LUCID: FROZEN UNSUPERVISED DIFFUSION FEATURES FOR EEG DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models pretrain on large corpora to learn general representations that transfer to downstream decoding tasks. Yet they are typically evaluated after full fine-tuning, entangling what pretraining encodes with what task labels supply. To separate the two, we present *LUCID* (Latent Unsupervised Characterisation from Internal Diffusion): an unsupervised diffusion model trained from scratch on each dataset, then frozen, with its internal activations read by a small linear probe. Two properties of EEG shape *LUCID*'s denoiser. Along the time axis, the signal is densely sampled and autocorrelated, with rhythms spanning tens of milliseconds to seconds; patching or tokenizing it, as most foundation models do, coarsens these dynamics. Across electrodes, channel power and inter-channel coupling shift from window to window – a nonstationary covariance that no fixed spatial operator can follow. We therefore build a time-preserving denoiser from structured long convolutions and, within each block, introduce **CoSMO**, a spatial-modulation operator that adapts to each window's own covariance. We evaluate *LUCID* on ten EEG decoding tasks across twelve datasets against seven task-specific supervised networks, a diffusion-based representation model, and six fully fine-tuned foundation models. On five of the twelve datasets, its frozen features tie the best of these foundation models. Re-evaluating each foundation model frozen, with only its head trained, reveals a pattern: *LUCID* ties where these models gain little from fine-tuning and trails where they gain most. It is also label-efficient, matching its full-label accuracy with a quarter of the labels on half of the datasets and keeping at least four fifths of it on the rest. These findings argue for evaluating EEG representations frozen, not only after full fine-tuning. Beyond decoding, *LUCID* generates synthetic EEGs that capture much of the spectral, temporal, and spatial structure of real recordings. Our source code is made available, anonymized.[1]

1 INTRODUCTION

Driven by large open corpora (Obeid & Picone, 2016) and by architectures carried over from vision and language, foundation models for electroencephalograms (EEG) – BIOT (Yang et al., 2023a), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025a), CSBrain (Zhou et al., 2026), REVE (El Ouahidi et al., 2026) and CodeBrain (Ma et al., 2026) – promise a single pretrained backbone that transfers across subjects and tasks. With the architectures came their training recipe: the waveform is discretised into patches or codebooks and a Transformer is trained by masked reconstruction or contrastive prediction. This models relations between discrete chunks rather than the continuous evolution of the signal, and masking parts of a time series can destroy the fine temporal variations it carries (Dong et al., 2023).

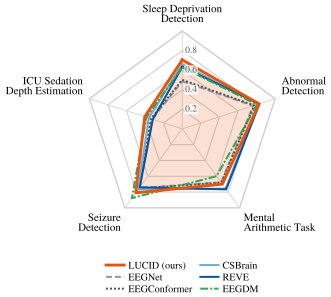


Figure 1: Balanced accuracy on five tasks (others in Table 2).

Generation offers a different pretext. A diffusion model (Ho et al., 2020) corrupts the whole multichannel signal and learns to reverse it, so it must capture the recording's continuous, multi-scale

[1] https://anonymous.4open.science/r/LUCID-AF74

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.

structure. In vision, a frozen denoiser's activations are linearly separable (Xiang et al., 2023) and serve as segmentation and classification features (Baranchuk et al., 2021; Mukhopadhyay et al., 2023). In EEG, diffusion has largely stopped at generation (Vetter et al., 2024; Torma & Szegletes, 2023) and augmentation (Huang et al., 2025; Wang et al., 2024; 2026). Only recently have generative features served EEG decoding: in B[FM]^2, a flow-matching foundation model (Hwang et al., 2026), and in two EEGDM studies, one training a large latent-fusion Transformer on frozen diffusion features (Puah et al., 2025), the other fine-tuning the encoder of a latent diffusion model (Wang et al., 2025b). We continue this line but keep the representation label-free: the denoiser trains without labels and stays frozen, and we ask whether it is linearly readable as it is. Further related work is in Appendix A.

What none of this settles is what the pretraining itself captured. All of these models reach their downstream number through labels – by fine-tuning, or, for EEGDM, a head trained between backbone and task – and published tables report almost only fine-tuned rows, the rare frozen tier sitting behind a trained head, as in REVE (El Ouahidi et al., 2026). This conflates what a backbone *encodes* with what a *trained head* supplies, and fine-tuning can itself distort the features it started from (Kumar et al., 2022). The question is therefore open on EEG. The answer matters most where labels are scarce and the signal is noisy: in clinical EEG, long resting-state recordings carry signatures of sedation (Purdon et al., 2013) and of consciousness (Sitt et al., 2014), yet a cohort numbers a few dozen patients, each recording is labelled once by a clinician, and a model must generalise to patients it never saw.

To settle it we introduce **LUCID** (Latent Unsupervised Characterisation from Internal Diffusion): one unconditional diffusion denoiser per dataset, trained from scratch and without labels on that dataset's own recordings, then frozen. All its blocks' activations, pooled over time and concatenated, are read at one noise level by a small ℓ_2-regularised logistic-regression probe (Section 2, Figure 2). Two properties of EEG shape the denoiser. Along the time axis the signal is densely sampled and autocorrelated, with rhythms spanning tens of milliseconds to seconds, which patching or tokenizing coarsens. Across electrodes, channel power and inter-channel coupling – the covariance that is the classic descriptor of EEG spatial structure (Barachant et al., 2011) – drift over trials, sessions and subjects (Shenoy et al., 2006; Zanini et al., 2017), a nonstationary covariance beyond the reach of any fixed spatial operator. We therefore build a time-preserving denoiser from structured long convolutions (Li et al., 2022b) and, within each block, introduce **CoSMO**, a spatial-modulation operator that adapts to each window's own covariance. By design, LUCID is not a foundation model and claims no cross-dataset transfer. An untrained copy of the same network, every block of which starts as the identity, isolates what training added. We call this control the *identity-map null*.

On ten EEG decoding tasks across twelve datasets, under a strict patient-disjoint protocol – among them a private ICU cohort of comatose patients (Section 3) – we compare LUCID with six foundation models under two read-outs, backbone frozen with only its own head trained (Y) or fully fine-tuned (FT), together with seven task-specific supervised networks, EEGDM with its trained head and the identity-map null. LUCID's unsupervised frozen features lie inside the tie band of the best fully fine-tuned foundation model on five of the twelve datasets (Figure 1) and stand 0.04–0.34 in balanced accuracy above its untrained copy on all twelve, while trailing the best fine-tuned model by up to 0.38 on the rest. Where fine-tuning adds a foundation model little, LUCID matches it. Where fine-tuning adds most, LUCID falls behind. This *fine-tuning gain*, fine-tuned minus frozen accuracy, orders the twelve datasets: an ordering, not a law. LUCID is also label-efficient, keeping most of its full-label accuracy with a quarter of the labels on every dataset (Section 3.5). Our contributions:

- We introduce **LUCID**, an unsupervised generative pretraining framework for EEG decoding that freezes a denoiser trained on each dataset's own recordings and reads all its blocks with a small linear probe, and we propose **CoSMO**, a covariance-conditioned spatial-modulation operator that adapts every block to each window's covariance.
- We show that training writes a linearly readable representation: against an exact identity-map null, LUCID gains on all twelve datasets (Section 3.2) and not through feature width (Section 3.4). It ties the best fully fine-tuned foundation model where fine-tuning adds little (Section 3.3).
- We evaluate six foundation models both frozen and fine-tuned on the same splits of twelve datasets, and show that their fine-tuning gain orders where LUCID ties and trails (Section 3.3, Figure 4). Fine-tuned results come from our protocol on some datasets and from published reports on others, a provenance confound we confront.
- We compare the generated EEGs with each dataset's real recordings on spectral, temporal and spatial statistics and show that they are faithful to them (Section G).

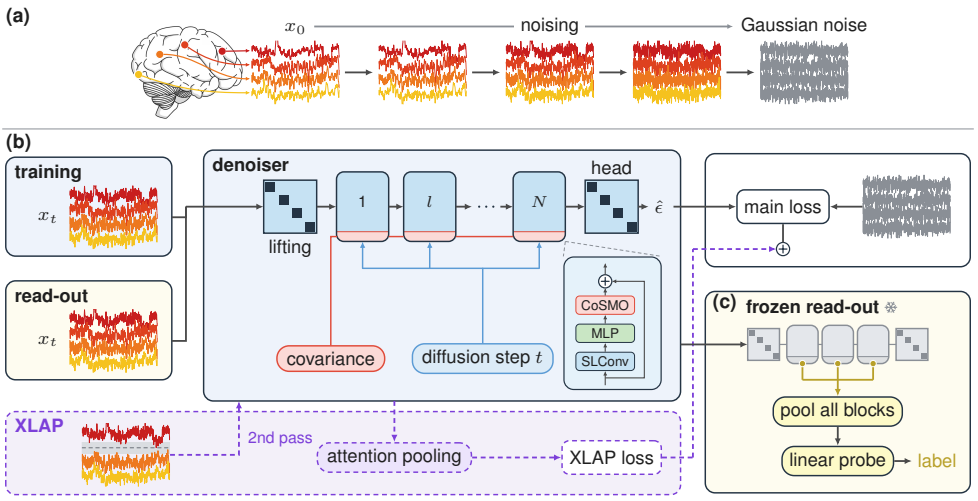


Figure 2: **Diffusion pretraining and frozen read-out with LUCID. (a)** EEG windows are corrupted with Gaussian noise, up to pure noise. **(b)** One unconditional denoiser per dataset predicts the noise from the noisy window x_t and the diffusion step t: a per-electrode convolution lifts the window, N residual blocks mix time and electrodes, keeping that shape, and a per-electrode head returns $\hat{\epsilon}_\theta(x_t, t)$; every block reads the window's covariance through CoSMO. The dashed band is XLAP, a training-only objective: a second pass on a channel-masked window pools all blocks with one learned query to predict the masked channels' covariance; its loss adds to the main loss. **(c)** The denoiser is frozen; all block outputs at one noise level are pooled over time and read by a small linear probe.

2 METHOD

LUCID trains one unconditional diffusion denoiser per dataset, freezes it, and reads all its blocks with a small linear probe (Figure 2). Its two new elements are **CoSMO**, a covariance-conditioned spatial modulation operator inside every block of a time-preserving backbone (Section 2.2, Figure 3), and **XLAP**, a training-only objective that spreads information across depth (Section 2.3).

2.1 PRELIMINARY: DIFFUSION PRETEXT

A window is $x_0 \in \mathbb{R}^{C\times T}$, C electrodes by T samples, normalised per dataset (Section B). A denoising diffusion probabilistic model (Ho et al., 2020) corrupts it with same-shape Gaussian noise ϵ and trains a network $\hat{\epsilon}_\theta$ to recover that noise,

$$\mathcal{L}_{\text{main}} = \left\| \epsilon - \hat{\epsilon}_\theta(x_t, t) \right\|_2^2, \qquad x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad t \sim \mathcal{U}\{1, \ldots, T_{\text{diff}}\}, \tag{1}$$

where $\bar{\alpha}_t$ decreases along a fixed variance schedule of T_{diff} steps, so x_t, the window corrupted to level t, is reached in one draw. The model is unconditional: no label, subject or electrode-position information enters, each denoiser is trained from scratch, and besides x_t and t the network sees only two statistics of x_t itself (Section 2.2.2).

2.2 TIME-PRESERVING DENOISER BACKBONE

The denoiser preserves the time axis (Figure 3a). A per-electrode 1×1 convolution (`conv_in`) lifts each electrode's trace to m hidden features, so $H = Cm$; N identical residual blocks keep the shape (H, T); and a per-electrode head (`conv_out`) returns the noise estimate $\hat{\epsilon}$. Nothing is patched, strided or downsampled: every block's output $h^{(l)} \in \mathbb{R}^{H\times T}$ is sample-aligned with the input and can be read (Section 2.4); as no layer is tied to T, the same network applies to windows of any length.

Each block follows the Diffusion Transformer block (Peebles & Xie, 2023), with self-attention replaced by SLConv, in two sub-layers (Section B): time mixing gives each hidden feature its own long convolution; electrode mixing combines hidden features, and with them electrodes, in three gated branches, a pointwise MLP and CoSMO's two window-specific terms. Each sub-layer

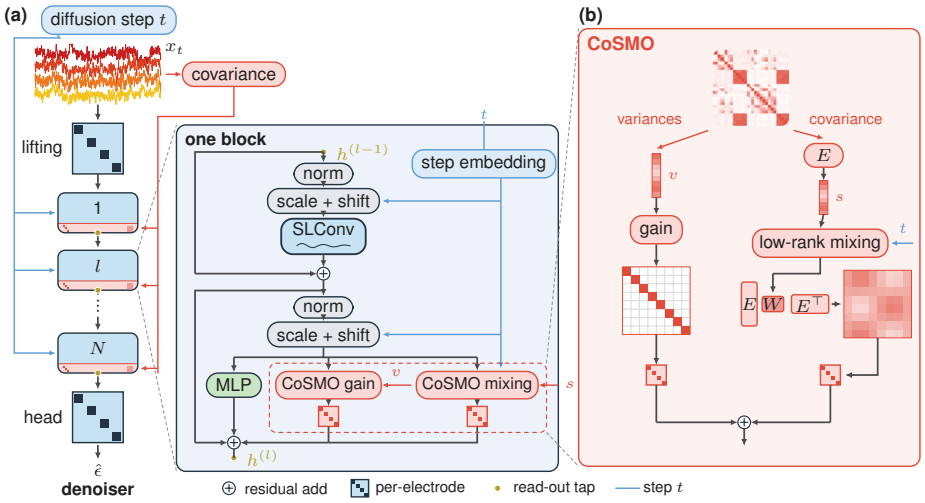


Figure 3: **The LUCID denoiser. (a)** The network of Figure 2b: a per-electrode convolution lifts the noisy window, N residual blocks keep that shape, and a per-electrode head returns the noise estimate; the input's covariance, reduced once to two statistics, and the diffusion step t feed every block. Inside a block, time mixing applies a structured long convolution and electrode mixing a pointwise MLP and CoSMO; every branch is gated, with gates initialised at zero so that each block starts as the identity. **(b)** CoSMO reads two statistics of the input covariance, the per-electrode variances and its projection through a small learned electrode embedding, and turns them into a per-electrode gain and a low-rank electrode mixing; the mixing also reads t. XLAP is drawn in Figure 2.

normalises its input, applies a scale and shift, and adds each branch back through its own gate, all zero-initialised maps of the step embedding τ_t. As in adaLN-Zero, each block therefore starts as the identity and the untrained network computes $\texttt{conv_out}(\texttt{conv_in}(x_t))$, a per-electrode affine map: the identity-map null (Section 2.4).

2.2.1 Time mixing: structured long convolutions

Densely sampled neurophysiological series carry dependencies over hundreds of samples within a second, beyond ordinary convolution kernels. Structured long convolutions (SLConv), introduced by Li et al. (2022b) as structured global convolutions, close that gap; Vetter et al. (2024) use them, under this name, as the temporal operator of a diffusion model for neurophysiological recordings; across the stack the receptive field covers the whole window.

2.2.2 Electrode mixing: CoSMO

The MLP's map is the same for every window; the spatial structure it faces is not. The channel covariance is the classic descriptor of EEG spatial structure (Blankertz et al., 2008; Barachant et al., 2011); its drift over trials and sessions (Shenoy et al., 2006) and across subjects (Zanini et al., 2017; He & Wu, 2019) is what adaptive classifiers correct for and alignment methods reduce before decoding. A denoiser has reason to follow that drift: for correlated electrodes the optimal noise estimate depends on their covariance, so the denoising loss alone can reward its use. CoSMO therefore turns the covariance into a conditioning signal, in the spirit of feature-wise modulation (Perez et al., 2018; Peebles & Xie, 2023) and of channel gains from an input's own pooled statistics (Hu et al., 2018), adding to every block's MLP a gain term $\mathcal{A}$ and a mixing term $\mathcal{C}$. Once per forward pass, the network reduces the covariance of its own input x_t to two quantities shared by all blocks (Figure 3b): $v \in \mathbb{R}^C$, each electrode's log-variance standardised across electrodes, and $s \in \mathbb{R}^{d_E^2}$, the covariance scaled by its mean variance, $\bar{\Sigma}$, projected through a small learned electrode embedding $E \in \mathbb{R}^{C \times d_E}$ shared by the network, then standardised, so s has a fixed size whatever C. Both terms act on the electrode axis of $\hat{Y} \in \mathbb{R}^{C \times mT}$, the sub-layer's normalised, scaled and shifted input:

$$\mathcal{A}(\hat{Y}) = \operatorname{diag}\big(a(v)\big)\,\hat{Y}, \qquad \mathcal{C}(\hat{Y}) = E\,W(s,\tau_t)\,E^\top \hat{Y}. \tag{2}$$

In each block a small network maps v to a gain $a \in \mathbb{R}^C$, scaling each electrode by the variance pattern of all electrodes; another maps s and τ_t to a window-specific update of a learned base, giving $W(s, \tau_t) \in \mathbb{R}^{d_E \times d_E}$, which E frames into a low-rank electrode mixing identical at every feature and sample. W reads the step because the covariance of x_t blends the window's with the noise floor in proportions set by t; the gain reads v alone. CoSMO adds nothing at initialisation and grows only as Equation (4) rewards it. Each block thus re-weights and re-couples electrodes per window, as a dynamic hypernetwork emits weights (Ha et al., 2016), rather than through spatial filters fixed after training (Lawhern et al., 2018).

2.3 XLAP: A TRAINING-ONLY CROSS-LAYER OBJECTIVE

A read-out that concatenates every block wants information spread across depth, not concentrated in the last blocks; yet a masked autoencoder's last layers specialise to reconstruction when its decoder is shallow (He et al., 2022). REVE (El Ouahidi et al., 2026) counters this in a Transformer with a secondary loss: a learned query attention-pools the outputs of all encoder layers and a small head reconstructs the masked patches from the pooled token, so each layer receives its own gradient path through one shared head rather than the per-layer heads of deep supervision (Lee et al., 2015). XLAP adapts this loss to a convolutional denoiser (Figure 2), which masks nothing, so it first creates something to predict: at every training step each channel of x_t is zeroed independently with a fixed probability and the same network makes a second pass on the masked window at the same t.

The backbone has no tokens to pool, so each block's output in that pass is summarised over time as in the read-out (Section 2.4); a learned query attends over the N summaries (Lee et al., 2019) and a two-layer head maps the pooled vector to a predicted covariance $\hat{\Sigma} \in \mathbb{R}^{C \times C}$ and log-variances $\hat{\nu} \in \mathbb{R}^C$, whose targets $\bar{\Sigma}^0$, ν^0 are CoSMO's $\bar{\Sigma}$, before its projection through E, and v on the clean window. As in masked feature prediction (Wei et al., 2022), the target is a descriptor of the hidden part, not its raw values; here it keeps the task well posed: a zeroed channel's second-order statistics are absent from the input and must be inferred from the visible ones. An L1 loss therefore scores only what the mask hides,

$$\mathcal{L}_{\mathrm{XLAP}} = \left\langle \left| \hat{\nu}_c - \nu_c^0 \right| \right\rangle_{c \in \mathcal{M}} + \left\langle \left| \hat{\Sigma}_{cc'} - \bar{\Sigma}^0_{cc'} \right| \right\rangle_{c \in \mathcal{M},\, c' \in \mathcal{V}}, \tag{3}$$

with $\mathcal{M}$, $\mathcal{V}$ the masked and visible channels and $\langle \cdot \rangle$ the set mean. XLAP acts only through the second pass's gradient; off outside training and its head discarded, the denoiser keeps the parameters and inference cost of one trained without it.

