ZERO-SHOT GENERATION OF THE NEXT MOMENT OF BRAIN ACTIVITY BY PREDICTING ITS DISTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG is a continuous, highly stochastic signal dominated by broadband noise and non-stationary fluctuations, so a pointwise autoregressive model chases noise it cannot reliably estimate. We therefore regard brain activity as intrinsically stochastic: the next moment of EEG is governed by a probability distribution. We propose **GenBrain**, a generative autoregressive model of EEG that predicts this distribution directly and generates brain activity by sampling from it. Grounded in the computational-neuroscience account of EEG as rhythmic oscillations riding on an aperiodic $1/f$ background, we parameterize the distribution with Rayleigh amplitudes, von Mises phases, phase-amplitude coupling, and a $1/f$ spectral slope, so that every predicted parameter carries a neurophysiological meaning. To capture transient bursts and spikes beyond this parameterization, a dual-head design pairs distribution prediction with zero-initialized raw-signal reconstruction; both heads share a block-causal Transformer backbone and train jointly, while neurophysiological priors enter as differentiable soft regularizers. Pretrained on 14 public EEG corpora and evaluated zero-shot on 10 held-out datasets spanning six paradigms, GenBrain attains lower teacher-forced negative log-likelihood than marginal, persistence, and autoregressive baselines on all 10 datasets (32–62% relative improvement, up to 7.0 nats below the strongest baseline); probability-integral-transform residuals localize where the parameterized family's tails fall short; and autoregressive generation improves spectral fidelity on 9 of 10 datasets (PSD correlation up to 0.94) while holding amplitude calibration at or below the real–real floor on 8 of 10. GenBrain recasts the modeling of brain activity as distribution prediction, offering a novel solution to the brain's intrinsic stochasticity.

1 INTRODUCTION

Electroencephalography (EEG) records the synchronous activity of cortical neural populations with millisecond resolution and underpins brain-computer interfaces, neurological diagnosis, and cognitive neuroscience (Wolpaw et al., 2002; Chaudhary et al., 2016; Lebedev & Nicolelis, 2017). Translating EEG into robust, general-purpose tools is bottlenecked by three problems: data are scarce and expert labels costly (Lotte et al., 2018; Keutayeva & Abibullaev, 2024; Eldawlatly, 2024); inter-subject variability, heterogeneous montages, and sampling rates erode cross-subject, cross-task, and cross-device generalization (Saha & Baumert, 2020; Shen et al., 2023; Wang et al., 2024b); and the dynamics are nonlinear, non-stationary, and organized into cross-frequency coupled rhythms that point-based models struggle to represent (Breakspear, 2017; Strogatz, 2024; Buzsáki, 2006; Canolty et al., 2006).

Costly labels have driven a shift toward foundation models pretrained on large unlabeled EEG corpora (Jiang et al., 2024b; Yang et al., 2023; Wang et al., 2025; 2026). Their pretraining objectives, however, remain predominantly discriminative: inherited from masked language modeling (Devlin et al., 2018) and masked autoencoders (He et al., 2022), they reconstruct occluded segments, do not characterize the generative process that produces EEG, and offer no natural pathway to sampling new signals (Appendix A).

A different paradigm transformed natural language processing: decoder-only autoregressive models predict the distribution of the next token given history, one objective that yields both strong representations and fluent generation (Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023).

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.

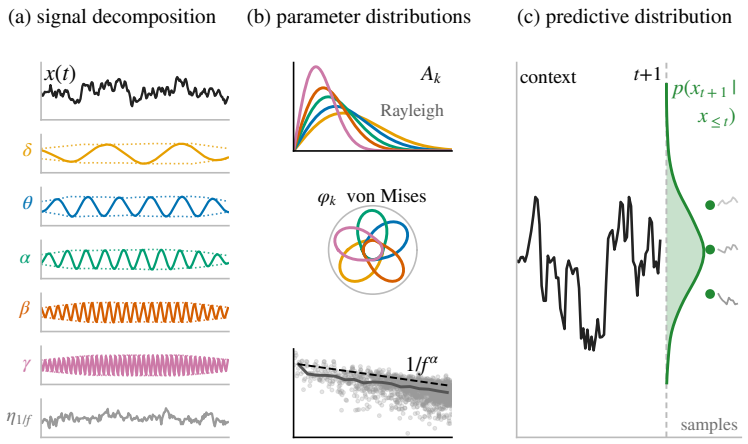

Figure 1: The generative view of EEG underlying GenBrain. **(a)** One time step of EEG decomposes into oscillations in five canonical frequency bands (δ, θ, α, β, γ) riding on a $1/f$ background (Equation 1). **(b)** The parameters of this decomposition are random variables with neurophysiologically grounded distributions: oscillatory envelopes follow Rayleigh distributions, phases in coherent regimes follow von Mises distributions, and the background spectrum obeys a power law. **(c)** Autoregressively predicting these parameters yields a full predictive distribution over the next time step, from which waveforms are sampled rather than regressed.

