as Figure 6.

Across the six evaluations in Table 7, pink noise yields the smallest mean SPOT-to-reconstruction retention ratio and the smallest ratio on five evaluations. The exception is HSP staging, where white noise gives a slightly smaller ratio (1.14 versus 1.18). This ordering is consistent with the

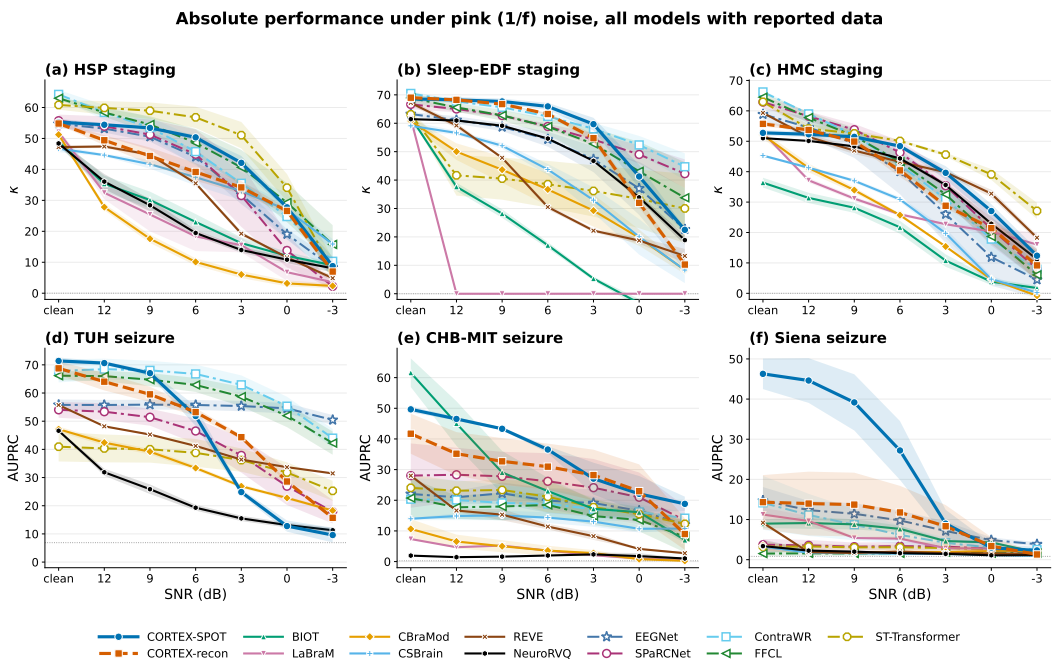


Figure 7: **Absolute performance under pink $(1/f)$ noise**, sleep-staging and seizure-detection evaluations, every model with reported data (the source of Table 8), in the same layout, colours and markers as Figure 6.

proposed role of contextual predictability, although Theorem 1 does not establish an ordering across contaminants.

Table 7: **Relative floor-adjusted retention across contaminants.** Values are the ratio of SPOT to reconstruction grid retention; values above one favor SPOT. The final column gives the arithmetic mean across six evaluations. Apnea is excluded because reconstruction retains only 1–2%, making the ratio unstable. Siena is excluded because reconstruction's clean score (6.9 ± 7.3 AUPRC) exceeds its floor (0.82) by less than one seed standard deviation.

Contaminant	HSP staging	Arousal	Sleep-EDF	HMC	TUH	CHB-MIT	**mean**
Muscle band (main text)	1.44	2.33	3.14	2.09	1.61	1.43	**2.01**
White	1.14	2.66	3.27	2.31	1.62	1.33	**2.05**
Pink $(1/f)$	1.18	1.52	2.21	1.48	1.31	1.22	**1.49**

Performance retention. For model m, we define floor-adjusted retention at SNR ρ as

$$r_m(\rho) = \frac{v_m(\rho) - b}{v_m^{\text{clean}} - b},$$

where $v_m(\rho)$ is the contaminated score, v_m^{clean} is the clean score, and b is the task-specific trivial-performance floor (Table 4). Retention equals one at clean performance and zero at the floor. We average retention over the six contaminated SNR levels $\{-3, 0, 3, 6, 9, 12\}$ dB. Table 7 reports the ratio of these averages for SPOT and reconstruction; values above one favor SPOT.