2.4 OBJECTIVE, GENERATION AND READ-OUT

Objective. The objective, minimised with AdamW (Loshchilov & Hutter, 2019), is

$$\mathcal{L} = \mathcal{L}_{\mathrm{main}} + \lambda \, \mathcal{L}_{\mathrm{XLAP}}, \tag{4}$$

up to a fixed time-dependent weighting (Section B); λ is set per dataset by gradient balance, equal norms of the XLAP and denoising gradients on the parameters both losses share at a trained checkpoint, and never tuned on downstream results. We keep the final weights. Samples come from ancestral DDPM sampling over all T_{diff} steps (Ho et al., 2020) from Gaussian noise (Section B) and are checked against real windows on spatial and spectral metrics before any feature is read (Section G).

Feature extraction and read-out. A frozen denoiser's activations at chosen blocks and timesteps carry semantic information, shown for image segmentation (Baranchuk et al., 2021); in vision both choices are made by grid search (Baranchuk et al., 2021; Mukhopadhyay et al., 2023; Xiang et al., 2023) or learned, by a trainable model or a learning agent (surveyed in Fuest et al., 2026). Ours spans every block, so only the noise level remains. The frozen denoiser runs on the dataset's windows, normalised as in training and noised to level t with one Gaussian draw per window, fixed across levels as in Baranchuk et al. (2021). Every block's output is pooled over time to its mean and standard deviation. The N summaries, concatenated and standardised with training statistics, feed a small logistic-regression probe with inverse ℓ_2 penalty C. Both choices are made per dataset on validation: the noise level $t^\star$ first, then the penalty $C^\star$ at that level; each probe is then fit on the training split and scored once on test (grids and seeds: Section 3); the identity-map null of Section 2.2 is read identically.

Table 1: Overview of the EEG decoding tasks and datasets.

Task	Dataset	# Channels	Duration	# Samples	Rate	# Classes
ICU Sedation Depth Estimation	COMA	11	8 s	32,685	256 Hz	4
Sleep Deprivation Detection	SLEEP	58	8 s	4,615	250 Hz	2
Abnormal Detection	TUAB	16	10 s	409,455	250 Hz	2
Event Type Classification	TUEV	16	5 s	113,353	250 Hz	6
Mental Arithmetic Task	MAT	20	5 s	1,707	500 Hz	2
Motor Imagery	BCIC-IV-2a	22	4 s	5,184	250 Hz	4
	PhysioNet-MI	64	4 s	9,837	160 Hz	4
Mental Disorder Diagnostic	Mumtaz	19	5 s	7,143	256 Hz	2
Emotion Recognition	SEED-V	62	1 s	117,744	1000 Hz	5
	FACED	32	10 s	10,332	250 Hz	9
Imagined Speech Classification	BCIC2020-3	64	3 s	6,000	256 Hz	5
Seizure Detection	Siena	29	10 s	51,307	512 Hz	2

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Decoding tasks and datasets. We evaluate on ten EEG decoding tasks across twelve datasets, spanning sedation depth, sleep deprivation, abnormality, event type, mental arithmetic, motor imagery, mental disorder, emotion, imagined speech and seizure detection (Table 1). On the ten datasets shared with prior work we adopt the preprocessing and train/validation/test splits of BIOT (Yang et al., 2023a), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025a) and REVE (El Ouahidi et al., 2026), so that our results are directly comparable; on the two datasets new to this comparison, COMA (Bouchereau et al., 2023; 2025) and SLEEP (Xiang et al., 2024), splits are patient- or subject-disjoint. Task definitions are in Section C.1, dataset specifics and preprocessing in Section D.

Baselines. We compare LUCID against three groups (Table 2). Seven task-specific networks trained from scratch on each dataset: EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2022), SPaRCNet (Jing et al., 2023), ContraWR (Yang et al., 2023b), CNN-Transformer (Peh et al., 2022), FFCL (Li et al., 2022a) and ST-Transformer (Song et al., 2021). Six pretrained EEG foundation models, BIOT (Yang et al., 2023a), LaBraM-Base (Jiang et al., 2024), CBraMod (Wang et al., 2025a), CSBrain (Zhou et al., 2026), REVE-Base (El Ouahidi et al., 2026) and CodeBrain (Ma et al., 2026), each read two ways: backbone frozen with only its own head trained (Y) and fully fine-tuned (FT)[2]. And EEGDM (Puah et al., 2025), the diffusion model closest to ours, whose frozen denoiser features feed a 6.9M-parameter head trained on each task[3]. An untrained copy of our denoiser, the identity-map null of Section 2, completes the table. Fine-tuned and supervised numbers are taken from REVE (El Ouahidi et al., 2026), CSBrain (Zhou et al., 2026) and EEGDM (Puah et al., 2025) where published, on the same splits we use. The frozen (Y) rows, every row on COMA and SLEEP, and the rows those papers do not cover are computed by us (Section C.2).

Metrics. All results are means $\pm$ standard deviation over five random seeds. The main text reports balanced accuracy. Section D adds AUC-PR and AUROC for *binary tasks* or Cohen's κ and weighted F1 for *multi-class tasks*; the metrics are defined in Section C.3. Validation balanced accuracy selects the noise level LUCID's probe reads and the probe's penalty.

3.2 MAIN RESULTS

Performance comparison with baselines. In Table 2, gaps within one seed standard deviation of either row are ties. Read this way, LUCID's frozen features, through a linear probe, tie the best fully fine-tuned foundation model on five datasets, COMA, SLEEP, TUAB, Siena and MAT, with the highest mean of the COMA and SLEEP columns, and trail it on the seven others, narrowly on

[2]The flow-matching foundation model B[FM][2] (Hwang et al., 2026) is not reported: its code had not been released at the time of writing.

[3]A second EEGDM (Wang et al., 2025b), built on a latent diffusion model, has no public code.

Table 2: Balanced accuracy (± std over five seeds) on the twelve EEG decoding tasks; best per dataset in bold, second best underlined. Foundation models are fully fine-tuned; their frozen-backbone results are in Figure 4. *Params*: backbone size (the pretrained encoder for foundation models and EEGDM, the frozen denoiser for LUCID, the whole network for supervised models), as a range over the datasets when it depends on their input shape.

	Methods	Params	COMA	SLEEP	TUAB	TUEV	MAT	BCIC-IV-2a
Task-specific Models	EEGNet	0.002–0.004M	0.3228±0.0283	0.4883±0.0176	0.7642±0.0036	0.3876±0.0143	0.6770±0.0116	0.4482±0.0094
	EEGConformer	0.6–1.5M	0.3363±0.0387	0.5019±0.0024	0.7758±0.0049	0.4074±0.0164	0.6805±0.0123	0.4696±0.0106
	SPaRCNet	1.0–1.2M	0.3419±0.0175	0.6267±0.0124	0.7896±0.0018	0.4161±0.0262	0.6879±0.0107	0.4635±0.0117
	ContraWR	1.6M	0.2957±0.0225	0.6024±0.0523	0.7746±0.0041	0.4384±0.0349	0.6631±0.0097	0.4678±0.0125
	CNN-Transformer	3.2M	0.3134±0.0285	0.5966±0.0305	0.7777±0.0022	0.4087±0.0161	0.6779±0.0268	0.4600±0.0108
	FFCL	2.4–2.5M	0.3108±0.0210	0.5967±0.0300	0.7848±0.0038	0.3979±0.0104	0.6798±0.0142	0.4470±0.0143
	ST-Transformer	3.4–3.5M	0.3748±0.0173	0.5967±0.1018	0.7966±0.0023	0.3984±0.0228	0.6631±0.0173	0.4575±0.0145
Foundation Models	BIOT	3.2M	0.4022±0.0161	0.5801±0.0233	0.7959±0.0057	0.5281±0.0225	0.6875±0.0186	0.4748±0.0093
	LaBraM-Base	5.8M	0.3977±0.0429	0.6700±0.0528	0.8140±0.0019	0.6409±0.0065	0.6909±0.0125	0.4869±0.0085
	CBraMod	4.9M	0.3362±0.0118	0.5795±0.0452	0.8289±0.0022	0.6671±0.0107	0.7256±0.0132	0.5138±0.0066
	CSBrain	8.9M	0.3603±0.0139	0.6624±0.0193	0.8172±0.0043	0.6903±0.0059	0.7558±0.0106	0.5657±0.0071
	REVE-Base	69.2M	0.3167±0.0157	0.6351±0.0214	**0.8315±0.0014**	0.6759±0.0229	**0.7660±0.0355**	**0.6396±0.0095**
	CodeBrain	15.2M	0.3728±0.0104	0.6404±0.0373	0.8294±0.0013	0.6428±0.0062	0.7514±0.0203	0.3847±0.0167
Diffusion Features	EEGDM (supervised)	5.9M	0.3865±0.0039	0.6312±0.0193	0.8051±0.0093	**0.7557±0.0147**	0.5986±0.0545	0.2700±0.0419
	LUCID (ours, untrained)	2.4–15.4M	0.2500±0.0000	0.5793±0.0000	0.4870±0.0008	0.3338±0.0000	0.5243±0.0000	0.4254±0.0000
	LUCID (ours, trained)	2.4–15.4M	**0.4038±0.0092**	**0.7071±0.0077**	0.8283±0.0042	0.5744±0.0104	0.7000±0.0731	0.5302±0.0234

	Methods	Params	PhysioNet-MI	Mumtaz	SEED-V	BCIC2020-3	Siena	FACED
Task-specific Models	EEGNet	0.002–0.011M	0.5814±0.0125	0.9232±0.0104	0.2961±0.0102	0.4413±0.0096	0.7487±0.0521	0.4090±0.0122
	EEGConformer	0.3–1.5M	0.6049±0.0104	0.9308±0.0117	0.3537±0.0112	0.4506±0.0133	0.7556±0.0210	0.4559±0.0125
	SPaRCNet	0.8–1.3M	0.5932±0.0152	0.9316±0.0095	0.2949±0.0078	0.4426±0.0156	0.6572±0.0381	0.4673±0.0155
	ContraWR	1.6M	0.5892±0.0133	0.9195±0.0115	0.3546±0.0105	0.4257±0.0162	0.6546±0.0311	0.4887±0.0078
	CNN-Transformer	3.2M	0.6053±0.0118	0.9305±0.0068	0.3678±0.0078	0.4533±0.0092	0.6982±0.0560	0.4697±0.0132
	FFCL	2.4–2.5M	0.5726±0.0092	0.9314±0.0083	0.3641±0.0092	0.4678±0.0197	0.6616±0.0391	0.4673±0.0158
	ST-Transformer	3.4–3.5M	0.6035±0.0081	0.9135±0.0103	0.3052±0.0072	0.4126±0.0122	0.7527±0.0381	0.4810±0.0079
Foundation Models	BIOT	3.2M	0.6153±0.0154	0.9358±0.0052	0.3837±0.0187	0.4920±0.0086	0.7352±0.0669	0.5118±0.0118
	LaBraM-Base	5.8M	0.6173±0.0122	0.9409±0.0079	0.3976±0.0138	0.5060±0.0155	0.7082±0.0329	0.5273±0.0107
	CBraMod	4.9M	0.6417±0.0091	0.9560±0.0056	0.4091±0.0097	0.5373±0.0108	0.7317±0.0647	0.5509±0.0089
	CSBrain	8.9M	0.6304±0.0090	0.9643±0.0155	**0.4197±0.0033**	0.6004±0.0187	0.7662±0.0471	0.5752±0.0042
	REVE-Base	69.2M	**0.6480±0.0140**	**0.9644±0.0097**	0.4048±0.0023	0.5635±0.0123	0.7398±0.0074	0.5646±0.0164
	CodeBrain	15.2M	0.6090±0.0070	0.8982±0.0084	0.4137±0.0023	**0.6101±0.0052**	0.6733±0.0273	**0.5941±0.0098**
Diffusion Features	EEGDM (supervised)	5.9M	0.3085±0.0474	0.8962±0.0032	0.3667±0.0038	0.2085±0.0139	**0.8770±0.0159**	0.2053±0.0167
	LUCID (ours, untrained)	7.6–18.4M	0.4106±0.0000	0.8582±0.0003	0.2424±0.0000	0.2533±0.0000	0.5105±0.0000	0.1275±0.0000
	LUCID (ours, trained)	7.6–18.4M	0.4805±0.0052	0.8995±0.0068	0.4028±0.0032	0.3296±0.0088	0.8102±0.0261	0.2121±0.0086

SEED-V and by the widest margins on BCIC2020-3 and FACED, where the probe is also below every supervised network and EEGDM's diffusion features do no better. Against the seven supervised networks the probe is above the best one on seven datasets, ties it on MAT and falls below it on four; against EEGDM it is above on eight datasets, ties on Mumtaz and FACED and trails on TUEV and Siena, where EEGDM is the best model in the table. LUCID is above its untrained copy on all twelve datasets, by 0.04–0.34. Section 3.3 asks what orders these outcomes.

3.3 The Fine-Tuning Gain

Figure 4 reads each foundation model frozen with its own head trained alone (Y) and fully fine-tuned (FT). Their difference, the model's *fine-tuning gain*, averaged over the six models (per model in Section E.1), orders the twelve datasets, and the probe's standing follows that order. The gain is small on SLEEP ($+0.002$), COMA ($+0.019$), TUAB ($+0.036$) and Siena ($+0.040$), four of the probe's five ties, and large on the seven others, from MAT ($+0.126$) to BCIC2020-3 ($+0.319$); the same five datasets carry the five largest gains and the probe's five largest deficits (Spearman $\rho = 0.91$ over the twelve). Mumtaz ($+0.100$) is inflated by one frozen outlier (CBraMod at 0.521); without it the mean is $+0.033$.

The ordering is not a law. TUAB separates nothing: every trained model lies within 0.1 of the others. SEED-V runs against the ordering in our favour: every model gains at least 0.045, yet the probe ends within 0.017 of the best fine-tuned model and above every frozen one. Mumtaz runs the other way: REVE gains nothing, yet the probe trails it by 0.065 and every supervised network. On SLEEP and Siena the frozen LaBraM and REVE backbones read above the probe. Where the gain is largest, the probe fails: on FACED and BCIC2020-3 it is below every supervised network.

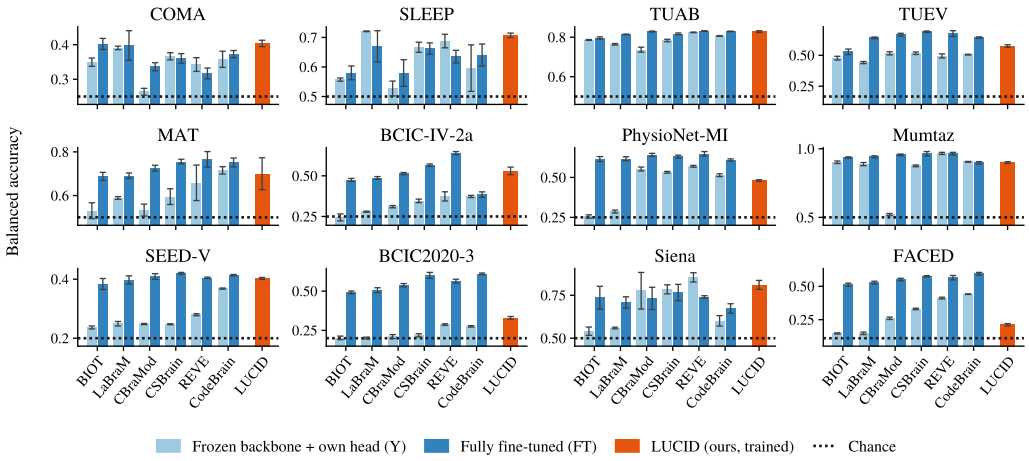


Figure 4: Frozen versus fine-tuned foundation models on twelve EEG datasets: balanced accuracy with the backbone frozen and read through the model's own head (Y, light) and fully fine-tuned (FT, dark), with LUCID (orange). Error bars: std over five seeds; dotted line: chance.

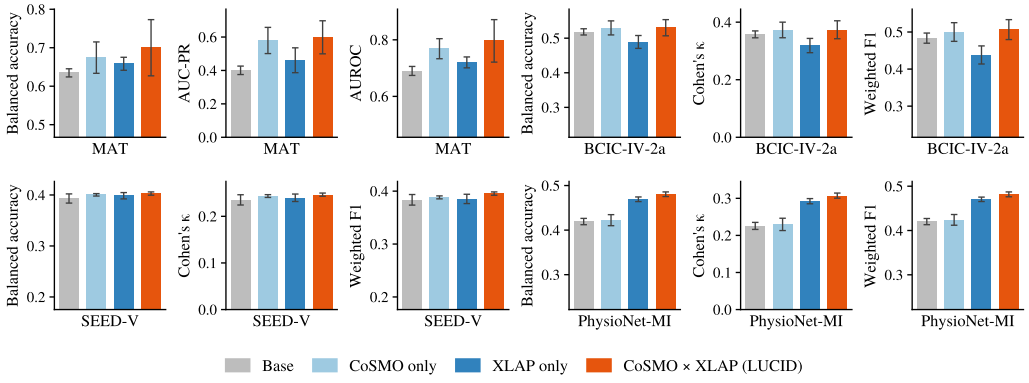


Figure 5: The 2×2 ablation of CoSMO and XLAP on four datasets, one panel per dataset and metric. Bars: the base denoiser, CoSMO only, XLAP only and CoSMO×XLAP, the Table 2 model; error bars: std over five seeds. XLAP only is trained at the gradient-balance weight, which is also the Table 2 model's except on MAT and SEED-V.

3.4 Ablations

CoSMO and XLAP. Figure 5 switches the two components off one at a time, every cell trained with five seeds and probed with its own $t^\star$ and $C^\star$. Neither component is decisive alone. XLAP is the larger effect where it helps: alone, it is above the base denoiser on MAT and PhysioNet-MI, by 0.02 and 0.05, ties it on SEED-V and falls below it on BCIC-IV-2a, by 0.03. CoSMO alone ties the base denoiser on all four in balanced accuracy, but on MAT it lifts AUC-PR by 0.18 and AUROC by 0.08, the two threshold-free metrics. Together they reach the highest balanced accuracy on all four datasets, beyond the seed spread of the base denoiser on SEED-V and PhysioNet-MI and of XLAP alone on BCIC-IV-2a and PhysioNet-MI: CoSMO adds 0.01 on top of XLAP on PhysioNet-MI and recovers on BCIC-IV-2a what XLAP alone loses. The gains are modest, 0.01 to 0.07 over the base denoiser, and the combination is never below it.

PCA width. Section E.3.1 reduces the embedding, 2,560 to 360,448 dimensions, to 256 components by PCA: the probe loses under 0.09 of balanced accuracy on every dataset, nothing on five, and

stays above the full-width untrained copy on all twelve (Figure 9), so width alone does not produce Table 2.