Migrating this idea to EEG meets a fundamental obstacle: EEG is a continuous, highly stochastic signal dominated by broadband noise and non-stationary fluctuations, so a model that predicts the next sample value pointwise is forced to chase high-frequency noise it cannot reliably estimate.

Our central observation is that the autoregressive target should not be the next sample value, but the *distribution* that the next time step of EEG follows. We draw this distribution from the computational-neuroscience account of EEG as band-limited rhythmic oscillations riding a $1/f$ spectral background, with cross-frequency coupling linking low-frequency phase to high-frequency amplitude (Buzsáki, 2006; Canolty et al., 2006). A single time step of EEG can thus be cast as

$$x(t) = \sum_{b} \sum_{k \in b} A_k \cos(2\pi f_k t + \varphi_k) + \eta_{1/f}(t), \tag{1}$$

where A_k, φ_k, and the spectral slope α of the $1/f$ noise $\eta_{1/f}$ are random variables governed by neurophysiological mechanisms (Figure 1): narrow-band oscillatory envelopes follow Rayleigh distributions, phases in coherent regimes follow von Mises distributions, and the background spectrum obeys a power law. Autoregressively predicting the parameters of this compound distribution, rather than the waveform itself, gives the model a well-defined, likelihood-computable, sampleable target grounded in neural mechanism.

Building on this idea, we propose **GenBrain**, a generative autoregressive model of EEG that, once pretrained, operates zero-shot across unseen subjects, montages, and datasets. GenBrain predicts a neuro-mechanistically parameterized distribution of the next time step and is trained by maximum likelihood; generation reduces to sampling from the predicted distribution and rolling the sampled segment back into the context. A *block-causal architecture* preserves strict causality across time steps while letting channels recorded at the same instant attend bidirectionally, and neurophysiological priors enter as differentiable soft regularizers. Finally, to capture transient bursts and spikes that lie beyond any parameterized family, a *dual-head* design lets a distribution-prediction head and a zero-initialized raw-signal reconstruction head share one backbone.

Our contributions are fourfold:

- **GenBrain: a generative autoregressive model of EEG.** We propose **GenBrain**, which recasts EEG pretraining from masked reconstruction to *distribution prediction*: the model predicts the probability distribution that the next time step of EEG follows and generates brain activity by sampling from it, so that a single pretrained model unifies representation,

physiologically plausible generation, and interpretable prediction—a generative capability absent from discriminative foundation models.

- **Model architecture.** We design a *block-causal Transformer* in which channels recorded at the same time step attend bidirectionally and time steps remain strictly causal, preserving autoregressive semantics while capturing cross-channel structure; a *dual-head* design pairs a distribution-prediction head with a zero-initialized reconstruction head on one backbone.
- **Pretraining objective.** We define a *neuro-mechanistically parameterized EEG distribution family*—Rayleigh amplitudes, von Mises phases, a $1/f$ aperiodic background, and phase-amplitude coupling—in which every predicted parameter carries a neurophysiological meaning while remaining likelihood-computable and sampleable, and build the pretraining objective by coupling *distribution prediction*, *raw-signal reconstruction* that absorbs out-of-family transients, and *neurophysiological constraints* as differentiable soft priors, trained in a single stage.
- **Evaluation.** We validate GenBrain zero-shot on 10 held-out datasets with a unified generative protocol: one-step negative log-likelihood against statistical baselines, probability-integral-transform (PIT) residual calibration, and autoregressive generation fidelity against persistence, surrogate, and real–real floor references, with single-head ablations isolating each head's contribution.