Under pink noise, SPOT achieves the highest grid-mean score among frozen encoders on six evaluations. REVE leads on HMC staging, while CORTEX-recon leads on TUH seizure detection (Table 8).

H.5 Additional contaminants and recorded muscle artifacts

Table 10 compares the two CORTEX arms on TUH seizure detection under two synthetic contaminants and expert-labelled muscle artifacts from TUAR.

Table 8: **Absolute performance under pink ($1/f$) noise**, in the format of Table 2: 12 dB / whole-grid mean, in each task's native metric. **Bold** marks the best result in each column among frozen-probe models. Values are computed to full precision from the same underlying sweeps as Table 2 and then rounded, rather than rounded twice.

	Staging			HSP		Seizure		
Model	HSP	Sleep-EDF	HMC	Arousal	Apnea	TUH	CHB-MIT	Siena
CORTEX-SPOT	**54.4/39.5**	68.2/**54.2**	52.3/38.5	**57.1/36.7**	**28.6/20.0**	**70.6**/39.5	**46.5/32.4**	**44.6/20.9**
CORTEX-recon	49.4/33.4	**68.2**/49.2	**53.7**/33.9	30.5/20.5	16.3/15.9	64.0/**44.2**	35.2/26.5	14.0/ 8.7
BIOT	35.4/21.0	37.7/12.7	31.4/16.3	21.5/16.5	15.4/14.2	22.5/16.3	45.0/22.9	9.1/ 6.1
LaBraM	32.5/17.0	0.0/ 0.0	37.2/25.6	13.5/14.4	11.6/ 5.4	30.9/14.8	4.7/ 2.8	9.6/ 4.6
CBraMod	27.8/11.2	50.0/31.3	41.2/20.0	27.1/16.8	11.4/ 9.4	42.4/30.5	6.6/ 3.2	1.9/ 2.0
CSBrain	44.6/33.3	56.6/35.6	41.2/22.2	23.5/18.4	12.9/ 8.0	50.2/38.7	14.9/13.1	1.7/ 1.4
REVE	47.4/27.2	59.2/32.0	51.4/**38.7**	28.7/22.5	16.0/15.6	48.2/39.4	16.7/ 9.8	1.9/ 1.7
NeuroRVQ[§]	36.1/19.5	61.0/45.7	50.1/35.4	25.9/20.5	16.0/14.3	31.9/19.5	1.4/ 1.7	2.3/ 1.6
EEGNet	53.2/34.5	61.1/46.9	54.6/31.1	23.9/21.9	15.7/13.5	55.9/54.8	21.9/19.3	12.4/ 8.3
SPaRCNet	53.9/32.9	65.0/55.3	58.3/37.2	49.3/28.0	12.6/ 7.5	53.3/38.9	29.0/24.1	3.6/ 3.1
ContraWR	58.4/38.0	67.7/58.5	59.0/35.0	50.5/33.7	19.9/17.1	68.9/61.3	21.4/18.7	11.1/ 6.0
FFCL	58.4/41.2	65.6/52.8	57.8/35.0	52.0/35.8	20.3/15.9	66.1/57.8	20.3/17.3	1.6/ 1.6
ST-Transformer	59.8/44.7	41.7/36.7	54.2/44.8	43.8/31.9	16.2/16.1	45.4/39.7	24.2/19.8	3.3/ 2.8

Table 9: **Absolute performance under white noise**, in the format of Table 2: 12 dB / whole-grid mean, in each task's native metric. **Bold** marks the best result in each column among frozen-probe models.