XLAP loss. Section E.3.2 sweeps the XLAP weight λ from a tenth to fifty times the gradient ratio it is set for and the masking ratio ρ from a tenth to nine tenths (Figures 10 and 11): within a decade of the balance point, and for ρ from a third up, the probe moves within its seed spread or by at most 0.015, while a weight beyond ten times the balance costs accuracy on all four datasets and masking only a tenth of the channels costs it on MAT.

3.5 Label Efficiency

Refitting only the probe on a fraction of the labelled training windows (Figure 8, Section E.2), a quarter of the labels matches the full-label balanced accuracy, within seed spread, on six datasets, COMA, TUAB, TUEV, MAT, Mumtaz and Siena, and keeps half to nine tenths of the margin over chance on the six others, where every added label still helps; a twentieth already puts the probe above its untrained copy on every dataset.

Further analyses. Section F visualises what the probe reads, by electrode attribution and by embedding projection; Section G compares the denoiser's samples with real windows on spectral, spatial and temporal statistics; Section H reports parameters and FLOPs.

4 Limitations and Future Work

Limitations. LUCID is evaluated in domain. Each denoiser is trained on the unlabelled training recordings of the dataset it is probed on, so the comparison measures what a dataset's own recordings yield to a label-free model, not cross-dataset transfer, which we neither claim nor test. The cost is one denoiser per dataset, where a foundation model is pretrained once. The features are also uneven. The probe ties the best fine-tuned model on five datasets and trails it on seven, and on FACED and BCIC2020-3 it is below every supervised network. The fine-tuning gain orders these outcomes, but over twelve datasets it is a correlation with two exceptions, not a law, and it covaries with what a label marks, the number of classes and the size of the dataset. The identity-map null isolates training, not the pretext. Corpus, architecture and objective all differ between LUCID and the foundation models, so a diffusion pretext is shown to write a readable representation, not to surpass a masked or contrastive one on the same backbone. Error bars are seed standard deviations, on few test subjects for some datasets, and the label-efficiency curves subsample windows, not subjects.

Future work. A masked or contrastive pretext on the same backbone and recordings would isolate the objective, which the identity-map null does not. Pretraining one denoiser across datasets needs a montage-agnostic input, since the hidden width follows the number of electrodes, and is the step we have not taken. MEG and intracranial recordings are open to the same pretext.

5 Conclusion

We asked what EEG pretraining captures before any label is used, and answered with LUCID: one unconditional diffusion denoiser per dataset, trained without labels and frozen, whose blocks a small linear probe reads. The denoiser keeps the time axis whole and, through CoSMO, adapts every block to the window's own covariance. On twelve datasets under one patient-disjoint protocol, training writes a linearly readable, label-efficient representation. The probe is above the untrained network on every dataset, ties the best fully fine-tuned foundation model on five of the twelve, and trails it where fine-tuning adds those models most. The ordering, not any single number, is the result. It argues for reporting the frozen tier beside the fine-tuned one, and an untrained control beside a label-free in-domain model, before a gain is credited to pretraining. What a denoiser learns from one dataset it may learn from many. Trained across cohorts, montages and modalities and read the same way, the same architecture could become a foundation model. Whether its features would survive away from the recordings that trained them is the question we leave open.

AI USE STATEMENT

We used AI assistants in two limited ways. First, to aid and polish writing: grammar, phrasing and formatting. Second, for retrieval and discovery, such as surfacing candidate related work that we then read and assessed ourselves. The methods, experiments, analyses and claims are our own. We verified every citation and every reported number against its primary source.

ETHICS STATEMENT

We have read the ICLR Code of Ethics and our work adheres to it. All of our experiments use human EEG recordings. Most cohorts are public datasets, reused under their licenses and data-use agreements, and they range from recordings of healthy volunteers to clinical EEG corpora. One cohort is a private ICU cohort of comatose patients. It is de-identified, was collected under the ethics approvals of the study that produced it and is not redistributed with this paper. We evaluate each cohort on its own and claim no diagnostic accuracy or readiness for clinical deployment. Any bedside use would require prospective, cohort-specific validation well beyond what we report here.

REPRODUCIBILITY STATEMENT

The method is specified in Section 2, and the denoiser's objective and architecture in Section B. All hyperparameters of the denoiser, its training and the probe are listed per dataset in Section B.3. The evaluation protocol is described in Section 3.1, including the data splits, the random seeds and validation-only model selection. Dataset processing is described in Section D, ablations in Section E and generation metrics in Section G. We release anonymized code as supplementary material, with the configuration files and environment specification needed to reproduce the reported numbers. The public cohorts reproduce end to end. The private ICU cohort is not redistributed, but its processing is documented so the pipeline can be re-run on any cohort with comparable structure.

REFERENCES

Bruno Aristimunha, Pierre Guetschel, Martin Wimpff, Lukas Gemein, Cedric Rommel, Hubert Banville, Maciej Sliwowski, Daniel Wilson, Simon Brandt, Théo Gnassounou, Joseph Paillard, Bruna Junqueira Lopes, Sara Sedlar, Thomas Moreau, Sylvain Chevallier, Alexandre Gramfort, and Robin Tibor Schirrmeister. Braindecode: toolbox for decoding raw electrophysiological brain data with deep learning models. Zenodo. URL `https://github.com/braindecode/braindecode`.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL `https://arxiv.org/abs/1607.06450`.

Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M Patel. Ddpm-cd: Denoising diffusion probabilistic models as feature extractors for change detection. *arXiv preprint arXiv:2206.11892*, 2022.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2011.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Konstantinos Barmpas, Na Lee, Alexandros Koliousis, Yannis Panagakis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Neurorvq: Multi-scale eeg tokenization for generative large brainwave models. *arXiv preprint arXiv:2510.13068*, 2025.

Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Muller. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*, 25 (1):41–56, 2008.

Eléonore Bouchereau, Angela Marchi, Bertrand Hermann, Estelle Pruvost-Robieux, Eléonore Guinard, Camille Legouy, Caroline Schimpf, Aurélien Mazeraud, Jean-Claude Baron, Céline Ramdani, et al. Quantitative analysis of early-stage eeg reactivity predicts awakening and recovery of consciousness in patients with severe brain injury. *British journal of anaesthesia*, 130(2): e225–e232, 2023.

Eléonore Bouchereau, Estelle Pruvost-Robieux, Shidasp Siami, Cendrine Chaffaut, Adrien Bouglé, Martine Gavaret, Nicholas Heming, Sivanthiny Sivanandamoorthy, Julie Zyss, Vincent Degos, et al. Altered lower brainstem neurophysiological response is associated with mortality in deeply sedated critically ill patients. *Intensive Care Medicine*, 51(6):1050–1061, 2025.

Clemens Brunner, Robert Leeb, and Gernot Müller-Putz. Bci competition 2008–graz data set a, 2024. URL https://dx.doi.org/10.21227/katb-zv89.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M Sandino, and Joseph Y Cheng. Maeeg: Masked auto-encoder for eeg representation learning. *arXiv preprint arXiv:2211.02625*, 2022.

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.

Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36:29996–30025, 2023.

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*, 38:22541–22577, 2026.

Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Diffusion models and representation learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks, 2016. URL https://arxiv.org/abs/1609.09106.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Junjie Huang, Mingyang Li, and Wanzhong Chen. Sad-ver: A self-supervised, diffusion probabilistic model-based data augmentation framework for visual-stimulus eeg recognition. *Advanced Engineering Informatics*, 65:103298, 2025.

Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23115–23127, 2024.

Jaedong Hwang, Kathleen Zhang, Wei Dai, Konstantinos Kontras, Maarten Vanmarcke, Maarten De Vos, Ila Fiete, and Paul Pu Liang. B[fm]^2: Brain foundation model via flow matching with splitunet, 2026. URL https://arxiv.org/abs/2606.20812.

Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 international brain–computer interface competition: A review. *Frontiers in human neuroscience*, 16:898300, 2022.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17): e1750–e1762, 2023.

Soowon Kim, Young-Eun Lee, Seo-Hyun Lee, and Seong-Whan Lee. Diff-e: Diffusion-based learning for decoding imagined speech eeg. *arXiv preprint arXiv:2307.14389*, 2023.

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.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

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.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pp. 562–570. Pmlr, 2015.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, 2022a.

Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, and Debadeepta Dey. What makes convolutional models great on long sequence modeling? *arXiv preprint arXiv:2210.09298*, 2022b.

Lequan Lin, Zhengkun Li, Ruikun Li, Xuliang Li, and Junbin Gao. Diffusion models for time-series applications: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25(1): 19–41, 2024.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, (2):715–729, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for eeg foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 83526–83574, 2026.

Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion based representation learning. In *International conference on machine learning*, pp. 24963–24982. PMLR, 2023.

Manuel Molano-Mazon, Arno Onken, Eugenio Piasini, and Stefano Panzeri. Synthesizing realistic neural population activity patterns using generative adversarial networks. *arXiv preprint arXiv:1803.00338*, 2018.

Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification. *arXiv preprint arXiv:2307.08702*, 2023.

Wajid Mumtaz. MDD Patients and Healthy Controls EEG Data (New). 11 2016. doi: 10.6084/m9.figshare.4244171.v2. URL https://figshare.com/articles/dataset/EEG_Data_New/4244171.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

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

Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10):805–815, 2018.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python, 2018. URL https://arxiv.org/abs/1201.0490.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602. IEEE, 2022.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jia Hong Puah, Sim Kuan Goh, Ziwei Zhang, Zixuan Ye, Chow Khuen Chan, Kheng Seang Lim, Si Lei Fong, Kok Sin Woon, and Cuntai Guan. Eegdm: Eeg representation learning via generative diffusion model. *arXiv preprint arXiv:2508.14086*, 2025.

Patrick L Purdon, Eric T Pierce, Eran A Mukamel, Michael J Prerau, John L Walsh, Kin Foon K Wong, Andres F Salazar-Gomez, Priscilla G Harrell, Aaron L Sampson, Aylin Cimenser, et al. Electroencephalogram signatures of loss and recovery of consciousness from propofol. *Proceedings of the National Academy of Sciences*, 110(12):E1142–E1151, 2013.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Pradeep Shenoy, Matthias Krauledat, Benjamin Blankertz, Rajesh PN Rao, and Klaus-Robert Müller. Towards adaptive classification for bci. *Journal of neural engineering*, 3(1):R13–R23, 2006.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shrikumar17a.html.

Jacobo Diego Sitt, Jean-Remi King, Imen El Karoui, Benjamin Rohaut, Frederic Faugeras, Alexandre Gramfort, Laurent Cohen, Mariano Sigman, Stanislas Dehaene, and Lionel Naccache. Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, 137(8):2258–2270, 2014.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

Szabolcs Torma and Luca Szegletes. Eegwave: a denoising diffusion probabilistic approach for eeg signal generation. *EasyChair Preprint*, 10275, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Julius Vetter, Jakob H Macke, and Richard Gao. Generating realistic neurophysiological time series with denoising diffusion probabilistic models. *Patterns*, 5(9), 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 *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025a.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. Eegdiffuser: Label-guided eeg signals synthesis via diffusion model for bci applications. *Neurocomputing*, 670:132636, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.132636. URL https://www.sciencedirect.com/science/article/pii/S0925231226000330.

Shaocong Wang, Tong Liu, Yihan Li, Ming Li, Kairui Wen, Pei Yang, Wenqi Ji, Minjing Yu, and Yong-Jin Liu. Eegdm: Learning eeg representation with latent diffusion model. *arXiv preprint arXiv:2508.20705*, 2025b.

Yilin Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Benyan Luo, Tao Li, and Gang Pan. Diffmdd: A diffusion-based deep learning framework for mdd diagnosis using eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:728–738, 2024.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14648–14658. IEEE, 2022.

Chuqin Xiang, Xinrui Fan, Duo Bai, Ke Lv, and Xu Lei. A resting-state eeg dataset for sleep deprivation. *Scientific Data*, 11(1):427, 2024.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15756–15766. IEEE, 2023.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023a.

Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2(1):e46769, 2023b.

Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2017.

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. *Advances in Neural Information Processing Systems*, 38: 87150–87195, 2026.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A Related Work

A.1 EEG Decoding and Foundation Models

EEG decoding moved from hand-crafted features, such as the band power of covariance-based spatial filters (Blankertz et al., 2008) or the covariance matrices themselves (Barachant et al., 2011), to networks trained end-to-end on the raw signal (Lawhern et al., 2018; Song et al., 2022). Self-supervision followed, one pretext at a time: contrastive in BENDR (Kostas et al., 2021), masked in MAEEG (Chien et al., 2022). Foundation models pretrain one backbone on large corpora. The six we compare with share per-channel patching and differ in the target. LaBraM predicts the codes of a spectral tokenizer (Jiang et al., 2024) and CodeBrain those of decoupled temporal and spectral codebooks (Ma et al., 2026), with newer tokenisers refining the codebook itself (Barmpas et al., 2025). CBraMod and CSBrain reconstruct masked patches, under criss-cross attention (Wang et al., 2025a) or cross-scale tokens with structured sparse attention (Zhou et al., 2026), and REVE masks patches positioned by electrode coordinates (El Ouahidi et al., 2026). BIOT is pretrained contrastively and with supervision (Yang et al., 2023a). Their electrode mixing uses attention or fixed convolutions and is never conditioned on the window's own covariance (Section 2.2.2). Only B[FM]^2 learns a generative model of the waveform itself, by flow matching on the unpatched signal (Hwang et al., 2026). LUCID is not a foundation model: one denoiser per dataset, trained on its own recordings, with no transfer claimed. It asks what the foundation models' frozen backbones carry (Section A.4).

A.2 Diffusion Models as Representation Learners

Masked reconstruction and denoising are two corruption pretexts of self-supervised learning. A masked autoencoder hides patches and reconstructs them (He et al., 2022). A diffusion model (Ho et al., 2020) adds Gaussian noise at every level and predicts it, so its activations vary with noise level as well as depth. One line of work trains an encoder through the diffusion loss (Mittal et al., 2023; Hudson et al., 2024). Another reads the activations of a pretrained denoiser, for segmentation (Baranchuk et al., 2021), classification (Xiang et al., 2023; Mukhopadhyay et al., 2023) and change detection (Bandara et al., 2022). The most useful features lie at intermediate blocks and low to moderate noise (Baranchuk et al., 2021; Xiang et al., 2023; Mukhopadhyay et al., 2023). The exact block and level are not known in advance, so they are searched for or learnt (Fuest et al., 2026). LUCID keeps this read-out: one forward pass on a real input noised to a level chosen on validation. The sampling trajectory is never run, so feature quality is a separate question from sample quality, as for any generative model (Theis et al., 2015). LUCID applies this read-out to a denoiser trained on one small EEG cohort rather than on images.

A.3 Generative Models of Neural Recordings

Generative models of neural activity predate diffusion. Sequential autoencoders infer single-trial population dynamics (Pandarinath et al., 2018), and adversarial networks synthesise population spike patterns (Molano-Mazon et al., 2018). Diffusion models then spread from images and audio to time-series forecasting, imputation and generation (Lin et al., 2024), and to neurophysiology. There, they generate recordings with realistic spectra (Vetter et al., 2024) and enlarge small training sets (Torma & Szegletes, 2023; Huang et al., 2025; Wang et al., 2024; 2026). In Diff-E, a denoiser's reconstruction errors train an autoencoder, and a classifier reads its latent (Kim et al., 2023). In all of these the denoiser is a means, not the representation. Two works named EEGDM learn the representation by diffusion pretraining, without labels. The first freezes a DDPM denoiser run on each channel separately and reads every layer on the clean signal at a small diffusion step. It then fuses the layers with a supervised Transformer trained on each task's labels (Puah et al., 2025). The second trains an encoder to condition a denoiser in a PCA space of the signal. It then fine-tunes the encoder on each task, without the denoiser (Wang et al., 2025b). LUCID shares with the first the source of its features: the activations of a frozen DDPM denoiser trained on the signal rather than on a latent. It differs in the denoiser. Ours models all electrodes jointly, preserves the time axis and is conditioned on each window's covariance (Section 2). It also differs in the read-out. EEGDM's is that supervised Transformer, larger than the frozen backbone it reads. LUCID's is a noise level chosen on validation, every block pooled over the whole window, and a small logistic probe. Table 2 therefore lists EEGDM as supervised. B[FM]^2 is trained by flow matching, not diffusion. It reads the penultimate layer of its velocity network on the clean input at a fixed flow time, through a linear

head, and fine-tunes the whole network end-to-end on each task (Hwang et al., 2026). It thus asks whether a generative backbone is a better start for fine-tuning. LUCID asks whether a frozen denoiser is linearly readable as it is.

A.4 Frozen versus Fine-Tuned Evaluation

In vision, the linear-probe and fine-tuning rankings of self-supervised encoders need not agree (He et al., 2022). Fine-tuning can also distort the pretrained features and underperform a linear probe out of distribution, when the features are good and the shift is large (Kumar et al., 2022). The same study proposes a two-step remedy: fit a linear probe on the frozen encoder, then fine-tune the whole network from that probe. EEG foundation model papers report the frozen tier unevenly. REVE adopts that two-step recipe and reports linear probing as a first-class result (El Ouahidi et al., 2026). LaBraM's partial fine-tuning study places linear probing below full fine-tuning (Jiang et al., 2024). CBraMod concludes that its backbone cannot yet serve as a fixed feature extractor (Wang et al., 2025a). CodeBrain reports linear probing only as a low-resource setting (Ma et al., 2026). CSBrain and B[FM]^2 report fine-tuned rows only (Zhou et al., 2026; Hwang et al., 2026). Where a frozen tier appears, it covers a few models or datasets. No paper therefore measures, for every model and dataset, what fine-tuning adds over the model's own frozen backbone, hence what generic EEG representation its pretraining alone learnt. The fine-tuning gain of Section 3.3 is that quantity (Figure 4).

B Additional Training Details

B.1 Denoising Objective

Table 3: Training settings shared by all datasets. Exceptions are in parentheses. The settings that differ between datasets are in Table 4.

	Setting	**Value**
Input	Normalisation	Per-channel z-score with training statistics (TUAB, MAT, FACED: per-window 95th-percentile scaling)
Diffusion	Steps T_{diff}	1000
	Prediction target	ϵ
	Variance schedule	Linear from 10^{-4} to 0.02 (TUAB, BCIC-IV-2a: quadratic to 0.04 and 0.033; TUEV: cosine with offset 0.008 (Nichol & Dhariwal, 2021))
CoSMO	Electrode-embedding width d_E	8
XLAP	Channel-masking probability	0.33
	Weight λ	Per dataset (Table 4)
Optimisation	Optimiser	AdamW, moment decay rates 0.9 and 0.999
	Weight decay	10^{-3} (SLEEP: 10^{-2})
	Learning-rate schedule	Cosine annealing from η to $\eta/100$ over all epochs, stepped once per epoch, no warm-up
	Peak rate η, batch size, epochs	Per dataset (Table 4)
	Gradient clipping	None (TUEV, MAT: max norm 1; COMA, Mumtaz, Siena, FACED: 10^6)
	Weights kept	Last epoch, no EMA

Forward process, objective and sampling. The forward process is a Markov chain, $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$, with a fixed variance schedule $\{\beta_t\}_{t=1}^{T_{\text{diff}}}$ (Table 3) and $\bar{\alpha}_t = \prod_{s \le t}(1-\beta_s)$; its marginal is the Gaussian of Equation (1), so any level is reached in one draw, almost clean at the first step and essentially pure noise at T_{diff}. The denoiser minimises Equation (1) in expectation over x_0, t and ϵ, the simplified objective of Ho et al. (2020), summed over the $C \times T$ entries. Samples (Section G) come from ancestral sampling (Ho et al., 2020) from $x_{T_{\text{diff}}} \sim \mathcal{N}(0, I)$,

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\hat{\epsilon}_\theta(x_t, t)\Big) + \sigma_t z_t, \qquad z_t \sim \mathcal{N}(0, I), \tag{5}$$

with $\alpha_t = 1 - \beta_t$, $\sigma_t^2 = \beta_t\,(1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)$ and no noise at the last step. Feature extraction never runs Equation (5): one noising

to the selected t, one forward pass (Section 2.4).