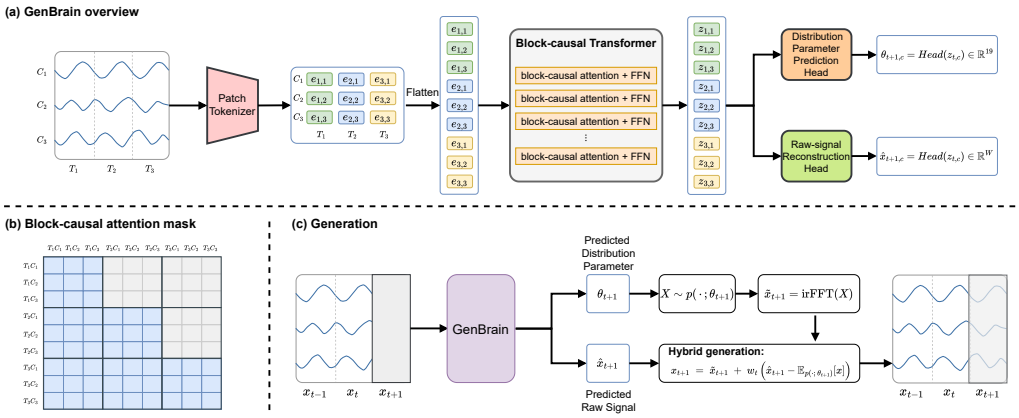

Figure 2: Overview of GenBrain. **(a)** A multichannel EEG recording is partitioned into time steps and tokenized into channel-specific tokens arranged in temporal blocks. A block-causal Transformer processes the sequence; at every position, a distribution-prediction head outputs the parameters of a neuro-mechanistically parameterized distribution for the next time step, while a zero-initialized reconstruction head regresses the next-step waveform. Training minimizes the negative log-likelihood of real EEG under the predicted distribution, jointly with a raw-signal reconstruction loss, regularized by soft neurophysiological priors. **(b)** The block-causal attention mask enforces strict causality across time steps while allowing full attention within each block. **(c)** At inference, hybrid generation blends a sample from the predicted distribution with the reconstruction head's transient residual to produce the next segment, which is rolled back into the context autoregressively.

2 METHOD

GenBrain is a generative autoregressive model of EEG. Instead of reconstructing masked signal fragments, it predicts the parameters of a neuro-mechanistically parameterized probability distribution that the next time step of EEG is assumed to follow, and generates brain activity by sampling from this distribution. Section 2.1 formalizes the problem, Section 2.2 defines the distribution family, Section 2.3 presents the block-causal architecture, Section 2.4 derives the pretraining objective that couples distribution prediction, raw-signal reconstruction, and neurophysiological constraints, and Section 2.5 describes generation and downstream adaptation. The full pipeline is illustrated in Figure 2.

2.1 PROBLEM FORMULATION

Given a multichannel EEG recording $X \in \mathbb{R}^{C\times T}$ with C channels and T samples, we partition it along the temporal axis into N non-overlapping time steps of length W (we use $W = 1\,\text{s}$):

$$X = [x_1, x_2, \ldots, x_N], \qquad x_t \in \mathbb{R}^{C\times W}. \tag{2}$$

The channel count C may vary across recordings; the formulation below is invariant to this.

GenBrain is a block-causal autoregressive model. Conditioned on the history $x_{\le t}$, it predicts a parameter vector θ_{t+1} that specifies the probability distribution of the next time step:

$$\theta_{t+1} = f_\phi(x_{\le t}), \qquad x_{t+1} \sim p(\,\cdot\,;\,\theta_{t+1}), \tag{3}$$

where f_ϕ is the GenBrain network and $p(\,\cdot\,;\,\theta_{t+1})$ is a neuro-mechanistically parameterized EEG distribution (Section 2.2). Training maximizes the likelihood of the observed x_{t+1} under $p(\,\cdot\,;\,\theta_{t+1})$; at inference, sampling from $p(\,\cdot\,;\,\theta_{t+1})$ produces a realization of the next time step, which is appended to the context to extend generation autoregressively.

2.2 A NEURO-MECHANISTICALLY PARAMETERIZED EEG DISTRIBUTION FAMILY

The core of GenBrain is a family of EEG distributions whose parameters carry explicit neurophysiological meaning, admit a tractable likelihood, and support differentiable sampling. For a single channel and time step, we adopt the generative model of Equation 1: narrowband rhythmic oscillations riding on a $1/f$ background, with $\mathcal{B} = \{\delta, \theta, \alpha, \beta, \gamma\}$ indexing the standard frequency bands.

Frequency-domain parameterized distribution. Applying the real-valued FFT to x yields the spectrum $X = \mathrm{rFFT}(x)$, with each frequency bin written in polar form as $X_k = A_k\, e^{i\varphi_k}$. Under the narrowband random-oscillator approximation, bins within a band are treated as conditionally independent given the band-level parameters, and the amplitude and phase follow distributions that arise physically for a narrowband Gaussian process (Rice, 1945; Mardia & Jupp, 2000):

$$A_k \sim \mathrm{Rayleigh}(\sigma_b), \quad b = b(k), \tag{4}$$

$$\varphi_k \sim \mathrm{vonMises}(\mu_b, \kappa_b), \quad b = b(k). \tag{5}$$

The Rayleigh distribution is the envelope distribution of a narrowband Gaussian process (Rice, 1945); the von Mises distribution is the maximum-entropy distribution on the circle for a given mean direction and concentration (Mardia & Jupp, 2000). The noise floor is modeled by the $1/f^\alpha$ law (Miller et al., 2009), and we additionally allow low-frequency phase to modulate high-frequency amplitude via phase-amplitude coupling (PAC) (Canolty et al., 2006), with coupling strength $\rho_{b_1,b_2} \in [0, 1]$ for selected band pairs.