	Staging			HSP		Seizure		
Model	HSP	Sleep-EDF	HMC	Arousal	Apnea	TUH	CHB-MIT	Siena
CORTEX-SPOT	**53.7**/29.2	**67.6/44.0**	**51.8/31.2**	**55.3/31.7**	**28.6/19.4**	**70.8/48.3**	**46.6/31.1**	**45.7/22.8**
CORTEX-recon	39.4/25.3	65.8/37.8	48.1/19.1	25.8/18.9	16.5/15.9	57.1/37.1	31.9/21.2	14.3/ 9.0
BIOT	20.4/11.4	28.1/15.2	22.2/ 7.4	13.1/12.2	16.1/14.6	14.3/11.7	30.2/18.3	2.4/ 1.5
LaBraM	19.1/ 8.6	0.0/ 0.0	20.4/14.2	13.3/13.6	5.9/ 2.1	17.5/12.0	4.2/ 1.4	3.0/ 2.0
CBraMod	20.5/ 5.4	32.6/14.9	30.9/10.3	26.0/17.7	13.5/ 9.7	45.8/37.6	10.0/ 4.7	2.4/ 2.2
CSBrain	45.2/**40.0**	54.1/39.6	41.9/28.2	23.1/18.5	15.8/14.7	49.1/42.0	13.2/13.7	1.6/ 1.8
REVE	43.3/15.3	36.2/15.3	35.1/15.3	25.4/17.8	16.8/16.0	51.0/42.8	18.7/ 6.1	3.7/ 4.7
NeuroRVQ[§]	48.6/35.2	59.2/34.8	45.6/31.2	27.9/22.8	17.1/17.0	41.1/26.7	1.7/ 1.4	2.6/ 1.8
EEGNet	50.9/24.2	56.8/31.9	41.8/14.8	23.7/21.1	15.5/13.0	55.8/55.5	23.3/22.1	12.8/ 7.5
SPaRCNet	55.5/35.6	60.9/43.8	48.2/22.3	46.5/28.1	12.8/ 8.5	55.7/55.1	29.5/28.7	3.6/ 3.1
ContraWR	52.8/32.9	63.6/49.3	42.7/18.3	45.5/32.0	17.6/14.9	69.7/63.1	21.8/20.1	11.7/ 7.0
FFCL	54.0/43.3	59.3/40.0	42.8/18.9	46.9/32.6	17.6/14.6	65.8/59.2	23.1/18.2	1.6/ 1.5
ST-Transformer	59.9/46.1	41.2/35.5	48.8/40.3	44.1/33.8	16.3/16.3	45.8/43.5	25.0/24.1	3.2/ 3.1

Table 10: **TUH seizure performance differences under contamination.** Values report the AUPRC$\times$100 difference between SPOT and reconstruction, all three rows from the same matched CORTEX-SPOT/CORTEX-recon checkpoint pair. White and EMG results, at all three SNR levels, come from the same sweeps as Tables 9 and 3 respectively; TUAR results are a separate 3-seed sweep with expert-labelled muscle artifact from the TUH Artifact Corpus in place of synthetic noise.

Contaminant	**Origin**	Δ (SPOT $-$ reconstruction)		
		12 dB	9 dB	6 dB
TUAR muscle artifacts	recorded	$+7.4$	$+10.3$	$+10.1$
EMG-band noise	synthetic, 20–100 Hz	$+11.7$	$+11.8$	$+13.1$
White noise	synthetic, broadband	$+13.7$	$+15.8$	$+17.7$

SPOT achieves higher contaminated AUPRC in every reported condition. For recorded muscle artifacts, the advantage ranges from 7.4 to 10.3 points across the evaluated SNR levels. These results extend the comparison beyond synthetic noise.

I Additional objective-transfer results

This appendix extends the comparisons in §4.5 to all six evaluations under muscle-band, pink, and white noise. Figures 9, 10, and 11 show the full SNR curves for CORTEX and LaBraM.