Cross-layer auxiliary objective. XLAP adapts REVE's secondary loss (El Ouahidi et al., 2026) and, where EEG pretraining masks time spans (Kostas et al., 2021; Chien et al., 2022) or channel–time patches (Jiang et al., 2024; Wang et al., 2025a), masks whole channels, as the channel drop-out within REVE's block masking does: each electrode of x_t is zeroed independently with the probability in Table 3, before the CoSMO statistics and `conv_in`, for a second pass at the same t. Each block output of that pass is pooled over time into its mean and standard deviation; one learned query attends, single-headed, over the N pooled vectors after a layer normalisation and linear key and value maps, and a two-layer head with a GELU (Hendrycks & Gimpel, 2023) maps the result to the predicted covariance and log-variances. The targets are CoSMO's two statistics computed on the clean window, scored by Equation (3). In the per-window objective the XLAP term is weighted by $\sqrt{\bar{\alpha}_t}$, which fades it at high noise, and by the number of entries $C\,T$, which matches the units of the summed denoising loss; Equation (4) omits this fixed weighting. XLAP's gradient reaches every parameter but `conv_out`; XLAP writes nothing into the residual stream and is off at validation, generation and read-out.

Gradient balance. The weight λ of Equation (4) is not tuned on downstream results but set per dataset (Table 4) so that the two losses pull equally hard. On a trained checkpoint, for each batch and at the same noise level and noise draw, we compute the norm of the gradient of $\mathcal{L}_{\text{main}}$ and the norm of the gradient of the XLAP term at $\lambda = 1$, both on the parameters the two losses share, and take their ratio r. λ is chosen so that the median of r over batches equals one ($r^\star = 1$): the auxiliary gradient is then as large as the denoising gradient, neither dominating.

Optimisation. AdamW (Loshchilov & Hutter, 2019) over all parameters, XLAP head included, with a per-epoch cosine learning-rate schedule, gradient-norm clipping and gradient accumulation where used (learning rate, weight decay, moments, epochs, batch sizes, clipping: Tables 3 and 4); the last-epoch weights are kept, without exponential moving average.

B.2 Architecture

Backbone. Windows are normalised as in Table 3. `conv_in` is a 1×1 convolution masked block-diagonally: electrode c feeds only its own m features, so $H = Cm$; `conv_out` carries the transposed mask, m features back to one trace per electrode. N blocks keep (H, T) (N, m: Table 4). Each follows adaLN-Zero (Peebles & Xie, 2023): a linear map of $\mathrm{SiLU}(\tau_t)$ (Hendrycks & Gimpel, 2023) emits the two sub-layers' six shifts, scales and gates, and a second map of τ_t, without SiLU, emits the two CoSMO gates; both start at zero. Layer normalisation (Ba et al., 2016) is parameter-free over the whole (H, T) window. τ_t is a sinusoidal embedding of t through a two-layer MLP; where used (Table 4), a same-form embedding of the sample index joins τ_t on the adaLN bus (all eight maps, not CoSMO's W). The MLP is two 1×1 convolutions with a GELU, dense or with a masked cross-electrode part (per dataset).

Structured long convolutions. SLConv (Li et al., 2022b; Vetter et al., 2024) is depthwise: each hidden feature owns one long kernel made of two learned segments of equal length (Table 4), the nearer weighted twice the farther, divided by its initial norm (fixed thereafter) and applied circularly around the window through the fast Fourier transform in $O(T \log T)$ time, plus a per-feature skip of its input.

CoSMO. Once per forward pass, from x_t centred over time, the network forms the channel covariance and reduces it to CoSMO's two statistics: v, the log-variance of each electrode standardised across electrodes, and s, the covariance divided by its mean variance, projected on both sides through the learned electrode embedding E of width d_E (Table 3) and standardised; both standardisations are parameter-free layer normalisations. In each block, a small two-layer network with a SiLU maps v to the gain a, and another maps s and τ_t to an update of a learned base W_0, giving $W(s, \tau_t)$; each small network's output is scaled by a per-block scalar (κ_a, κ_c) that starts at zero. The two terms of

Equation (2) then pass through their own per-electrode $m \times m$ projection and, like the MLP, enter through their own gate. Generation recomputes v and s from the chain state at every step.

B.3 Per-Cohort Configurations

Table 4: Settings that differ between datasets. Denoiser: number of blocks N, hidden features per electrode m, and long-convolution kernel length in samples at the training rate (200 Hz, or 256 Hz on COMA and 250 Hz on SLEEP). Training: peak learning rate η, effective batch size, epochs and XLAP weight λ. Probe: noise level $t^\star$ and inverse ℓ_2 penalty $C^\star$. †With the optional positional embedding. ‡Block outputs pooled over 64 (BCIC-IV-2a) and 16 (PhysioNet-MI) equal time chunks before the probe.

	Denoiser			**Training**				**Probe**	
Dataset	N	m	Kernel	η ($\times 10^{-4}$)	Batch	Epochs	λ	$t^\star$	$C^\star$
COMA	16	16	768	20	64	300	0.0165	200	10^{-5}
SLEEP	12	8	500	16	128	2045	0.02	200	10^{-6}
TUAB	8	16	768	4	64	84	0.051	50	10^{-4}
TUEV	12	24	240	4	64	1000	0.013	200	10^{-6}
MAT	4	16	300	16	32	2400	0.186	50	10^{-4}
BCIC-IV-2a†‡	8	16	240	4	64	2200	0.0129	50	10^{-3}
PhysioNet-MI†‡	8	8	240	4	64	1010	0.055	200	10^{-3}
Mumtaz	16	24	300	2	64	1300	0.0195	50	10^{-3}
SEED-V	8	8	60	4	64	184	0.74	1	10^{-2}
FACED	8	12	600	20	64	940	0.397	200	10^{-3}
BCIC2020-3†	8	8	220	4	64	350	0.0308	1	10^{-3}
Siena	8	16	600	8	64	168	0.043	1	10^{-3}

Each denoiser is trained on the training split of its dataset, without labels. The five seeds of a dataset share its configuration. The probe's noise level $t^\star$ is selected in $\{1, 50, 200\}$ at a fixed penalty, then its inverse penalty $C^\star$ among the powers of ten from 10^{-5} to 10^{2}, extended by one decade when the best value lies at an edge. Both are selected on the validation balanced accuracy averaged over the five seeds, so the seeds of a dataset also share $t^\star$ and $C^\star$. The probe weights each class by its inverse frequency. The parameter count of each denoiser is given with its results in Section D.

B.4 Training Loss Curves

Figure 6 shows the training loss (Equation (4), averaged over the steps of each epoch) against epochs for one training run of each configuration, replotted from the training logs. The curves are reported for completeness; nothing in the protocol selects on them.

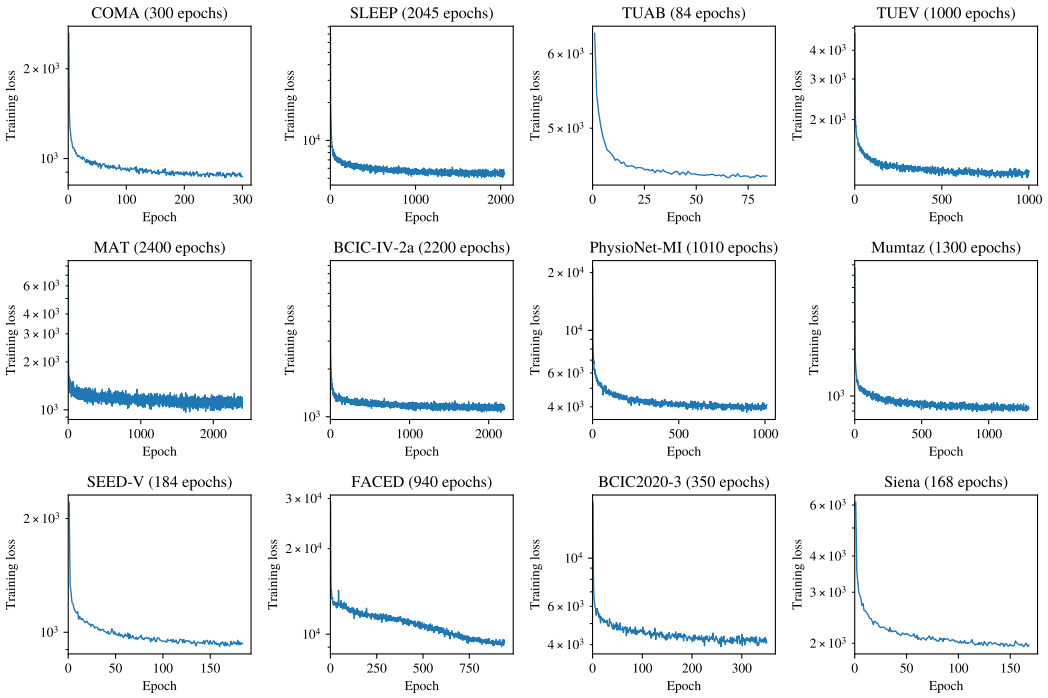


Figure 6: Training loss of the denoisers against epochs, one training run per configuration, logarithmic y-axis. The number of epochs is given in each panel title.

C Additional Experimental Settings

C.1 EEG Decoding Tasks

We evaluate on the ten tasks of Table 1, listed here in the same order; the datasets, their preprocessing and their splits are described in Section D.

- **ICU sedation depth estimation** rates the depth of sedation of a comatose patient from scalp EEG on four levels, from no to heavy sedation. It is a clinical task on COMA (Bouchereau et al., 2023; 2025), our private cohort.
- **Sleep deprivation detection** tells whether a resting recording of a healthy adult was made after a normal night or after a night without sleep (SLEEP; Xiang et al., 2024); the two sessions of every subject are the two classes.
- **Abnormal detection** decides whether a clinical recording is normal or abnormal, as labelled by neurologists (TUAB; Obeid & Picone, 2016).
- **Event type classification** assigns each annotated window of a clinical recording to one of six event types: spike and sharp wave, generalized and lateralized periodic epileptiform discharges, eye movement, artifact and background (TUEV; Obeid & Picone, 2016).
- **Mental arithmetic task** separates EEG recorded at rest from EEG recorded during a serial-subtraction task, taken as without and with mental stress (MAT; Zyma et al., 2019).
- **Motor imagery** decodes which movement a subject imagines from the EEG of the trial: left hand, right hand, both feet or tongue on BCIC-IV-2a (Brunner et al., 2024), and left fist, right fist, both fists or both feet on PhysioNet-MI (Goldberger et al., 2000).
- **Mental disorder diagnostic** separates patients with major depressive disorder from healthy controls on resting EEG (Mumtaz; Mumtaz, 2016).
- **Emotion recognition** identifies the emotion elicited by a film clip: five emotions on SEED-V (Liu et al., 2021), nine on FACED (Chen et al., 2023). SEED-V is split by trial within subject, FACED by subject.
- **Imagined speech classification** identifies which of five phrases a subject silently imagines saying (BCIC2020-3; Jeong et al., 2022); the split is within subject.

- **Seizure detection** flags the windows of a clinical recording that contain a seizure (Siena; Detti et al., 2020), a heavily imbalanced binary task.

C.2 BASELINES

Foundation models. Six pretrained EEG encoders with released weights, each read two ways on identical splits: frozen, with only its own default head trained (Y), and fully fine-tuned from the paper's released recipe (FT).

- **BIOT** (Yang et al., 2023a) is a linear transformer over per-channel spectral patch tokens, pretrained with mixed supervised and unsupervised objectives. Its checkpoint embeds at most 18 channels; in our runs, the missing channel embeddings are drawn from the mean and standard deviation of the pretrained table, the rest of the network staying pretrained.
- **LaBraM-Base** (Jiang et al., 2024) trains a neural tokenizer by vector-quantised spectrum prediction, then a transformer that predicts the codes of masked patches.
- **CBraMod** (Wang et al., 2025a) is a criss-cross transformer with separate spatial and temporal attention, pretrained by masked patch reconstruction.
- **CSBrain** (Zhou et al., 2026) alternates cross-scale spatiotemporal tokenisation with structured sparse attention.
- **REVE-Base** (El Ouahidi et al., 2026) is a masked autoencoder with a 4D positional encoding, pretrained on 92 datasets and 25,000 subjects.
- **CodeBrain** (Ma et al., 2026) pairs a decoupled time–frequency tokenizer with a multi-scale state-space architecture.

The frozen rows are computed by us on every dataset, with each model's default head. The fine-tuned rows are computed by us on COMA and SLEEP, and for CodeBrain on BCIC-IV-2a, PhysioNet-MI, Mumtaz and Siena, which its paper does not cover; elsewhere they are the numbers published by REVE (El Ouahidi et al., 2026), CSBrain (Zhou et al., 2026) and CodeBrain (Ma et al., 2026) under their own recipes. In every run of ours, learning rate, weight decay and batch size are selected on the validation split and five seeds are scored once on the test split.

Diffusion-based features.

- **EEGDM** (Puah et al., 2025) is the published method closest to LUCID: a signal-space diffusion model whose frozen denoiser activations serve as the representation. The denoiser stacks 20 DiffWave blocks built on bidirectional S4D layers, applied to each channel separately, and is trained as a DDPM (v-prediction, cosine schedule, 50 steps) on μ-law companded signals. The read-out is a latent fusion transformer trained with the labels of each task, in which learned fusion tokens cross-attend to the activations of every block pooled over 1 s windows: 6.9M supervised parameters, more than the 5.9M backbone.
- **Latent-diffusion EEGDM** (Wang et al., 2025b) is a second method of the same name, built on a latent diffusion model. Its code is not public; its published TUAB number is the TUAB row.

LUCID differs from EEGDM on both sides: its denoiser is multivariate and models all channels jointly, and its read-out is a small linear probe, whereas EEGDM's head is a 6.9M-parameter network trained by supervised learning on each task, larger than the frozen backbone it reads. Every EEGDM number is therefore the outcome of supervised training on top of the diffusion features, and Table 2 lists it as supervised. No released EEGDM checkpoint covers our datasets: we train the denoiser on each dataset's training split with the paper's settings, then the head, supervised, on the training labels, with its learning rate, weight decay and batch size selected on the validation split; MAT and Siena, the two imbalanced binary tasks, use the class-ratio loss weighting of the paper's seizure recipe. The TUEV row is the paper's published number under its own protocol.

Supervised baselines. Seven networks trained from scratch on each dataset, without pretraining.

- **EEGNet** (Lawhern et al., 2018) is a compact CNN of depthwise and separable convolutions.
- **EEGConformer** (Song et al., 2022) is a convolutional front-end followed by self-attention.

- **SPaRCNet** (Jing et al., 2023) is a 1-D CNN with dense residual connections.
- **ContraWR** (Yang et al., 2023b) is a 2-D ResNet on multi-channel spectrograms.
- **CNN-Transformer** (Peh et al., 2022) extracts convolutional features and passes them through transformer layers.
- **FFCL** (Li et al., 2022a) runs parallel CNN and LSTM branches, fused by a fully connected layer.
- **ST-Transformer** (Song et al., 2021) applies attention over both the channel and the time dimension.

On COMA and SLEEP we train them ourselves, ST-Transformer, FFCL and CNN-Transformer from BIOT's public implementation (Yang et al., 2023a) and the other four from braindecode (Aristimunha et al.), under one fixed recipe, not tuned per network or per dataset (AdamW, learning rate 10^{-3}, weight decay 10^{-4}, cosine annealing, batch size 64, at most 100 epochs with early stopping, five seeds); elsewhere the rows are the numbers published by REVE (El Ouahidi et al., 2026) and CSBrain (Zhou et al., 2026). They are reference points under one fixed budget, read as a floor tier in Table 2.

C.3 Metrics

Following prior work (Yang et al., 2023a; Jiang et al., 2024; Wang et al., 2025a; Zhou et al., 2026), every model is scored on the test split with the following metrics.

- **Balanced accuracy** (bACC) is the mean of the per-class recalls, insensitive to class imbalance; it is the headline metric on every task.
- **AUC-PR** is the area under the precision–recall curve, computed as average precision, for binary tasks.
- **AUROC** is the area under the receiver operating characteristic curve, for binary tasks. Both areas are computed from the predicted probability of the positive class and do not depend on the decision threshold.
- **Cohen's κ** is the agreement between predicted and true labels corrected for chance agreement, for multi-class tasks.
- **Weighted F1** is the mean of the per-class F1 scores weighted by class support, for multi-class tasks.

All metrics are computed with scikit-learn (Pedregosa et al., 2018) and reported as mean $\pm$ standard deviation over five seeds. Validation balanced accuracy selects the noise level that LUCID's probe reads and the penalty of the probe.

D Detailed Results on EEG Decoding Tasks

This section describes the twelve datasets and reports, for each, the three-metric results behind Table 2 (mean $\pm$ std over five seeds). Unless stated otherwise, signals are resampled to 200 Hz. Filters are applied to the continuous recording before it is cut into windows. On the ten datasets shared with prior work, we use the provided train/validation/test splits. COMA and SLEEP use our own splits, with no patient or subject shared across training, validation and test. *Params* is the model size.

D.1 ICU Sedation Depth Estimation

Table 5: The results of different methods on COMA (sedation depth estimation, 4-class).