Collecting all parameters, a single time step and channel is governed by the parameter vector

$$\theta = \big\{(\sigma_b, \mu_b, \kappa_b)\big\}_{b\in\mathcal{B}} \cup \{\alpha\} \cup \{\rho_{b_1,b_2}\} \in \mathbb{R}^{19}, \tag{6}$$

comprising five bands times three rhythmic parameters ($5 \times 3 = 15$), one spectral slope, and three PAC coefficients (δ–θ, θ–γ, α–β); the θ and α in these pair names are band labels, distinct from the parameter vector θ and the spectral slope α. The distribution $p(x;\theta)$ is therefore not a Gaussian but a structured composite of Rayleigh (amplitude), von Mises (phase), and $1/f$ (noise) components, with each parameter carrying an explicit neurophysiological meaning (Table 8 in Appendix G). Each channel c carries its own $\theta^{(c)} \in \mathbb{R}^{19}$; inter-channel structure is introduced through the spatial consistency prior (Appendix H) and the 3D-coordinate channel encoding (Appendix D). The network f_ϕ that predicts θ from history is described in Section 2.3.

2.3 BLOCK-CAUSAL ARCHITECTURE

Why block-causal? A discriminative MAE-style model can flatten the $C \times N$ patch tokens into a single sequence, because bidirectional attention makes causality irrelevant. GenBrain, however, is autoregressive: predicting the distribution at step $t + 1$ must depend only on steps $\le t$. The naive alternative—flattening all tokens into a single sequence under a standard lower-triangular mask—both violates the simultaneity of channels recorded at the same step and misaligns the autoregressive target (Appendix E). Both issues force a redesign of the token layout and attention mask.

Time-first block-wise sequence and block-causal mask. The $C \times N$ tokens are arranged in time-first order, the C channel tokens of each time step forming a contiguous block (total length $L = C \cdot N$). We define the block-causal mask $M \in \{0, -\infty\}^{L \times L}$. For query token (t_i, c_i) and key token (t_j, c_j):

$$M_{i,j} = \begin{cases} 0 & \text{if } t_j \le t_i \quad \text{(past or same step: visible)}, \\ -\infty & \text{if } t_j > t_i \quad \text{(future step: masked)}. \end{cases} \tag{7}$$

Each time step forms a block: attention is *full within each block* and *strictly causal across blocks* (Table 7 in Appendix F for a C=3, N=3 visualization). Within-block full attention reflects the simultaneity of channels recorded at the same step—scalp potentials are passive superpositions of simultaneously active sources, with no causal ordering among them—and lets channel relationships such as volume-conduction similarity and functional connectivity be modeled inside the Transformer rather than compressed into the tokenizer; across-block strict causality preserves autoregressive semantics.

Block-causal Transformer decoder. GenBrain uses a Pre-LN Transformer decoder (Xiong et al., 2020; Dehghani et al., 2023) with multi-head self-attention, each layer applying the block-causal mask of Equation 7 in place of the per-token lower-triangular mask of GPT. Input tokens are produced by a format-agnostic patch tokenizer that normalizes sampling rates, shares weights across channels, and encodes each channel with its 3-D electrode coordinates, allowing arbitrary montages (Appendix D). Because the mask and the channel encoding are constructed from the per-batch (C, N) rather than stored as fixed parameters, the decoder accepts variable C without any change to its weights; a training-time random channel permutation and subsetting further reinforce invariance to channel order and count (Appendix D).

Distribution-prediction head. The decoder emits a hidden state $z_{t,c} \in \mathbb{R}^d$ at every token position (t, c). A shared prediction head maps each $z_{t,c}$ to the distribution parameters of the same channel at the next time step, $\theta_{t+1,c} = \mathrm{Head}(z_{t,c}) \in \mathbb{R}^{19}$—a shifted (next-step) prediction in the GPT sense. To enforce physiological ranges, the head applies biophysically motivated hard transforms to its raw outputs, ensuring $\sigma_b > 0$, $\mu_b \in [-\pi, \pi]$, $\kappa_b \ge 0$, $\alpha \in [1, 4]$, and $\rho_{b_1,b_2} \in [0, 1]$ (explicit forms in Appendix I).