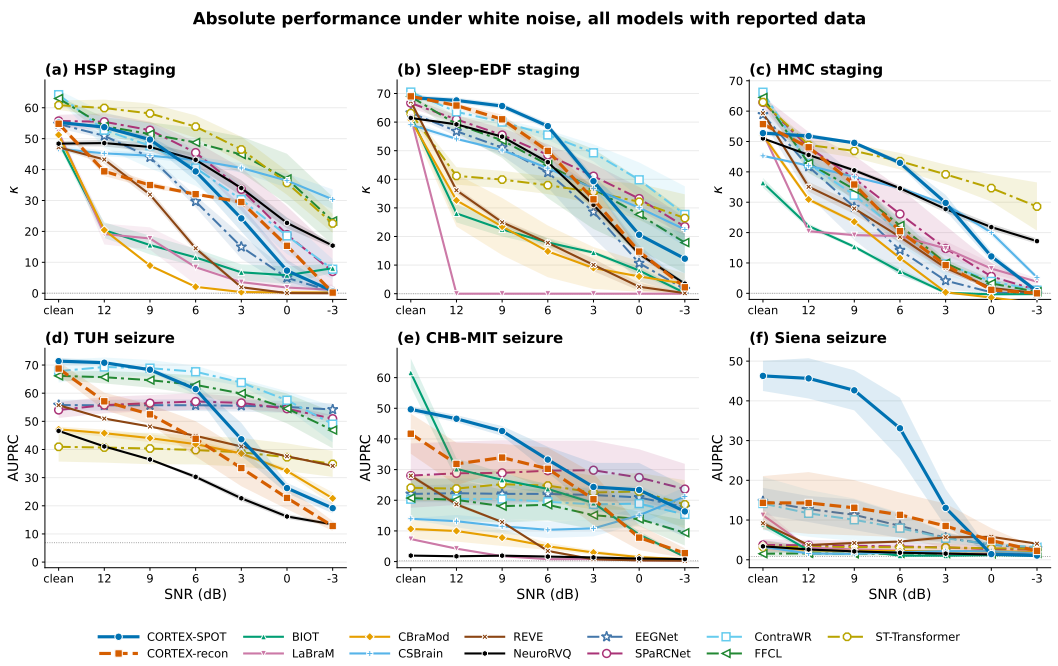


Figure 8: **Absolute performance under white noise**, sleep-staging and seizure-detection evaluations, every model with reported data (the source of Table 9), in the same layout, colours and markers as Figure 6.

Configuration checks. Within each backbone, the reconstruction and SPOT arms use the same encoder architecture, pretraining corpus, sampling order, training budget, and downstream probe protocol. The LaBraM encoders each contain 5,818,976 parameters, with matching `state_dict` keys and tensor shapes. Each pair of launch configurations differs only in the objective flag.

Results. Under muscle-band contamination, SPOT achieves higher grid-mean performance on all six tasks for both backbones and higher performance at 12 dB in eleven of the twelve comparisons (Table 3). Clean performance varies by task. LaBraM's apnea scores remain near the trivial macro-F1 floor, limiting interpretation of that comparison.

I.1 Model size

Table 11: **Model parameter counts.** The upper block lists pretrained encoders; the lower block lists supervised baselines. Supervised-model counts use the TUH input configuration (19 channels, 2,000 samples), as some architectures depend on input dimensions.

	BIOT	CBraMod	**CORTEX**	LaBraM	NeuroRVQ	CSBrain	REVE
Backbone params	3.44M	4.88M	**5.59M**	5.82M	5.87M	8.86M	69.19M

	EEGNet	SPaRCNet	ContraWR	FFCL	ST-Transformer
Params	0.003M	1.14M	1.57M	2.47M	3.43M

CORTEX has 5.59M parameters, making it the third smallest of the seven pretrained encoders (Table 11). The supervised baselines are smaller but update all parameters for each downstream task, whereas the pretrained encoders are evaluated with frozen representations. The matched comparisons in §4.5 assess the pretraining objective while holding encoder size fixed.

J Additional discussion of limitations

Scope of the theoretical analysis. Theorem 1 assumes jointly Gaussian variables, additive noise, and a basis that simultaneously diagonalizes the signal and noise covariances. These assumptions