	Method	Params	bACC	Cohen's κ	Weighted F1
Task-specific Models	EEGNet	0.004M	0.3228±0.0283	0.1327±0.0323	0.4574±0.0188
	EEGConformer	1.2M	0.3363±0.0387	0.1776±0.0499	0.4782±0.0339
	SPaRCNet	1.1M	0.3419±0.0175	0.1921±0.0225	0.4796±0.0098
	ContraWR	1.6M	0.2957±0.0225	0.1226±0.0235	0.4525±0.0175
	CNN-Transformer	3.2M	0.3134±0.0285	0.1545±0.0411	0.4684±0.0302
	FFCL	2.4M	0.3108±0.0210	0.1408±0.0276	0.4622±0.0149
	ST-Transformer	3.4M	0.3748±0.0173	0.2154±0.0207	0.4855±0.0133
Foundation Models	BIOT	3.2M	0.4022±0.0161	0.2669±0.0149	**0.5452±0.0085**
	LaBraM-Base	5.8M	0.3977±0.0429	**0.2697±0.0521**	0.5366±0.0332
	CBraMod	4.9M	0.3362±0.0118	0.1511±0.0163	0.4732±0.0097
	CSBrain	8.9M	0.3603±0.0139	0.1701±0.0149	0.4774±0.0091
	REVE-Base	69.2M	0.3167±0.0157	0.1379±0.0183	0.4622±0.0017
	CodeBrain	15.2M	0.3728±0.0104	0.2032±0.0186	0.4967±0.0112
Diffusion Features	EEGDM (supervised)	5.9M	0.3865±0.0039	0.2286±0.0112	0.5074±0.0088
	LUCID (ours, untrained)	4.7M	0.2500±0.0000	0.0000±0.0000	0.3888±0.0000
	LUCID (ours, trained)	4.7M	**0.4038±0.0092**	0.2498±0.0137	0.5168±0.0100

COMA (Bouchereau et al., 2023; 2025). A private ICU cohort of 159 comatose patients (200 recordings) with 11-channel scalp EEG at 256 Hz, kept at its native rate. The task is to estimate sedation depth on four levels: no, light, medium and heavy sedation. Signals are band-pass filtered at 0.5–45 Hz, notch filtered at 50 Hz and cut into 8 s windows, giving 32,685 windows. The patient-disjoint split puts patients 1–82 in training, 83–118 in validation and 119–159 in test (82/36/41 patients, 20,667/5,302/6,716 windows). Table 5 reports the results.

D.2 Sleep Deprivation Detection

Table 6: The results of different methods on SLEEP (sleep deprivation, 2-class).

	Method	Params	bACC	AUC-PR	AUROC
Task-specific Models	EEGNet	0.004M	0.4883±0.0176	0.5646±0.0441	0.6139±0.0415
	EEGConformer	1.3M	0.5019±0.0024	0.7456±0.0446	0.7383±0.0348
	SPaRCNet	1.2M	0.6267±0.0124	0.6938±0.0570	0.7033±0.0412
	ContraWR	1.6M	0.6024±0.0523	0.7257±0.0487	0.7507±0.0214
	CNN-Transformer	3.2M	0.5966±0.0305	0.6734±0.0437	0.7123±0.0366
	FFCL	2.5M	0.5967±0.0300	0.6732±0.0513	0.6896±0.0417
	ST-Transformer	3.5M	0.5967±0.1018	**0.7785±0.0554**	0.7861±0.0366
Foundation Models	BIOT	3.2M	0.5801±0.0233	0.6227±0.0123	0.6165±0.0239
	LaBraM-Base	5.8M	0.6700±0.0528	0.7617±0.0801	0.7834±0.0530
	CBraMod	4.9M	0.5795±0.0452	0.6424±0.0508	0.6789±0.0464
	CSBrain	8.9M	0.6624±0.0193	0.6828±0.0367	0.7247±0.0249
	REVE-Base	69.2M	0.6351±0.0214	0.7181±0.0472	0.7497±0.0408
	CodeBrain	15.2M	0.6404±0.0373	0.6892±0.0390	0.7130±0.0366
Diffusion Features	EEGDM (supervised)	5.9M	0.6312±0.0193	0.7270±0.0239	0.6847±0.0214
	LUCID (ours, untrained)	15.4M	0.5793±0.0000	0.6052±0.0000	0.6391±0.0000
	LUCID (ours, trained)	15.4M	**0.7071±0.0077**	0.7609±0.0328	**0.7882±0.0174**

SLEEP (Xiang et al., 2024). Eyes-open resting EEG from 71 healthy adults, recorded with 61 channels[4] at 500 Hz once after a normal night and once after sleep deprivation. The task is to tell the two sessions apart. Recordings are downsampled to 250 Hz, band-pass filtered at 0.5–45 Hz, notch filtered at 50 Hz and cut into 8 s windows, giving 4,615 windows. The subject-disjoint split puts

[4]We detected three corrupted or noisy electrodes (C3, F1 and O1) and excluded them, leaving 58 channels.

subjects 1–46 in training, 47–57 in validation and 58–71 in test (46/11/14 subjects, 2,912/764/939 windows). Table 6 reports the results.

D.3 ABNORMAL DETECTION

Table 7: The results of different methods on TUAB (abnormal EEG, 2-class).

	Method	Params	bACC	AUC-PR	AUROC
Task-specific Models	EEGNet	0.003M	0.7642±0.0036	0.8299±0.0043	0.8412±0.0031
	EEGConformer	1.5M	0.7758±0.0049	0.8427±0.0054	0.8445±0.0038
	SPaRCNet	1.1M	0.7896±0.0018	0.8414±0.0018	0.8676±0.0012
	ContraWR	1.6M	0.7746±0.0041	0.8421±0.0104	0.8456±0.0074
	CNN-Transformer	3.2M	0.7777±0.0022	0.8433±0.0039	0.8461±0.0013
	FFCL	2.5M	0.7848±0.0038	0.8448±0.0065	0.8569±0.0051
	ST-Transformer	3.4M	0.7966±0.0023	0.8521±0.0026	0.8707±0.0019
Foundation Models	BIOT	3.2M	0.7959±0.0057	0.8792±0.0023	0.8815±0.0043
	LaBraM-Base	5.8M	0.8140±0.0019	0.8965±0.0016	0.9022±0.0009
	CBraMod	4.9M	0.8289±0.0022	0.9258±0.0008	0.9227±0.0011
	CSBrain	8.9M	0.8172±0.0043	0.9005±0.0066	0.8957±0.0046
	REVE-Base	69.2M	**0.8315±0.0014**	**0.9281±0.0009**	**0.9245±0.0013**
	CodeBrain	15.2M	0.8294±0.0013	0.9100±0.0006	0.9030±0.0009
Diffusion Features	EEGDM (supervised)	5.9M	0.8051±0.0093	0.8981±0.0092	0.8856±0.0120
	LUCID (ours, untrained)	4.2M	0.4870±0.0008	0.4631±0.0003	0.4905±0.0003
	LUCID (ours, trained)	4.2M	0.8283±0.0042	0.9042±0.0023	0.8981±0.0019

TUAB (Obeid & Picone, 2016). The TUH Abnormal EEG Corpus (v3.0.1) contains 2,993 clinical recordings from 2,327 patients, each labelled normal or abnormal. Following the BIOT preprocessing (Yang et al., 2023a) shared by later foundation models, recordings are resampled from their native 250–512 Hz to 200 Hz, mapped to the 16 bipolar channels of the double-banana montage, band-pass filtered at 0.3–75 Hz with a 60 Hz notch, and cut into 10 s windows, giving 409,455 windows. The corpus-provided evaluation set is the test set (253 patients, 36,945 windows). The remaining patients are split 80/20 into training and validation (1,510/564 patients, 302,058/70,452 windows). Table 7 reports the results.

D.4 EVENT TYPE CLASSIFICATION

Table 8: The results of different methods on TUEV (event type classification, 6-class).

	Method	Params	bACC	Cohen's κ	Weighted F1
Task-specific Models	EEGNet	0.004M	0.3876±0.0143	0.3577±0.0155	0.6539±0.0120
	EEGConformer	0.8M	0.4074±0.0164	0.3967±0.0195	0.6983±0.0152
	SPaRCNet	1.0M	0.4161±0.0262	0.4233±0.0181	0.7024±0.0104
	ContraWR	1.6M	0.4384±0.0349	0.3912±0.0237	0.6893±0.0136
	CNN-Transformer	3.2M	0.4087±0.0161	0.3815±0.0134	0.6854±0.0293
	FFCL	2.4M	0.3979±0.0104	0.3732±0.0188	0.6783±0.0120
	ST-Transformer	3.4M	0.3984±0.0228	0.3765±0.0306	0.6823±0.0190
Foundation Models	BIOT	3.2M	0.5281±0.0225	0.5273±0.0249	0.7492±0.0082
	LaBraM-Base	5.8M	0.6409±0.0065	0.6637±0.0093	0.8312±0.0052
	CBraMod	4.9M	0.6671±0.0107	0.6772±0.0096	0.8342±0.0064
	CSBrain	8.9M	0.6903±0.0059	0.6833±0.0047	0.8333±0.0057
	REVE-Base	69.2M	0.6759±0.0229	0.6783±0.0199	0.8451±0.0129
	CodeBrain	15.2M	0.6428±0.0062	0.6912±0.0101	0.8362±0.0048
Diffusion Features	EEGDM (supervised)	5.9M	**0.7557±0.0147**	**0.7423±0.0136**	**0.8688±0.0066**
	LUCID (ours, untrained)	9.8M	0.3338±0.0000	0.2276±0.0000	0.5880±0.0000
	LUCID (ours, trained)	9.8M	0.5744±0.0104	0.4302±0.0067	0.7025±0.0051

TUEV (Obeid & Picone, 2016). The TUH EEG Events Corpus (v2.0.1) annotates 518 recordings from 337 patients with six event types: spike and sharp wave, generalized periodic epileptiform discharges, periodic lateralized epileptiform discharges, eye movement, artifact and background. With the same BIOT preprocessing as TUAB (16 bipolar channels, 0.3–75 Hz band-pass, 60 Hz notch, 250 Hz resampled to 200 Hz), one 5 s window is cut around each annotation, giving 113,353 labelled windows.[5] The corpus-provided evaluation set is the test set (63 patients, 29,421 windows). The remaining patients are split 80/20 into training and validation (185/89 patients, 67,141/16,791 windows). Table 8 reports the results.

D.5 Mental Arithmetic Task

Table 9: The results of different methods on MAT (mental arithmetic task, 2-class).

	Method	Params	bACC	AUC-PR	AUROC
Task-specific Models	EEGNet	0.002M	0.6770±0.0116	0.5763±0.0102	0.7321±0.0108
	EEGConformer	0.8M	0.6805±0.0123	0.5829±0.0134	0.7424±0.0128
	SPaRCNet	1.0M	0.6879±0.0107	0.5825±0.0193	0.7418±0.0132
	ContraWR	1.6M	0.6631±0.0097	0.5787±0.0164	0.7332±0.0082
	CNN-Transformer	3.2M	0.6779±0.0268	0.5777±0.0285	0.7258±0.0336
	FFCL	2.4M	0.6798±0.0142	0.5786±0.0266	0.7330±0.0198
	ST-Transformer	3.4M	0.6631±0.0173	0.5672±0.0259	0.7132±0.0174
Foundation Models	BIOT	3.2M	0.6875±0.0186	0.6004±0.0195	0.7536±0.0144
	LaBraM-Base	5.8M	0.6909±0.0125	0.5999±0.0155	0.7721±0.0093
	CBraMod	4.9M	0.7256±0.0132	0.6267±0.0099	0.7905±0.0073
	CSBrain	8.9M	0.7558±0.0106	0.6696±0.0221	0.8478±0.0297
	REVE-Base	69.2M	**0.7660±0.0355**	**0.7470±0.0807**	0.8450±0.0514
	CodeBrain	15.2M	0.7514±0.0203	0.7177±0.0421	**0.8707±0.0209**
Diffusion Features	EEGDM (supervised)	5.9M	0.5986±0.0545	0.3936±0.0169	0.6693±0.0217
	LUCID (ours, untrained)	2.4M	0.5243±0.0000	0.3173±0.0000	0.4967±0.0000
	LUCID (ours, trained)	2.4M	0.7000±0.0731	0.5984±0.0989	0.7963±0.0751

MAT (Zyma et al., 2019). The MAT dataset contains 20-channel EEG at 500 Hz from 36 healthy subjects, recorded at rest and during a serial-subtraction task. The two recordings are labelled without and with mental stress. Signals are resampled to 200 Hz, high-pass filtered at 0.5 Hz, notch filtered at 50 Hz and cut into 5 s windows, giving 1,707 windows. We use the published subject split 1–28 / 29–32 / 33–36 (1,343/172/192 windows). Table 9 reports the results.

[5] Prior work reports 112,491 windows (Jiang et al., 2024; Wang et al., 2025a; Zhou et al., 2026; El Ouahidi et al., 2026). The difference comes from a change of corpus version. The change does not affect the test set.

D.6 MOTOR IMAGERY CLASSIFICATION

Table 10: The results of different methods on BCIC-IV-2a (motor imagery, 4-class).

	Method	Params	bACC	Cohen's κ	Weighted F1
Task-specific Models	EEGNet	0.003M	0.4482±0.0094	0.2693±0.0121	0.4226±0.0108
	EEGConformer	0.6M	0.4696±0.0106	0.2924±0.0141	0.4533±0.0128
	SPaRCNet	1.0M	0.4635±0.0117	0.2847±0.0147	0.4432±0.0126
	ContraWR	1.6M	0.4678±0.0125	0.2905±0.0160	0.4413±0.0142
	CNN-Transformer	3.2M	0.4600±0.0108	0.2800±0.0148	0.4460±0.0114
	FFCL	2.4M	0.4470±0.0143	0.2627±0.0176	0.4238±0.0139
	ST-Transformer	3.4M	0.4575±0.0145	0.2733±0.0198	0.4471±0.0142
Foundation Models	BIOT	3.2M	0.4748±0.0093	0.2997±0.0139	0.4607±0.0125
	LaBraM-Base	5.8M	0.4869±0.0085	0.3159±0.0154	0.4758±0.0103
	CBraMod	4.9M	0.5138±0.0066	0.3518±0.0094	0.4984±0.0085
	CSBrain	8.9M	0.5657±0.0071	0.4209±0.0093	0.5637±0.0087
	REVE-Base	69.2M	**0.6396±0.0095**	**0.5194±0.0126**	**0.6339±0.0110**
	CodeBrain	15.2M	0.3847±0.0167	0.1796±0.0223	0.3452±0.0219
Diffusion Features	EEGDM (supervised)	5.9M	0.2700±0.0419	0.0266±0.0558	0.1568±0.0661
	LUCID (ours, untrained)	5.6M	0.4254±0.0000	0.2338±0.0000	0.4209±0.0000
	LUCID (ours, trained)	5.6M	0.5302±0.0234	0.3736±0.0312	0.5065±0.0269

BCIC-IV-2a (Brunner et al., 2024). Nine healthy subjects imagined four movements (left hand, right hand, both feet, tongue) over two sessions of 288 trials each, recorded with 22 electrodes at 250 Hz. Signals are band-pass filtered at 0.3–40 Hz with a 50 Hz notch. The imagery interval of each trial, from 2 to 6 s, is cut and resampled to 200 Hz, giving 5,184 windows of 4 s. We use the published subject split A01–A05 / A06–A07 / A08–A09 (2,880/1,152/1,152 windows). Table 10 reports the results.

Table 11: The results of different methods on PhysioNet-MI (motor imagery, 4-class).

	Method	Params	bACC	Cohen's κ	Weighted F1
Task-specific Models	EEGNet	0.004M	0.5814±0.0125	0.4468±0.0199	0.5796±0.0115
	EEGConformer	0.7M	0.6049±0.0104	0.4736±0.0171	0.6062±0.0095
	SPaRCNet	1.3M	0.5932±0.0152	0.4564±0.0234	0.5937±0.0147
	ContraWR	1.6M	0.5892±0.0133	0.4527±0.0248	0.5918±0.0116
	CNN-Transformer	3.2M	0.6053±0.0118	0.4725±0.0223	0.6041±0.0105
	FFCL	2.4M	0.5726±0.0092	0.4323±0.0182	0.5701±0.0079
	ST-Transformer	3.5M	0.6035±0.0081	0.4712±0.0199	0.6053±0.0075
Foundation Models	BIOT	3.2M	0.6153±0.0154	0.4875±0.0272	0.6158±0.0197
	LaBraM-Base	5.8M	0.6173±0.0122	0.4912±0.0192	0.6177±0.0141
	CBraMod	4.9M	0.6417±0.0091	0.5222±0.0169	0.6427±0.0100
	CSBrain	8.9M	0.6304±0.0090	0.5071±0.0120	0.6308±0.0095
	REVE-Base	69.2M	**0.6480±0.0140**	**0.5306±0.0187**	**0.6484±0.0170**
	CodeBrain	15.2M	0.6090±0.0070	0.4786±0.0094	0.6098±0.0077
Diffusion Features	EEGDM (supervised)	5.9M	0.3085±0.0474	0.0781±0.0633	0.2586±0.0677
	LUCID (ours, untrained)	11.3M	0.4106±0.0000	0.2140±0.0000	0.4057±0.0000
	LUCID (ours, trained)	11.3M	0.4805±0.0052	0.3073±0.0069	0.4816±0.0055

PhysioNet-MI (Goldberger et al., 2000). The PhysioNet Motor Movement/Imagery dataset contains 64-channel EEG at 160 Hz from 109 subjects. We keep the six motor-imagery runs of each subject, with four classes: left fist, right fist, both fists and both feet. Following CBraMod (Wang et al., 2025a), signals are common-average referenced, high-pass filtered at 0.3 Hz, notch filtered at 60 Hz and resampled to 200 Hz. One 4 s window is cut at each cue onset, giving 9,837 windows. We use the published subject split 1–70 / 71–89 / 90–109 (6,300/1,734/1,803 windows). Table 11 reports the results.

D.7 MENTAL DISORDER DIAGNOSTIC

Table 12: The results of different methods on Mumtaz (major depressive disorder versus healthy controls, 2-class).

	Method	Params	bACC	AUC-PR	AUROC
Task-specific Models	EEGNet	0.002M	0.9232±0.0104	0.9626±0.0095	0.9639±0.0093
	EEGConformer	0.8M	0.9308±0.0117	0.9684±0.0105	0.9702±0.0101
	SPaRCNet	1.0M	0.9316±0.0095	0.9754±0.0065	0.9781±0.0063
	ContraWR	1.6M	0.9195±0.0115	0.9589±0.0102	0.9621±0.0092
	CNN-Transformer	3.2M	0.9305±0.0068	0.9757±0.0074	0.9742±0.0059
	FFCL	2.4M	0.9314±0.0083	0.9717±0.0021	0.9753±0.0033
	ST-Transformer	3.4M	0.9135±0.0103	0.9578±0.0086	0.9594±0.0059
Foundation Models	BIOT	3.2M	0.9358±0.0052	0.9736±0.0034	0.9758±0.0042
	LaBraM-Base	5.8M	0.9409±0.0079	0.9798±0.0093	0.9782±0.0057
	CBraMod	4.9M	0.9560±0.0056	0.9923±0.0032	0.9921±0.0025
	CSBrain	8.9M	0.9643±0.0155	0.9936±0.0034	0.9956±0.0020
	REVE-Base	69.2M	**0.9644±0.0097**	**0.9961±0.0013**	**0.9957±0.0015**
	CodeBrain	15.2M	0.8982±0.0084	0.9715±0.0069	0.9680±0.0083
Diffusion Features	EEGDM (supervised)	5.9M	0.8962±0.0032	0.9758±0.0038	0.9732±0.0045
	LUCID (ours, untrained)	18.4M	0.8582±0.0003	0.8776±0.0001	0.9220±0.0000
	LUCID (ours, trained)	18.4M	0.8995±0.0068	0.9669±0.0053	0.9610±0.0065

Mumtaz (Mumtaz, 2016). The Mumtaz2016 dataset contains 19-channel resting EEG at 256 Hz from 34 patients with major depressive disorder and 30 healthy controls. Only the eyes-open and eyes-closed sessions are used. The task is to separate patients from controls. Following CBraMod, signals are resampled to 200 Hz, band-pass filtered at 0.3–75 Hz, notch filtered at 50 Hz and cut into 5 s windows, giving 7,143 windows. We use the split released with CBraMod and reused by CSBrain and REVE (4,891/1,041/1,211 windows). Table 12 reports the results.