Parallel raw-signal reconstruction head. Real EEG contains transient, out-of-family components such as bursts and spikes that the Rayleigh and von Mises components cannot express. We therefore attach a second head in parallel: a lightweight zero-initialized MLP that maps the same hidden state $z_{t,c}$ to a direct regression of the next-step waveform $\hat{x}_{t+1,c} \in \mathbb{R}^W$. The two heads are complementary: the distribution head captures the parameterizable rhythmic structure, while the reconstruction head captures the out-of-family transients (architecture details in Appendix I; the head is dropped when the pretrained checkpoint is loaded for downstream fine-tuning).

2.4 Training Objective

Negative log-likelihood. Given the ground-truth next time step x_{t+1}, we compute its rFFT to obtain observed amplitudes $\{A_k^{\mathrm{obs}}\}$ and phases $\{\varphi_k^{\mathrm{obs}}\}$. Two practical designs shape the likelihood. First, spurious frequency bins attenuated by upstream low-pass filtering in some corpora are excluded by an intra-band relative power mask, so that only informative bins enter the likelihood. Second, *PAC-modulated likelihood*: phase-amplitude coupling is not a sampling-time embellishment but enters the likelihood directly—the observed low-frequency phase $\psi_{b_{\mathrm{low}}}$ modulates the effective scale of the high-frequency band,

$$\sigma_b^{\mathrm{eff}} = \sigma_b \cdot \mathrm{clip}\big(1 + \rho_{b_{\mathrm{low}},b}\, \cos(\psi_{b_{\mathrm{low}}} - \mu_{b_{\mathrm{low}}}),\ 0.2,\ 1.8\big), \tag{8}$$

so that the physiological mechanism "high-frequency amplitude increases where low-frequency phase is locked" is explicitly constrained by the data and the couplings ρ_{b_1,b_2} are learned rather than postulated. The final NLL, summed over retained bins and normalized by the number of valid bins, is

$$\mathcal{L}_{\mathrm{NLL}} = -\frac{1}{|\mathcal{V}|} \sum_{k \in \mathcal{V}} \Big[\log \mathrm{Rayleigh}\Big(A_k^{\mathrm{obs}};\, \sigma_{b(k)}^{\mathrm{eff}}\Big) + \log \mathrm{vonMises}\big(\varphi_k^{\mathrm{obs}};\, \mu_{b(k)}, \kappa_{b(k)}\big) \Big] + \lambda_{1f}\, \mathcal{L}_{1/f\text{-fit}}, \tag{9}$$

where $\mathcal{V}$ is the retained-bin set, $b(k)$ is the band containing bin k, and $\mathcal{L}_{1/f\text{-fit}}$ is the log-log fitting error between the masked power spectral density and the $1/f^{\alpha}$ model. This is genuine maximum-likelihood training, not MSE regression; the bin-mask construction, the explicit Rayleigh and von Mises log-densities, and the numerically stable Bessel implementation are given in Appendix I.

Raw-signal reconstruction loss. The reconstruction head is trained with a complementary pair of losses,

$$\mathcal{L}_{\text{rec}} = w_{\text{time}}\,\text{Huber}_{\delta}(\hat{x}_{t+1},\, x_{t+1}) + w_{\text{spec}}\,\mathcal{L}_{\text{MR-STFT}}(\hat{x}_{t+1},\, x_{t+1})\,, \tag{10}$$

with $\delta = 0.1$. The time-domain Huber term is point-wise: sensitive to transient shape, latency, and polarity, with bounded gradients under outliers (Huber, 1992). The multi-resolution STFT term is phase-insensitive and matches time-frequency energy structure (Yamamoto et al., 2020). The two are complementary by construction: the absolute phase of the next second of EEG is intrinsically unpredictable, so a pure time-domain MSE would collapse to a blurry near-zero-mean prediction, and the MR-STFT term prevents precisely this mean regression (resolutions and spectral terms in Appendix I).

Neurophysiological prior constraints. Beyond the range-clamping hard transforms (Equation 20), we impose five differentiable soft regularizers encoding neurophysiological priors: a $1/f$ spectral-slope constraint (physiological range and cross-step smoothness of α), a band-power ratio constraint, a phase-synchrony bound on κ_b, a cross-frequency coupling bound on ρ_{b_1,b_2}, and a spatial consistency constraint penalizing divergence of θ between spatially adjacent channels under a volume-conduction distance kernel. All terms take hinge or Huber forms and vanish whenever the predicted parameters already satisfy physiological bounds, biasing the solution only where physiology would otherwise be violated (full derivations and ranges in Appendix H).