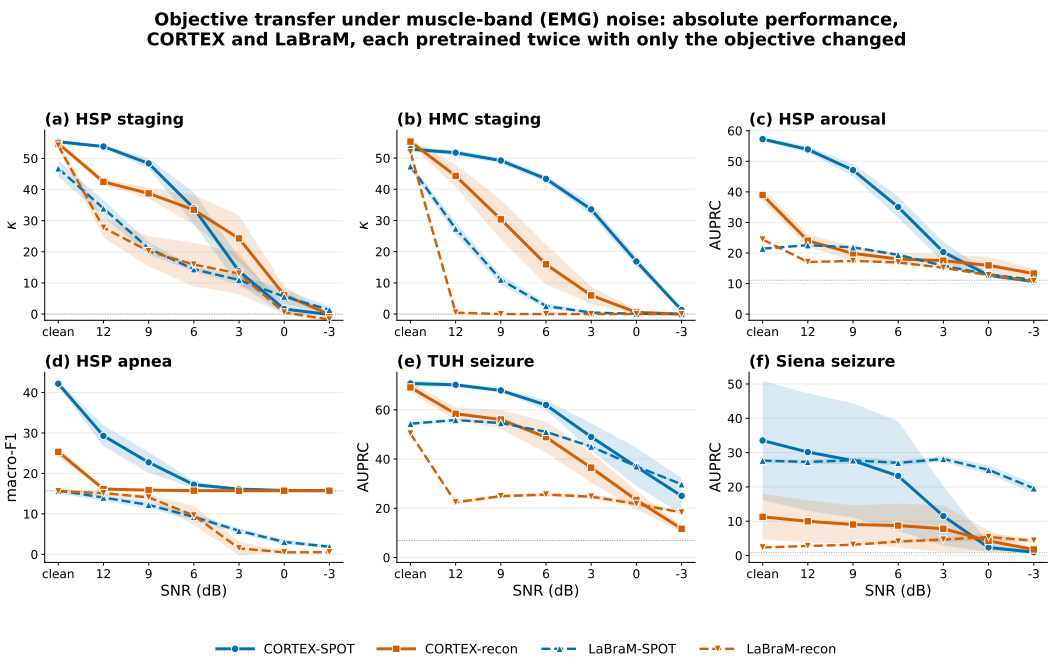


Figure 9: **Objective transfer under muscle-band noise.** Full SNR curves for both backbones across six evaluations. Blue denotes SPOT and vermilion denotes reconstruction; solid lines denote CORTEX and dashed lines denote LaBraM. Shading indicates ± 1 seed standard deviation. Table 3 reports the corresponding summary scores.

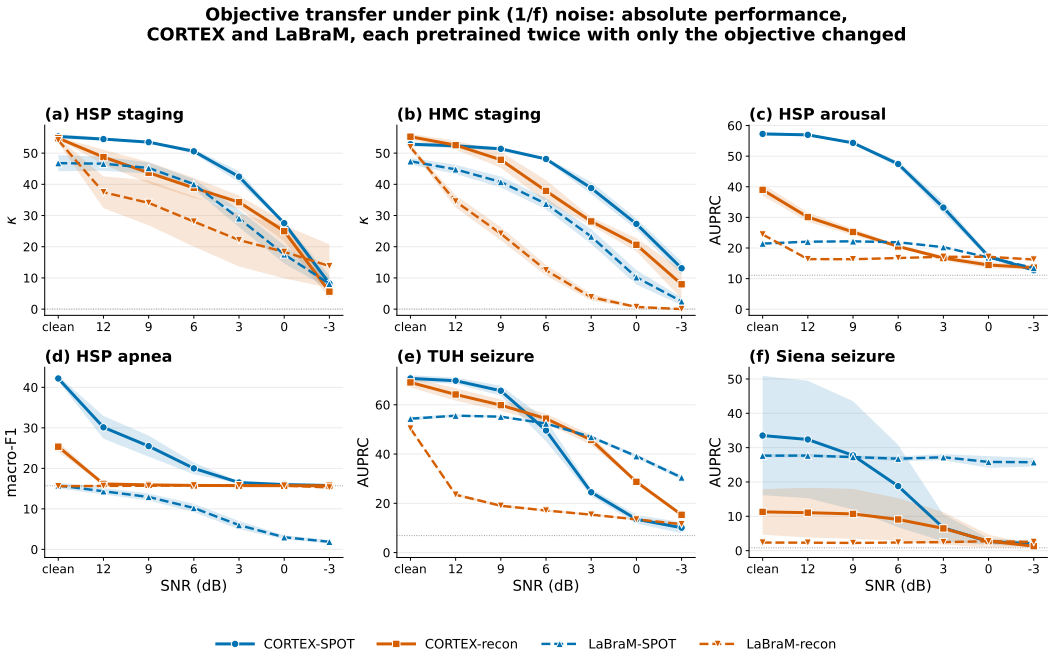


Figure 10: **Objective transfer under pink** $(1/f)$ **noise.** Evaluations and plotting conventions follow Figure 9.