D.8 EMOTION RECOGNITION

Table 13: The results of different methods on SEED-V (emotion recognition, 5-class).

	Method	Params	bACC	Cohen's κ	Weighted F1
Task-specific Models	EEGNet	0.003M	0.2961±0.0102	0.1006±0.0143	0.2749±0.0098
	EEGConformer	0.3M	0.3537±0.0112	0.1772±0.0174	0.3487±0.0136
	SPaRCNet	0.8M	0.2949±0.0078	0.1121±0.0139	0.2979±0.0083
	ContraWR	1.6M	0.3546±0.0105	0.1905±0.0188	0.3544±0.0121
	CNN-Transformer	3.2M	0.3678±0.0078	0.2072±0.0183	0.3642±0.0088
	FFCL	2.4M	0.3641±0.0092	0.2078±0.0201	0.3645±0.0132
	ST-Transformer	3.5M	0.3052±0.0072	0.1083±0.0121	0.2833±0.0105
Foundation Models	BIOT	3.2M	0.3837±0.0187	0.2261±0.0262	0.3856±0.0203
	LaBraM-Base	5.8M	0.3976±0.0138	0.2386±0.0209	0.3974±0.0111
	CBraMod	4.9M	0.4091±0.0097	0.2569±0.0143	0.4101±0.0108
	CSBrain	8.9M	**0.4197±0.0033**	**0.2785±0.0034**	**0.4280±0.0023**
	REVE-Base	69.2M	0.4048±0.0023	0.2622±0.0021	0.4145±0.0017
	CodeBrain	15.2M	0.4137±0.0023	0.2735±0.0032	0.4235±0.0022
Diffusion Features	EEGDM (supervised)	5.9M	0.3667±0.0038	0.2049±0.0044	0.3658±0.0058
	LUCID (ours, untrained)	9.9M	0.2424±0.0000	0.0511±0.0000	0.2390±0.0000
	LUCID (ours, trained)	9.9M	0.4028±0.0032	0.2465±0.0034	0.3954±0.0031

SEED-V (Liu et al., 2021). SEED-V contains 62-channel EEG at 1000 Hz from 16 subjects who each watched 15 film clips in three sessions, eliciting five emotions: disgust, fear, sad, neutral and happy. Signals are resampled to 200 Hz, band-pass filtered at 0.3–75 Hz (the recorder already notches the 50 Hz mains) and cut into 1 s windows, giving 117,744 windows. We use the provided split,

which divides the 15 trials of every session 5:5:5 by trial index into training, validation and test (34,432/42,960/40,352 windows). Table 13 reports the results.

Table 14: The results of different methods on FACED (emotion recognition, 9-class).

	Method	Params	bACC	Cohen's κ	Weighted F1
Task-specific Models	EEGNet	0.011M	0.4090±0.0122	0.3342±0.0251	0.4124±0.0141
	EEGConformer	1.5M	0.4559±0.0125	0.3858±0.0186	0.4514±0.0107
	SPaRCNet	1.2M	0.4673±0.0155	0.3978±0.0289	0.4729±0.0133
	ContraWR	1.6M	0.4887±0.0078	0.4231±0.0151	0.4884±0.0074
	CNN-Transformer	3.2M	0.4697±0.0132	0.4017±0.0168	0.4720±0.0125
	FFCL	2.5M	0.4673±0.0158	0.3987±0.0383	0.4699±0.0145
	ST-Transformer	3.4M	0.4810±0.0079	0.4137±0.0133	0.4795±0.0096
Foundation Models	BIOT	3.2M	0.5118±0.0118	0.4476±0.0254	0.5136±0.0112
	LaBraM-Base	5.8M	0.5273±0.0107	0.4698±0.0188	0.5288±0.0102
	CBraMod	4.9M	0.5509±0.0089	0.5041±0.0122	0.5618±0.0093
	CSBrain	8.9M	0.5752±0.0042	0.5204±0.0036	0.5796±0.0031
	REVE-Base	69.2M	0.5646±0.0164	0.5080±0.0191	0.5659±0.0172
	CodeBrain	15.2M	**0.5941±0.0098**	**0.5406±0.0084**	**0.5953±0.0113**
Diffusion Features	EEGDM (supervised)	5.9M	0.2053±0.0167	0.1082±0.0177	0.1966±0.0250
	LUCID (ours, untrained)	7.6M	0.1275±0.0000	0.0191±0.0000	0.1176±0.0000
	LUCID (ours, trained)	7.6M	0.2121±0.0086	0.1145±0.0098	0.2077±0.0085

FACED (Chen et al., 2023). FACED contains 32-channel EEG from 123 subjects watching 28 video clips that elicit nine emotions: anger, disgust, fear, sadness, neutral, amusement, inspiration, joy and tenderness. We start from the authors' processed release (250 Hz, band-pass filtered at 0.05–47 Hz, ICA-cleaned and common-average referenced). We add a 50 Hz notch, resample to 200 Hz and cut three 10 s windows per 30 s clip, giving 10,332 windows. We use the published subject split 1–80 / 81–100 / 101–123 (6,720/1,680/1,932 windows). Table 14 reports the results.

D.9 Imagined Speech Classification

Table 15: The results of different methods on BCIC2020-3 (imagined speech, 5-class).

	Method	Params	bACC	Cohen's κ	Weighted F1
Task-specific Models	EEGNet	0.004M	0.4413±0.0096	0.3016±0.0123	0.4413±0.0102
	EEGConformer	0.6M	0.4506±0.0133	0.3133±0.0183	0.4488±0.0154
	SPaRCNet	1.3M	0.4426±0.0156	0.3033±0.0233	0.4420±0.0108
	ContraWR	1.6M	0.4257±0.0162	0.3078±0.0218	0.4407±0.0182
	CNN-Transformer	3.2M	0.4533±0.0092	0.3166±0.0118	0.4506±0.0127
	FFCL	2.4M	0.4678±0.0197	0.3301±0.0359	0.4689±0.0205
	ST-Transformer	3.5M	0.4126±0.0122	0.2941±0.0159	0.4247±0.0138
Foundation Models	BIOT	3.2M	0.4920±0.0086	0.3650±0.0176	0.4917±0.0079
	LaBraM-Base	5.8M	0.5060±0.0155	0.3800±0.0242	0.5054±0.0205
	CBraMod	4.9M	0.5373±0.0108	0.4216±0.0163	0.5383±0.0096
	CSBrain	8.9M	0.6004±0.0187	0.5006±0.0233	0.6003±0.0192
	REVE-Base	69.2M	0.5635±0.0123	0.4543±0.0154	0.5633±0.0124
	CodeBrain	15.2M	**0.6101±0.0052**	**0.5127±0.0065**	**0.6101±0.0053**
Diffusion Features	EEGDM (supervised)	5.9M	0.2085±0.0139	0.0107±0.0174	0.1243±0.0271
	LUCID (ours, untrained)	11.2M	0.2533±0.0000	0.0667±0.0000	0.2521±0.0000
	LUCID (ours, trained)	11.2M	0.3296±0.0088	0.1620±0.0110	0.3290±0.0089

BCIC2020-3 (Jeong et al., 2022). Track 3 of the 2020 International BCI Competition contains 64-channel EEG at 256 Hz from 15 subjects silently imagining five phrases ("hello", "help me", "stop", "thank you", "yes"), 80 trials per phrase. Each 3 s trial is high-pass filtered at 0.5 Hz, notch filtered at 60 Hz and resampled to 200 Hz, giving 6,000 windows. We use the split provided by the

competition, 60/10/10 trials per phrase and subject for training, validation and test (4,500/750/750 windows). Table 15 reports the results.

D.10 SEIZURE DETECTION

Table 16: The results of different methods on Siena (seizure detection, 2-class).

	Method	Params	bACC	AUC-PR	AUROC
Task-specific Models	EEGNet	0.004M	0.7487±0.0521	0.3753±0.0867	0.8687±0.0527
	EEGConformer	1.5M	0.7556±0.0210	0.2091±0.0786	0.8159±0.0261
	SPaRCNet	1.1M	0.6572±0.0381	0.3164±0.0659	0.7334±0.0857
	ContraWR	1.6M	0.6546±0.0311	0.3711±0.0405	0.7819±0.0596
	CNN-Transformer	3.2M	0.6982±0.0560	0.3835±0.0709	0.8719±0.0403
	FFCL	2.5M	0.6616±0.0391	0.3938±0.0903	0.8154±0.1155
	ST-Transformer	3.4M	0.7527±0.0381	0.3636±0.0252	0.8884±0.0091
Foundation Models	BIOT	3.2M	0.7352±0.0669	0.3809±0.0892	0.9029±0.0304
	LaBraM-Base	5.8M	0.7082±0.0329	0.3122±0.0976	0.8814±0.0328
	CBraMod	4.9M	0.7317±0.0647	0.4107±0.0720	0.9038±0.0218
	CSBrain	8.9M	0.7662±0.0471	0.4871±0.0343	0.9076±0.0119
	REVE-Base	69.2M	0.7398±0.0074	0.3812±0.0531	0.8752±0.0296
	CodeBrain	15.2M	0.6733±0.0273	0.4276±0.0763	0.9206±0.0112
Diffusion Features	EEGDM (supervised)	5.9M	**0.8770±0.0159**	**0.5338±0.0297**	**0.9417±0.0037**
	LUCID (ours, untrained)	10.6M	0.5105±0.0000	0.1901±0.0000	0.8298±0.0000
	LUCID (ours, trained)	10.6M	0.8102±0.0261	0.4604±0.0363	0.8279±0.0171

Siena (Detti et al., 2020). The Siena Scalp EEG Database contains 41 clinical video-EEG recordings from 14 adult epilepsy patients at 512 Hz, with seizures annotated by clinicians. The task is binary seizure detection. Following CSBrain (Zhou et al., 2026), we keep the 29 channels common to all patients, band-pass filter at 0.5–75 Hz with a 50 Hz notch, cut 10 s windows and resample to 200 Hz, giving 51,307 windows, of which 637 contain a seizure. Patients PN16 and PN17 form the test set (3,679 windows). The other twelve are split 80/20 into training and validation within each patient (38,094/9,534 windows). Table 16 reports the results.

E Additional Analysis and Ablations

E.1 Frozen and Fine-Tuned Rows per Foundation Model

Figure 7 gives the fine-tuning gain of Section 3.3, FT − Y, per model rather than per dataset. Fine-tuning helps in 62 of the 72 cells, by 0.13 on average, and every loss falls on COMA, SLEEP, Mumtaz or Siena, four datasets of small gain (Section 3.3). The gain is a property of the model as much as of the dataset. BIOT, LaBraM and CBraMod gain 0.16 to 0.17 on average, CSBrain 0.13, and REVE and CodeBrain 0.08. These two read best frozen, on nine of the twelve datasets, and so have the least to gain from being updated, and REVE loses accuracy on four datasets.

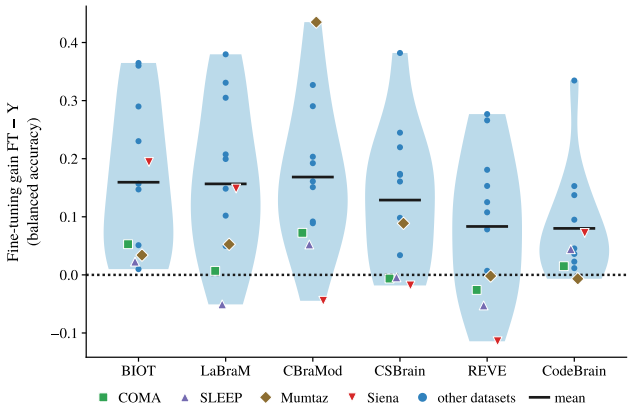


Figure 7: Fine-tuning gain per foundation model: FT − Y in balanced accuracy on the twelve datasets, from the values of Figure 4, one violin and twelve points per model; bar: mean over the datasets; dotted line: no gain, above it fine-tuning helps and below it hurts. COMA, SLEEP, Mumtaz and Siena, the four datasets on which a model loses accuracy, have their own marker; the other eight are plain dots.

E.2 Label Efficiency

Figure 8 refits the probe on a class-stratified fraction of the labelled training windows, with the noise level and penalty of Table 2, and probes the untrained copy on the same subsets. With a quarter of the labels, LUCID matches its full-label balanced accuracy, within seed spread, on COMA, TUAB, TUEV, MAT, Mumtaz and Siena. On COMA, TUAB and Mumtaz a twentieth is nearly enough. The six other datasets keep four fifths (BCIC2020-3) to nineteen twentieths (SLEEP, SEED-V) of their full-label accuracy. Balanced accuracy has a chance floor, so the share of the margin above chance that they keep is lower, from half on BCIC2020-3 to nine tenths on SEED-V. The two groups differ in what a label marks. A label that holds for a recording, or recurs within it, is shared by windows that largely repeat one another, so a few of them stand for the rest. A label that changes from trial to trial, as in motor imagery, imagined speech and emotion, makes each trial a new sample of a weak signal, and every added label still helps. SLEEP follows the second group although its label holds for a recording. In the figure, the untrained copy is below LUCID from the lowest fraction on every dataset, and the distance widens with the labels on the trial-level tasks. The copy starts as the identity, so its features are the per-channel mean and standard deviation of the noised window whatever the seed: its five seeds agree, and its full-label bars have almost no spread. Probed without class weighting, as in Table 2, it predicts the majority class on COMA and Siena, hence chance. Windows are subsampled, not subjects, so the curves measure fewer labels per subject, not fewer labelled subjects.

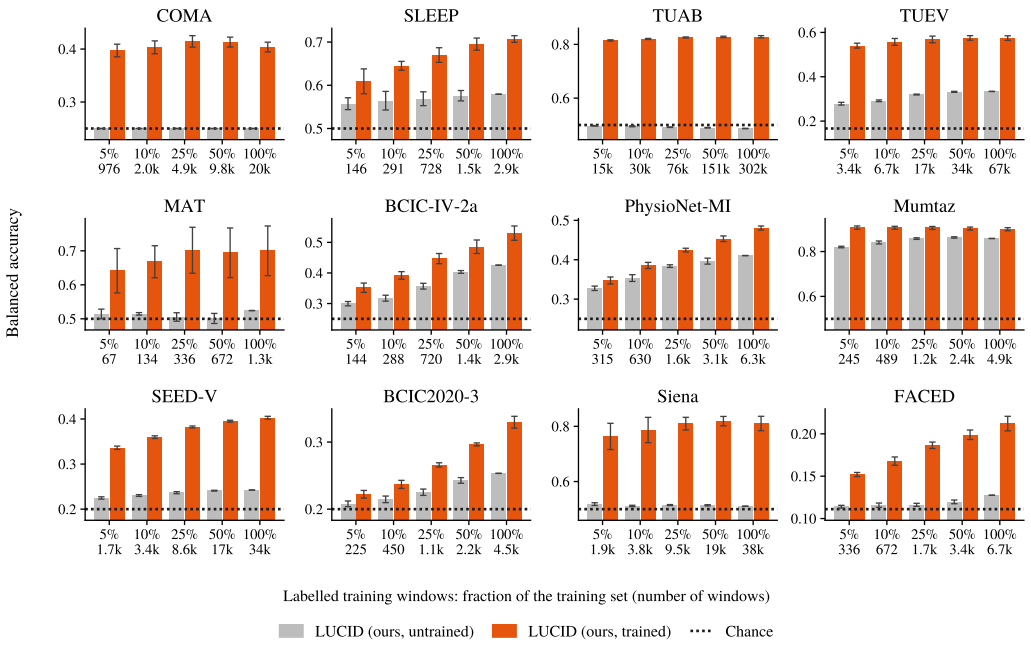


Figure 8: Label efficiency: balanced accuracy of LUCID (orange) and of its untrained copy (grey) against the fraction of labelled training windows (number of windows under each pair), one panel per dataset; mean and std over the five seeds. Both are probed on the same subsets at the noise level of Table 2, each with its own full-label penalty, fixed along the curve, the untrained copy without class weighting on COMA, Mumtaz and Siena; windows subsampled with class stratification, five subsets per seed; the 100% bars are the Table 2 values. Dotted line: chance.

E.3 Ablations

E.3.1 Embedding Width

Figure 9 asks whether the probe owes its accuracy to the width of the embedding. The width is 2,560 to 14,592 on ten datasets, 131,072 on PhysioNet-MI and 360,448 on BCIC-IV-2a, where the 16- and 64-chunk pooling of Table 4 adds a mean and a standard deviation per feature for every chunk. A PCA fitted on the training split, without labels, reduces each embedding, and the penalty is re-selected at each width. At 256 components, a tenth of the width or less, TUAB, Mumtaz, Siena, BCIC-IV-2a and MAT keep their full accuracy, and the seven other datasets keep half to four fifths of their margin over chance. The accuracy does not rest on the width. The sweep also shows where the signal sits. COMA, SLEEP, TUEV and BCIC2020-3 drop at the first cut, to 1,024 components, and not again, although that cut removes under two percent of the variance: their signal lies in many low-variance directions, which the probe sums and PCA discards first. PhysioNet-MI, SEED-V and FACED decline with every halving, their signal spanning directions of comparable weight. BCIC-IV-2a keeps its accuracy with under half of its variance, so most of its chunked width is redundant for the probe.

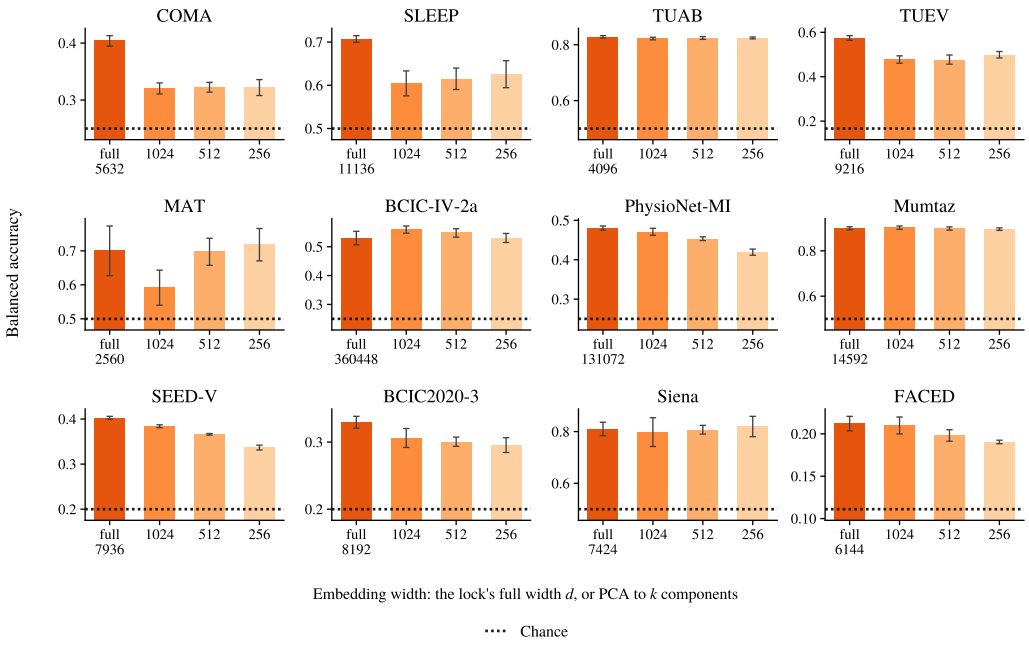


Figure 9: Embedding width: balanced accuracy of LUCID with its full cross-layer embedding (width d under the bar) and after a train-fitted PCA to 1024, 512 and 256 components, one panel per dataset; mean and std over five seeds, noise level fixed, penalty re-selected on validation for each width. Dotted line: chance.