Dual-head joint objective. The total objective combines the two heads with the prior constraints,

$$\mathcal{L} = w_{\text{nll}}\,\mathcal{L}_{\text{NLL}} + w_{\text{rec}}\,\mathcal{L}_{\text{rec}} + \sum_{i=1}^{5} \lambda_i\,\mathcal{L}_{\text{constraint},i}, \tag{11}$$

with $w_{\text{nll}} \equiv 1$; by default w_{rec} ramps linearly from 0 over the first 20% of training steps, while an uncertainty-based alternative (Kendall et al., 2018) learns the task weights jointly (schedules detailed in Appendix I and compared in Section 3). Training proceeds in a single stage from random initialization, with all parameters optimized jointly throughout under GPT-style shifted teacher forcing (Williams & Zipser, 1989); optimization details are given in Appendix I, and the complete hyperparameter configuration is listed in Appendix J.

2.5 Generation and Downstream Adaptation

Generation by sampling. Once θ_{t+1} is predicted, the next time step is synthesized by differentiable reparameterized sampling from $p(\,\cdot\,;\theta_{t+1})$ in the frequency domain, followed by inverse rFFT (Appendix K); the sampled x_{t+1} is fed back into the model as a new block, extending generation autoregressively to arbitrary lengths. The per-step cost is dominated by a single inverse FFT per channel and is real-time feasible.

Hybrid generation with transient injection. For the dual-head model, each sampled segment can optionally be refined with the out-of-family transient that the reconstruction head predicts. At rollout step t, we form

$$x_{t+1} \;=\; \tilde{x}_{t+1} \;+\; w_t\,\big(\hat{x}_{t+1} - \mathbb{E}_{p(\cdot;\,\theta_{t+1})}[x]\big), \qquad w_t = w\,\gamma^{\,t}, \tag{12}$$

where $\tilde{x}_{t+1}$ is the sampled segment, $\hat{x}_{t+1}$ is the reconstruction head's prediction, and $\mathbb{E}_p[x]$ is the closed-form mean waveform of the predicted distribution; the residue $\hat{x}_{t+1} - \mathbb{E}_p[x]$ isolates the transient structure that the parameterized family cannot express. Because deterministic predictions degrade as feedback accumulates, the injection weight decays geometrically (w=0.5, γ=0.85, fixed across all datasets in our experiments), preserving the spectral gains of injection while preventing amplitude overshoot on long horizons; where the transient residue is uninformative, the scheme degrades gracefully toward pure sampling.

Downstream adaptation. The pretrained representation also transfers to discriminative decoding through end-to-end fine-tuning: the decoder hidden states are aggregated or pooled and fed to a classification head trained jointly with the backbone (Appendix N).

3 EXPERIMENTS

3.1 SETUP

Pretraining corpora. We pretrain GenBrain on 14 public EEG corpora totaling 4.28M one-second segments (≈32,625 hours, 16,393 subjects after preprocessing): TUEG (Obeid & Picone, 2016), Siena Scalp EEG (Detti et al., 2020), the PhysioNet/CinC 2018 sleep corpus (Ghassemi et al., 2018; Goldberger et al., 2000), a raw community EEG archive (Trujillo, 2020), and ten OpenNeuro BIDS datasets (OpenNeuro, 2025; Melcón et al., 2025; Zhang et al., 2025; xx, 2025; X, 2025; anonymous, 2025; Kinley & Becker, 2025; Kim et al., 2025b;a; Computational Neuroimaging and Neuroengineering Lab ar the University of Oklahoma, 2025; Ghaffari et al., 2025). All corpora pass through one minimal preprocessing pipeline that preserves the signals' native statistics while guaranteeing cross-corpus comparability (Appendix B for the full corpus statistics).

Evaluation suite. All evaluations are zero-shot: the pretrained model is applied, without any parameter updates or dataset-specific adaptation, to the test splits of 10 datasets held out from the pretraining corpora, spanning six paradigms—emotion recognition, motor imagery, sleep staging, seizure detection, clinical psychiatry, and cognitive workload—with channel counts from 6 to 128 and classification targets from 2 to 9 classes (Appendix B for the full suite and dataset citations). Every experiment below runs on the full set of test records of each dataset, with no sampling or subsetting. Beyond the generative evaluations, Appendix N compares full fine-tuning decoding of the same checkpoint against supervised and foundation-model baselines on these datasets.

Model variants. *GenBrain* denotes the dual-head model. Two single-head ablations are pretrained from scratch under identical data, budget, and schedule (10 epochs): *dist-only*, which keeps the distribution head alone, and *signal-only*, which keeps the reconstruction head alone. The dist-only variant rolls out by within-family sampling; the signal-only variant has no predictive distribution (its NLL, where reported, is induced by moment matching its predicted waveform) and rolls out by deterministic waveform feedback.

Table 1: One-step teacher-forced NLL on the test splits of the 10 evaluation datasets (lower is better). Bold marks the best NLL per dataset. Δ = strongest baseline $-$ model, reported for GenBrain and the signal-only ablation.