simplify the structure of physiological recordings. The analysis characterizes a rate-allocation mechanism under this model; its quantitative predictions do not directly extend to real EEG. Moreover, part (i) establishes the existence of a predictive-optimal code that excludes predictive-null components, without guaranteeing that a trained encoder reaches this optimum (Remark 1). The theorem

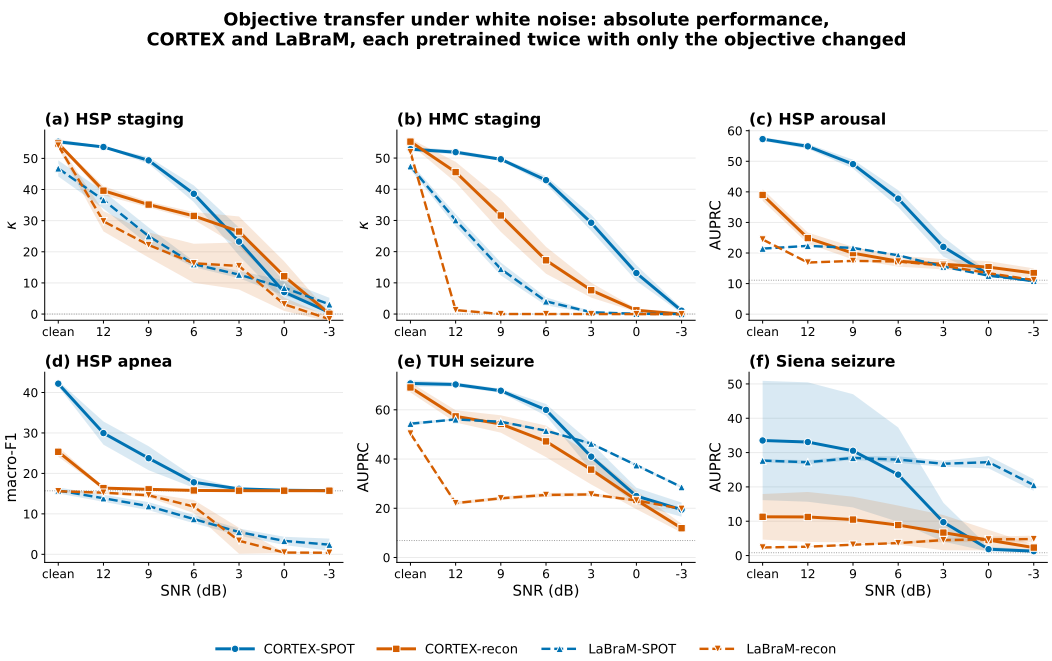


Figure 11: **Objective transfer under white noise.** Evaluations and plotting conventions follow Figure 9.

also does not establish downstream robustness or the removal of noise within context-predictive directions.

Architectural coverage and statistical variability. The controlled objective-transfer experiments cover CORTEX and LaBraM, providing evidence that the observed effect extends beyond the proposed encoder. The five-seed comparisons vary the downstream probe while holding a single pretrained checkpoint fixed for each arm. The resulting variability and significance tests therefore characterize probe training conditional on these checkpoints, rather than variation across independent pretraining runs.

Sensitivity to design choices. We do not independently ablate the candidate count K, FSQ depth, EMA decay, or context length. We also do not isolate the contribution of tube masking from that of the selective objective. The experiments therefore evaluate the specified SPOT configuration, while leaving the contribution of individual components and sensitivity to their settings unresolved.

Dependence on the contamination regime. The robustness advantage depends on the evaluation and contamination level. On HSP staging, CORTEX-SPOT leads at 12 dB, whereas NeuroRVQ achieves the highest grid-mean score (Table 2). Thus, performance at moderate contamination does not establish superiority across the full evaluated range.

Clinical applicability. Our evaluations use retrospective datasets with synthetic or injected contamination. Although the recorded-artifact experiments complement the synthetic tests, they do not capture the full range of artifacts and acquisition conditions encountered in clinical practice. Benchmark performance under these conditions does not establish diagnostic validity or reliability in prospective deployment. Performance across demographic groups and recording systems also requires further evaluation.