E.3.2 XLAP Loss

Loss weight. Figure 10 varies the XLAP weight λ from a tenth to fifty times the value at which the XLAP and denoising gradients balance ($r = 1$), on BCIC-IV-2a with the 64-chunk pooling of Table 4. From a tenth to ten times the balance, the probe moves within its seed spread on MAT, BCIC-IV-2a and SEED-V, and by at most 0.015 on BCIC2020-3. Beyond ten times, it falls: by 0.06 to 0.08 on BCIC2020-3 and MAT at fifty times, and by 0.02 on SEED-V, which holds until twenty-five times. The auxiliary objective then competes with the pretext for the parameters they share. BCIC-IV-2a dips at twenty-five times and is back at fifty, with another noise level selected. The Table 2 models sit on this plateau. PhysioNet-MI, where Figure 5 shows the clearest XLAP effect, is not among the four datasets swept.

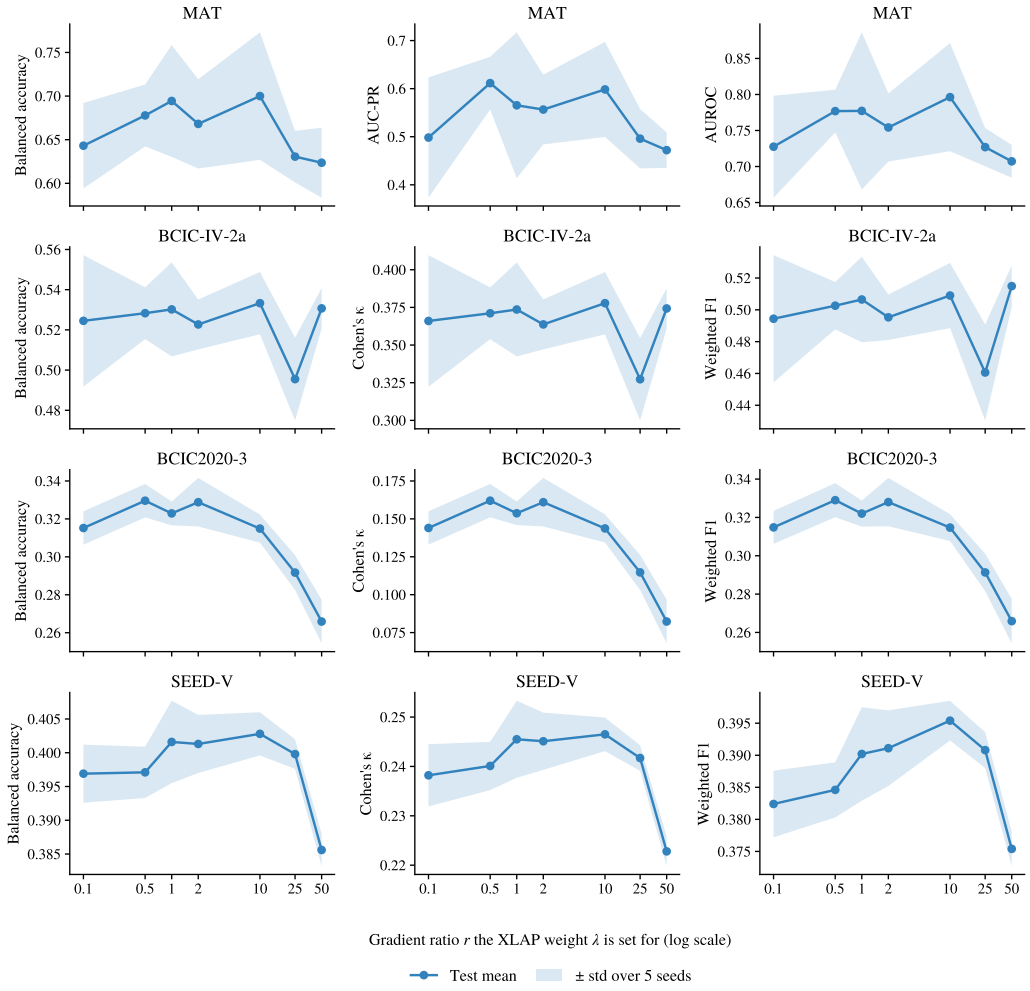


Figure 10: Weight λ of the XLAP loss: test performance against the gradient ratio r that λ is set for, one row per dataset and one panel per metric; line and band: mean and std over five seeds, each point probed with its own noise level and penalty; on BCIC-IV-2a with the 64-chunk pooling of Table 4.

Masking ratio. Figure 11 varies the fraction ρ of channels XLAP masks, with λ held at the value of the $r = 1$ model. From a third of the channels to nine in ten, the ratio makes no difference: on both datasets the probe moves within its seed spread. Masking a tenth is the one setting that costs, on MAT, by 0.10: with nearly every channel visible, the hidden statistics can be read off the neighbours, and the objective has little left to teach. SEED-V is indifferent throughout. Two datasets, one with four test subjects, bound what the sweep can say.

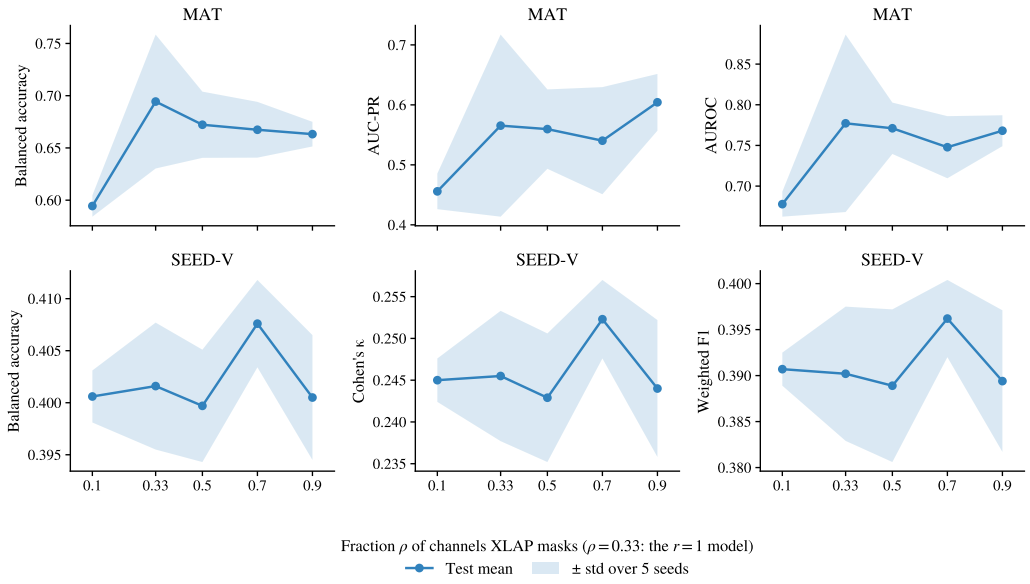
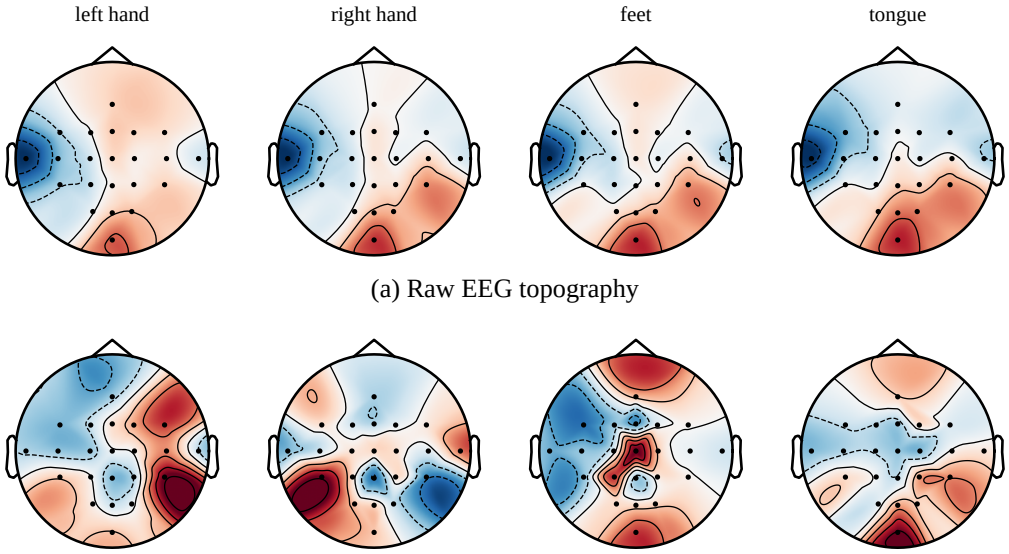


Figure 11: Masking ratio ρ of XLAP: test performance against the fraction of channels XLAP masks during training, with λ held at the value of the $r = 1$ model ($\rho = 0.33$ is that model), one row per dataset and one panel per metric; line and band: mean and std over five seeds, each point probed with its own noise level and penalty.

F Interpretability Analysis

F.1 Topography Visualisation

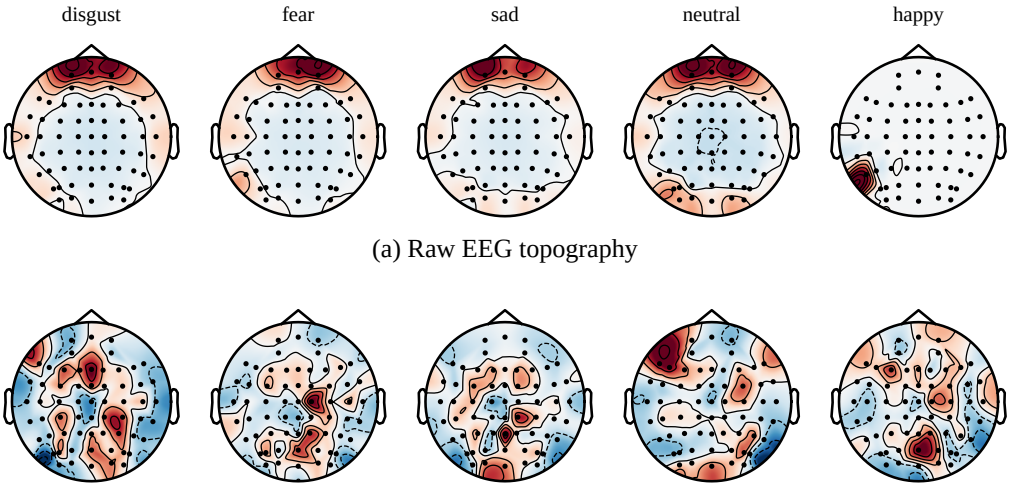
To see which electrodes LUCID's predictions rely on, Figures 12 and 13 show two topographies per class, averaged over the class's test windows: (a) the raw EEG energy, the time average of the squared unnormalised signal at each electrode; (b) LUCID's attribution, the time average of $|x_0 \odot \partial \ell_k / \partial x_0|$, where x_0 is the normalised window and ℓ_k the probe's logit for class k, differentiated through the frozen denoiser at the probe's noise level. This input×gradient attribution (Shrikumar et al., 2017) replaces the usual Grad-CAM (Selvaraju et al., 2017), which projects hidden activations onto electrodes: every block of the denoiser mixes electrodes, so after the first block a hidden unit no longer belongs to a single electrode. On BCIC-IV-2a, the raw energy has almost the same spatial pattern in the four classes, whereas LUCID's attribution changes from class to class and follows the somatotopy of the motor cortex: it peaks at the contralateral centro-parietal electrode for hand imagery, CP4 for the left hand and CP3 for the right, and at Cz, over the foot area, for feet imagery, while for tongue imagery it peaks further back, at POz. The contralateral hand pattern appears in all five seeds, and Cz is the largest electrode for feet in four of them. On SEED-V, the raw energy is dominated by the frontal-pole electrodes, the most exposed to eye movements, in four of the five classes, and in the fifth, happy, by a single posterior electrode, P7. LUCID's attribution never peaks at the frontal pole: its largest values lie over lateral frontal, central, parietal and occipital electrodes, and which of these dominates an emotion changes from seed to seed. Since each electrode is standardised before entering LUCID, part of this difference from the raw energy reflects the input normalisation rather than the model. LUCID's attribution thus takes a different shape for each task: focal, seed-stable peaks over the sensorimotor cortex for motor imagery, and broader, less stable



(a) Raw EEG topography

(b) LUCID attribution topography (|input × gradient|)

Figure 12: Topography on BCIC-IV-2a, one column per class, averaged over the class's test windows: (a) raw EEG energy; (b) LUCID's input×gradient attribution. Each map is centred on its own spatial mean, red above and blue below, and has its own colour scale, so panels compare by spatial pattern only. Maps from one of the five seeds.



(a) Raw EEG topography

(b) LUCID attribution topography (|input × gradient|)

Figure 13: Topography on SEED-V, one column per class, drawn as in Figure 12.

centro-parietal and occipital patterns for emotion. The linear probe on LUCID's frozen features therefore relies on spatially plausible, task-specific electrodes, most clearly on BCIC-IV-2a.

F.2 Embedding Visualisation

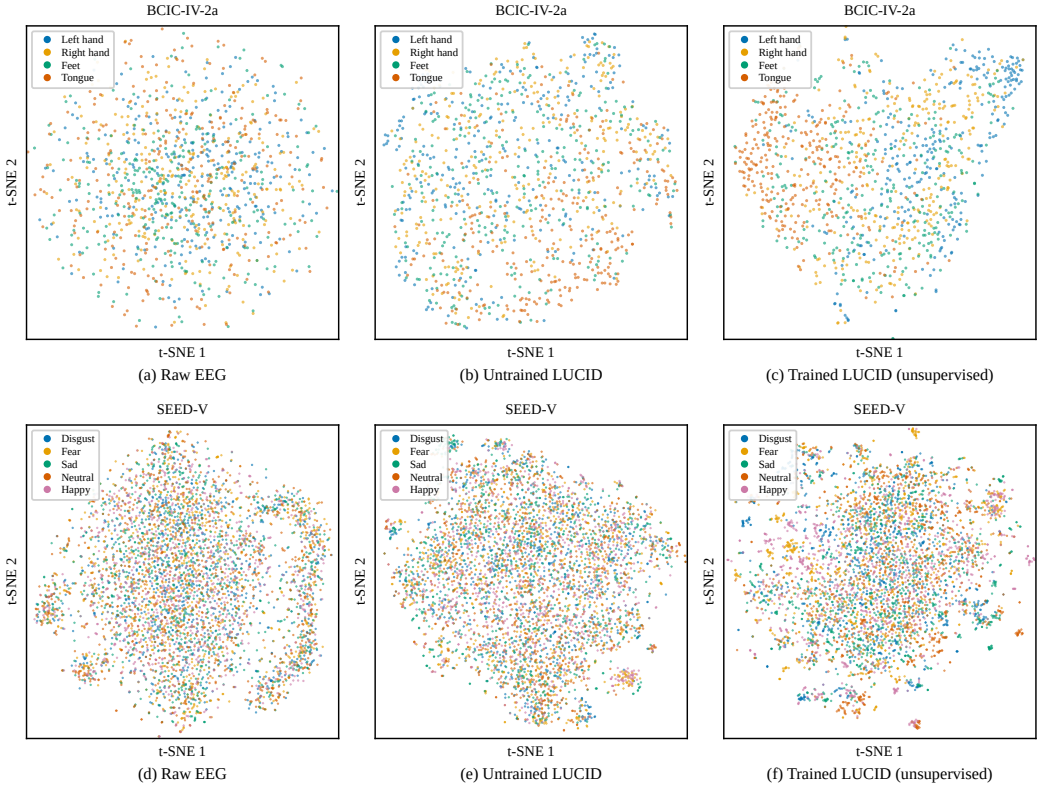


Figure 14: t-SNE of held-out test data, coloured by class: one point per trial on BCIC-IV-2a (top) and per 1 s window on SEED-V (bottom). BCIC-IV-2a: (a) raw EEG; (b) the untrained copy of LUCID; (c) LUCID, the features its probe reads. SEED-V: (d) raw EEG; (e) the untrained copy of LUCID; (f) LUCID. Labels only colour the points. Maps from one of the five seeds.

The top row of Figure 14 maps the 1,152 test trials of BCIC-IV-2a, from the two held-out subjects, for three representations built by the same label-free pipeline: (a) the raw EEG windows, (b) the untrained copy of LUCID, and (c) LUCID, trained without labels and frozen, whose features the probe reads. Features are standardised, each recording session's mean is removed, and PCA precedes t-SNE (van der Maaten & Hinton, 2008). In LUCID's representation, trials of the same class occupy broad regions of the map, whereas they largely mix in the raw EEG and in the untrained copy. The classes still overlap rather than form separate clusters: tongue and left-hand trials gather at opposite ends of the map, with right-hand and feet trials in between.

The bottom row, on 4,800 of SEED-V's 1 s test windows, 20 drawn at random from each of the 240 held-out film clips, shows the limit of this unsupervised embedding. The raw EEG (d) and the untrained copy (e) each form a largely mixed cloud; LUCID (f) breaks into small single-colour blobs around a mixed centre, but the blobs are the windows of single film clips, single-coloured only because each clip carries one emotion label. Trained without labels, LUCID's embedding thus reflects which recording a window comes from more than which emotion it carries. The emotion information is not absent, since the supervised linear readout decodes it above chance from the same frozen features (Table 13), but it barely shapes the label-free map. The contrast also shows why the BCIC-IV-2a map can be read as class structure: there, every trial is its own recording segment and the classes are balanced within each session, so segment identity cannot make a trial's neighbourhood look class-pure.

G GENERATION METRICS

Each LUCID denoiser is also a generator. Its samples can be compared with real windows of the same dataset on spectral (band powers, power spectra), temporal (autocorrelation) and spatial (inter-electrode coherence) statistics. Sample quality is not the object of this paper. The read-out never runs the sampling chain of Equation (5), and none of the quantities below selected a configuration, noise level or seed. We report one summary number on five datasets and a visual comparison on two of them.