			Baselines			Δ = best baseline $-$ model	
Dataset	GenBrain	Signal-only	Marginal	Persist.	AR(3)	GenBrain	Signal-only
BCIC-IV-2a	**4.400**	7.393	7.637	7.575	7.955	+3.17	+0.18
CHB-MIT	**5.523**	10.893	19.487	10.952	31.018	+5.43	+0.06
FACED	**4.777**	8.040	9.161	7.873	8.362	+3.10	−0.17
ISRUC	**4.308**	11.325	303.664	11.344	16.415	+7.04	+0.02
MODMA	**3.971**	6.108	6.775	6.333	7.332	+2.36	+0.23
MentalArithmetic	**4.511**	9.473	9.999	9.861	10.004	+5.35	+0.39
PhysioNet-MI	**5.050**	7.727	8.525	7.582	10.721	+2.53	−0.14
SEED-V	**4.057**	5.759	6.325	5.974	7.400	+1.92	+0.22
SEED-VII	**4.365**	5.938	8.194	22.378	23.289	+3.83	+2.26
SHU-MI	**2.932**	4.928	10.398	4.662	5.149	+1.73	−0.27

3.2 ONE-STEP PREDICTION QUALITY

Protocol. We measure the teacher-forced negative log-likelihood (NLL) of the true next time step under each forecaster's predictive distribution, using one likelihood function for all forecasters; the log score is a strictly proper scoring rule, so the comparison rewards only calibrated sharpness (Gneiting & Raftery, 2007). The baselines are the standard trio of probabilistic forecasting, all closed-form estimators in the same parameter space as GenBrain: a *marginal* forecaster returning the per-component median of moment-fitted parameters (a constant carrying no temporal information), *persistence*, which moment-fits the current step and copies it forward as the forecast, and an *AR(3)* predictor regressing each parameter on its three predecessors, solved once by least squares (order auto-reduced on short-trial datasets; Appendix L). The statistical baselines are fitted in-domain on held-out records of each test split, disjoint from the scored records—so they see downstream data

that GenBrain, evaluated fully zero-shot, never does. Every forecaster is scored under the identical likelihood (Appendix L details the scoring rule, the reconstruction SNR, and the baseline protocol).

Table 1 reports the results. Despite being fully zero-shot, GenBrain attains lower NLL than the strongest in-domain baseline on all 10 datasets, with margins of $+1.7$ to $+7.0$ nats (32–62% relative improvement); the largest gap is on ISRUC, where the marginal baseline collapses to 303.7 nats—between-subject spectral variability makes unconditional forecasting infeasible and conditional distribution prediction essential. The signal-only ablation is informative: moment matching its predicted waveform induces an implicit distribution that beats the baselines on 7 of 10 datasets yet trails GenBrain by 1.6 to 7.0 nats. Signal regression thus acquires part of the distributional structure implicitly, but far from all of it. Conversely, the dist-only ablation matches GenBrain's NLL within 0.07 nats on every dataset, and the dual-head reconstruction reaches an SNR comparable to the dedicated signal-only checkpoint while substantially exceeding a persistence reconstruction: the two heads are complementary rather than competing, each attaining its specialty at no cost to the other.

Table 2: PIT calibration diagnostics on the test splits. KS: distance of the amplitude PIT against uniform (lower is better); $P(z{>}3)$: tail exceedance against the 1.11% Rayleigh reference; excess kurtosis of standardized amplitude residuals for the dist-only ablation and GenBrain.

			Excess kurtosis	
Dataset	KS ↓	$P(z{>}3)$ (%)	Dist-only	GenBrain
BCIC-IV-2a	0.222	8.64	+6.0	+5.9
CHB-MIT	0.186	10.37	+27.9	+27.6
FACED	0.104	4.54	+9.7	+9.6
ISRUC	0.210	9.36	+94.0	+91.4
MODMA	0.143	5.97	+64.7	+69.6
MentalArithmetic	0.214	8.43	+18.6	+25.9
PhysioNet-MI	0.109	7.09	+20.2	+19.4
SEED-V	0.066	3.81	+62.9	+77.2
SEED-VII	0.081	3.34	+22.5	+22.6
SHU-MI	0.235	1.72	+369.4	+117.8

3.3 Residual Calibration

Protocol. Probability-integral-transform (PIT) residuals (Rosenblatt, 1952) diagnose calibration on held-out data: if the predicted distributions were correct, the PIT of amplitudes and the standardized amplitude residuals z would follow uniform and standard-normal references. We summarize the KS distance of the amplitude PIT against uniform, the tail exceedance $P(z{>}3)$ against the 1.11% Rayleigh reference, and the excess kurtosis of z (Gaussian reference 0; the Rayleigh family itself sits at -0.27). Appendix L gives the exact construction of these diagnostics and their reference values.