G.1 BAND-POWER WASSERSTEIN-1 DISTANCE

Following B[FM]^2 (Hwang et al., 2026), we draw from each denoiser one unconditional sample per training window (ten thousand on SEED-V). For each electrode and each of the five canonical bands, δ (0.5–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz) and γ (30–45 Hz), we compute the Wasserstein-1 (W_1) distance between the distributions of log band power of the generated windows and of the held-out real test windows, then average over electrodes. Each distance is read against a real-versus-real floor: the same distance between two disjoint sets of real windows, drawn from the training split at the sizes of the test comparison, since a finite-sample estimate is positive even between two real sets. The floor depends on the real data only, so it is one value per dataset. Table 17 reports both, per band and averaged over the five bands, as mean and standard deviation over the training seeds. Averaged over bands, the generators of BCIC2020-3, Mumtaz and MAT sit at their floor, within six percent of it, and those of SEED-V and BCIC-IV-2a within two and a half times it. The bands are not equally hard. On MAT, Mumtaz and SEED-V the largest distance is in γ, the band with the least power. On BCIC-IV-2a and BCIC2020-3 it is in α. BCIC-IV-2a shows the widest spread, from below its floor in γ to five times it in α, the band that motor imagery modulates and the largest distance of the table.

Table 17: Sample quality of LUCID on five datasets: Wasserstein-1 distance between the log band power of generated windows and of held-out real test windows, per band and averaged over the five bands (mean $\pm$ std over five training seeds). Real-vs-Real is the noise-floor baseline, the same distance between two disjoint sets of real windows. Ratio: the band-averaged distance over its floor, so one is the floor itself. Lower distances are better.

Dataset		$\delta\downarrow$	$\theta\downarrow$	$\alpha\downarrow$	$\beta\downarrow$	$\gamma\downarrow$	Mean $\downarrow$	Ratio
MAT	Real-vs-Real	0.098	0.115	0.174	0.131	0.192	0.142	–
	LUCID	0.084 ± 0.005	0.107 ± 0.002	0.204 ± 0.016	0.131 ± 0.015	0.227 ± 0.005	0.150 ± 0.006	1.06 ± 0.04
BCIC-IV-2a	Real-vs-Real	0.164	0.053	0.309	0.292	0.553	0.274	–
	LUCID	0.411 ± 0.012	0.649 ± 0.007	1.516 ± 0.022	0.464 ± 0.019	0.183 ± 0.011	0.644 ± 0.011	2.35 ± 0.04
Mumtaz	Real-vs-Real	0.263	0.308	0.263	0.275	0.363	0.294	–
	LUCID	0.129 ± 0.022	0.279 ± 0.033	0.249 ± 0.025	0.341 ± 0.015	0.539 ± 0.043	0.307 ± 0.018	1.04 ± 0.06
SEED-V	Real-vs-Real	0.072	0.067	0.104	0.120	0.161	0.105	–
	LUCID	0.228 ± 0.023	0.124 ± 0.012	0.197 ± 0.007	0.227 ± 0.020	0.361 ± 0.037	0.227 ± 0.007	2.17 ± 0.07
BCIC2020-3	Real-vs-Real	0.383	0.209	0.427	0.312	0.467	0.359	–
	LUCID	0.419 ± 0.009	0.235 ± 0.006	0.438 ± 0.034	0.290 ± 0.033	0.330 ± 0.057	0.342 ± 0.019	0.95 ± 0.05

G.2 VISUAL COMPARISON ON SEED-V AND MUMTAZ

Figures 15 to 18 compare held-out real windows with generated ones on two datasets of Table 17: SEED-V, whose generator is at twice its floor, and Mumtaz, whose generator sits at it. Each figure shows one training seed.

Band-power topomaps. Figure 15 maps the share of each band in the power of each electrode. On SEED-V the generated maps reproduce the real layout in every band: frontal δ with temporal minima, the two θ maxima, the bilateral parietal α maxima, and the temporal β and γ maxima. What shifts is the balance between bands. The generated windows put a little more of their power in δ and a little less in β and γ, a tilt towards low frequencies that the spectra below show as well. On Mumtaz the generated maps reproduce frontal δ, posterior α and the temporal β and γ maxima, with the same tilt. The real θ map is patchy, with a right-frontal maximum, where the generated one has a single fronto-central maximum. The real map of the highest band is a single frontal electrode, which the generated map does not single out.

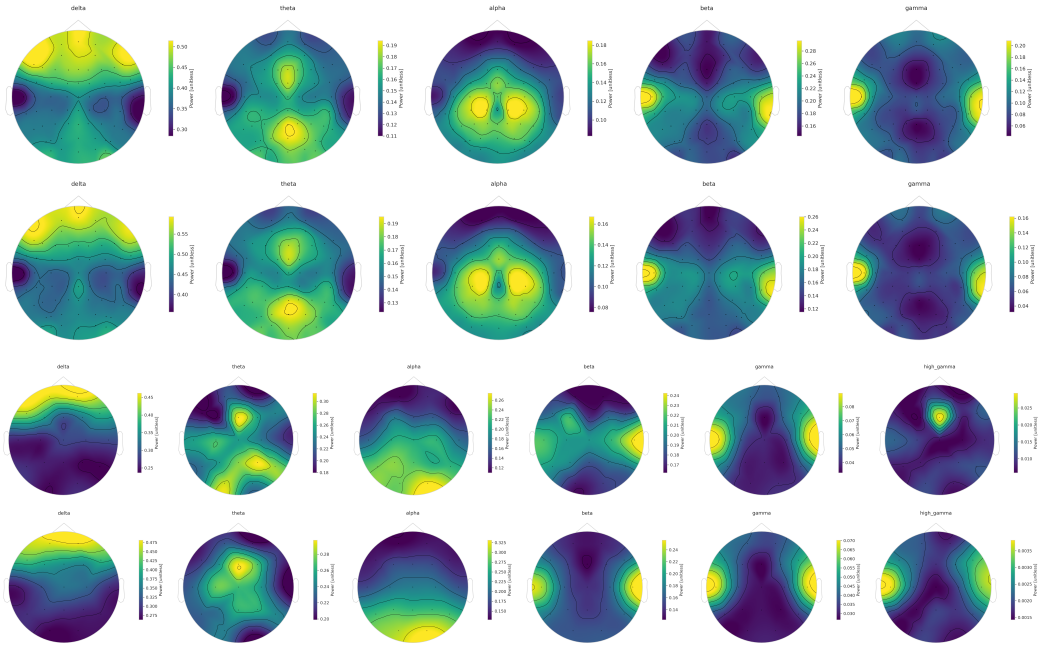


Figure 15: Band-power topomaps, real above generated: the share of each band in each electrode's power, averaged over held-out windows. SEED-V (upper pair) and Mumtaz (lower pair, with a 45–95 Hz band). One training seed.

Autocorrelation. Figure 16 shows the autocorrelation of real and generated windows, averaged over electrodes and windows, up to a quarter of the window length. SEED-V's 1 s windows leave a quarter of a second, within which both curves fall to zero, the generated one a little above the real one before 0.1 s and a little below it after. On Mumtaz the real curve carries a rhythm near 5 Hz, one lobe every 0.19 s. The generated curve misses the first lobe, dipping where the real curve rises, then follows the real curve to within the real-versus-real floor.

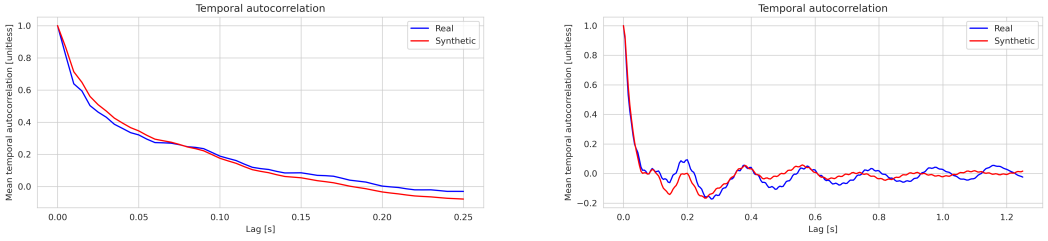


Figure 16: Temporal autocorrelation of real held-out windows (blue) and generated windows (red), averaged over electrodes and windows, against the lag. SEED-V (left) and Mumtaz (right). One training seed.

Power spectra. Figure 17 overlays the power spectral density of real and generated windows on every electrode, median and interquartile range over windows. On SEED-V the two coincide on nearly every electrode, the α bump of the parietal and occipital electrodes included. The tilt is slight, a shift of the spectrum's mean frequency of about two hertz, and on a few central and temporal electrodes the real range is wider than the generated one. On Mumtaz the medians coincide from 2 Hz to the 50 Hz notch. Two things differ. A narrow peak near 5 Hz, present in the upper quartile of the real windows on most electrodes, hence in a subset of windows, is absent from the generated ones. And below 2 Hz the generated windows carry less power than the real ones over their whole range.

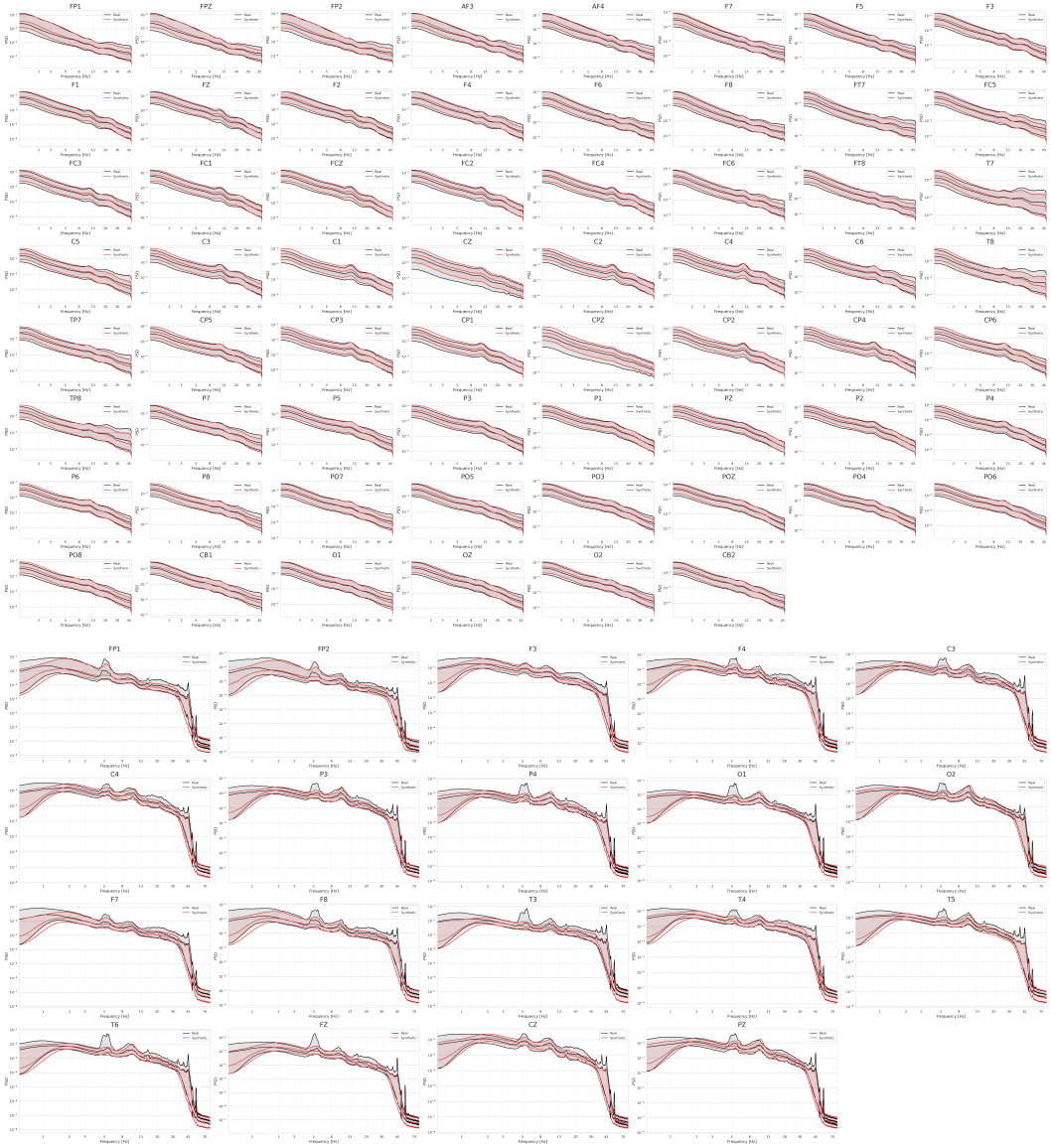


Figure 17: Power spectral density of real held-out windows (black) and generated windows (red) on every electrode: median and interquartile range over windows, log–log axes. SEED-V (top) and Mumtaz (bottom). One training seed.

Coherence topomaps. Figure 18 maps the mean coherence of each electrode with all the others, per band. On SEED-V the generated maps reproduce the central minimum of the three lower bands and the frontal and temporal minima of β and γ, at a level a few hundredths higher. They lack the deep minimum that the real maps show on one left temporo-parietal electrode. On Mumtaz they reproduce the central plateau of δ and α and the frontal and occipital minima of δ. The real β and γ maps are patchy and lean to the left, where the generated ones are smooth and central. Little can be read into that: two halves of the real data agree with each other no better than the generated maps agree with the real ones. Sample quality does not predict the probe. Mumtaz, at its floor, is among the datasets where training adds least over the untrained copy, and SEED-V, at twice its floor, among those where it adds most and where LUCID comes closest to the best fine-tuned model (Table 2).

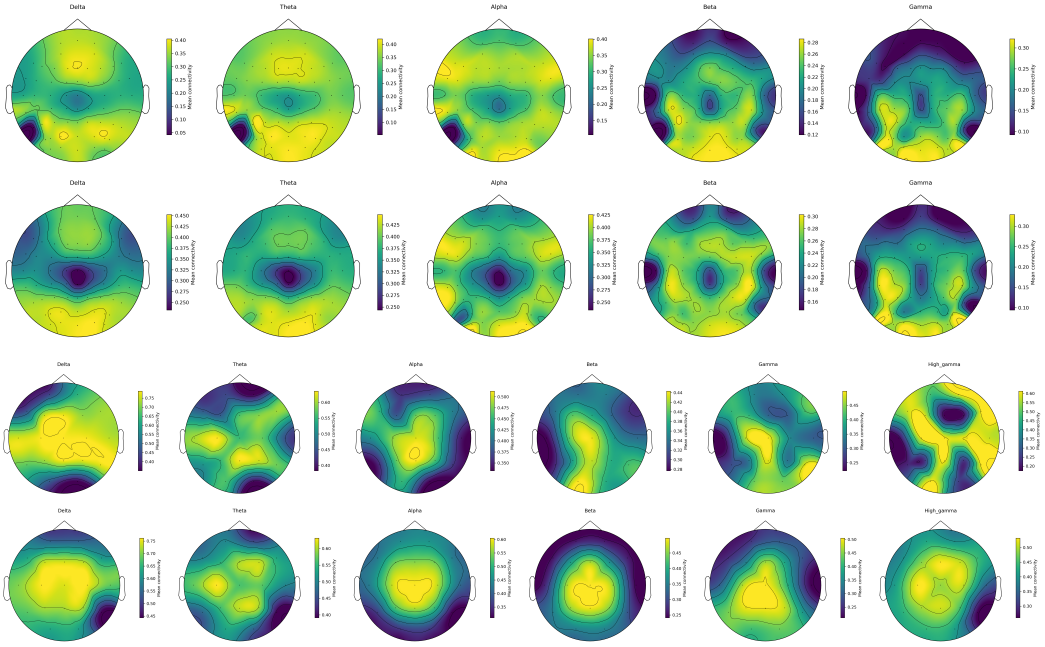


Figure 18: Coherence topomaps, real above generated: the mean coherence of each electrode with all the others, averaged over held-out windows. SEED-V (upper pair) and Mumtaz (lower pair, with a 45–95 Hz band). One training seed.

H Model Complexity

Taking SEED-V and BCIC-IV-2a as examples, the first with many channels and short windows, the second with fewer channels and longer windows, Table 18 compares the parameters and FLOPs[6] of LUCID and the baselines. We recompute both for every model rather than taking the published counts.[7]

On BCIC-IV-2a, LUCID's backbone (5.6M) is smaller than LaBraM-Base, CSBrain, CodeBrain, REVE-Base and EEGDM. On SEED-V (9.9M), it is smaller than CodeBrain and REVE-Base. REVE-Base (El Ouahidi et al., 2026), the largest model, has 7 and 12 times LUCID's parameters on SEED-V and BCIC-IV-2a, and costs 2.3 and 1.6 times its FLOPs. EEGDM (Puah et al., 2025), the other diffusion-based representation learner, costs 8.2 and 6.5 times LUCID's FLOPs. Its backbone runs every channel separately through 20 bidirectional S4 blocks, and its trained 6.9M-parameter fusion head is counted in its FLOPs but not in its parameters. The compact encoders (BIOT, LaBraM-Base, CBraMod, CSBrain and CodeBrain) are cheaper to run. LUCID costs 2.2 to 9.7 times their FLOPs, because its denoiser works at the full time resolution of the signal, whereas they compress each 1 s patch into a single token. This extra cost reflects what LUCID is: a generative model of the whole

[6] PyTorch's `FlopCounterMode` (torch 2.8, `https://github.com/pytorch/pytorch/blob/main/torch/utils/flop_counter.py`), reported as multiply–accumulate operations (MACs).

[7] CBraMod's checkpoint loads as 4.9M parameters, against the 4.0M given in its paper (Wang et al., 2025a). We count it the same way as the other models.

Table 18: Parameters and FLOPs on SEED-V (62 channels, 1 s) and BCIC-IV-2a (22 channels, 4 s). Supervised baselines are built at each dataset's input shape, so some of their sizes differ between the two datasets.

		SEED-V		BCIC-IV-2a	
	Method	Params	FLOPs	Params	FLOPs
Task-specific Models	EEGNet	0.003M	6.6M	0.003M	9.4M
	EEGConformer	0.3M	29.3M	0.6M	51.4M
	SPaRCNet	0.8M	16.9M	1.0M	50.7M
	ContraWR	1.6M	32.9M	1.6M	64.4M
	CNN-Transformer	3.2M	34.5M	3.2M	87.9M
	FFCL	2.4M	1.02G	2.4M	428.5M
	ST-Transformer	3.5M	10.5M	3.4M	40.8M
Foundation Models	BIOT	3.2M	200.8M	3.2M	498.8M
	LaBraM-Base	5.8M	382.7M	5.8M	551.8M
	CBraMod	4.9M	315.0M	4.9M	449.9M
	CSBrain	8.9M	681.5M	8.9M	961.0M
	REVE-Base	69.2M	4.37G	69.2M	6.26G
	CodeBrain	15.2M	867.1M	15.2M	1.26G
Diffusion Features	EEGDM	5.9M	15.9G	5.9M	25.6G
	LUCID w/o CoSMO & XLAP	8.3M	1.61G	4.8M	3.30G
	LUCID w/o XLAP	9.9M	1.94G	5.6M	3.95G
	LUCID	9.9M	1.94G	5.6M	3.95G

EEG signal, every channel and time step, trained with an unsupervised diffusion objective rather than fine-tuned to one prediction. It is justified by the results of Table 2, not by this table. Finally, CoSMO adds 17–19% parameters and about 20% FLOPs. XLAP, a training-only loss, adds nothing at inference, which is why the last two rows coincide.