Table 2 reports the diagnostics. Calibration is strongest on the high-density emotion and motor-imagery montages (KS 0.066–0.109) and weaker on the remaining datasets (0.14–0.24). The uniform diagnostic across all datasets is that the tails are too light: $P(z{>}3)$ exceeds the Rayleigh reference everywhere (1.7–10.4%), and the excess kurtosis of the standardized residuals is positive on all 10 datasets (up to $+369$ on SHU-MI)—real EEG contains transient bursts and spikes heavier than the parameterized family. This measurement both quantifies the family's expressiveness boundary and motivates the transient-injection pathway of Equation 12. The dual-head model rescues the most extreme tails—excess kurtosis on SHU-MI drops from $+369$ (dist-only) to $+118$ (-68%), though it remains above the dist-only ablation on a subset of datasets; the signal-only variant admits no PIT at all, since a deterministic point prediction defines no forecast distribution.

3.4 Autoregressive Generation Fidelity

Protocol. Each test sequence provides a context of a few seconds to one minute; the model generates the next M seconds autoregressively, and the generated segment is compared against the real future segment. We report the PSD correlation (psd_corr, higher is better) and the generated-to-real RMS ratio (rmsR; 1.0 is calibrated). Three references bracket the task: *persistence* repeats the last

Table 3: Autoregressive rollout fidelity on the test splits (context → rollout seconds; metrics computed against the real future segment). psd_corr: spectral-shape correlation (higher is better); rmsR: generated-to-real RMS ratio (1.0 calibrated; the real–real floor itself spans 1.05–1.53, so rmsR is read against the floor). GB = GenBrain by pure distribution sampling; Hybrid = transient injection (Equation 12) with geometrically decaying weight; Signal-only rolls out by deterministic waveform feedback. Persistence copies the context waveform (not generative); Surrogate keeps the context spectrum with random phase.

		GB		Hybrid (geo)		Signal-only		Persist.	Surrog.	Floor	
Dataset	ctx→M	psd	rmsR	psd	rmsR	psd	rmsR	psd	psd	psd	rmsR
BCIC-IV-2a	2→2	0.835	0.904	0.863	0.972	0.875	0.630	0.972	0.962	0.963	1.066
CHB-MIT	7→3	0.679	0.654	0.732	0.728	0.850	0.428	0.976	0.967	0.915	1.295
FACED	7→3	0.798	0.965	0.804	1.048	0.910	0.552	0.990	0.984	0.980	1.159
ISRUC	20→10	0.721	1.453	0.699	1.874	0.921	0.428	0.986	0.966	0.942	1.414
MODMA	10→5	0.863	0.801	0.887	0.911	0.848	0.453	0.981	0.968	0.951	1.101
MentalArithmetic	3→2	0.767	0.865	0.787	0.925	0.927	0.604	0.990	0.986	0.981	1.052
PhysioNet-MI	2→2	0.884	0.785	0.910	0.842	0.866	0.497	0.977	0.970	0.919	1.455
SEED-V	20→10	0.854	0.971	0.888	1.213	0.833	0.395	0.981	0.947	0.919	1.298
SEED-VII	3→2	0.913	0.859	0.940	0.917	0.838	0.560	0.959	0.948	0.927	1.378
SHU-MI	2→2	0.859	1.865	0.871	1.951	0.823	0.711	0.928	0.881	0.898	1.525

context patches (a waveform copy, not a generation); *surrogate* keeps the context amplitude spectrum with random phase (right spectrum, no dynamics); the *real–real floor* scores the future segment of a different record—perfect but unrelated EEG—and is the level at which a generated segment is statistically indistinguishable from real data. Note that the floor's own rmsR is 1.05–1.53, not 1.0: cross-record amplitudes are inherently mismatched, so rmsR must be read against the floor rather than against unity. Appendix L details the rollout metrics and reference systems.

Table 3 reports the results. Pure distribution sampling (GB) tracks the real future spectrum at psd_corr 0.68–0.91, approaching the surrogate reference on most datasets. The signal-only variant looks strong on psd_corr but collapses in amplitude: its rmsR falls to 0.40–0.71 as deterministic feedback errors accumulate and the waveform smooths toward zero—regression alone cannot serve as a generator. Hybrid generation with geometrically decaying injection (Equation 12) improves psd_corr on 9 of 10 datasets over pure sampling ($+0.006$ to $+0.053$) while keeping amplitude calibration at or below the real–real floor on 8 of 10 datasets: on SEED-V (the longest 10-second rollouts) the injected rmsR of 1.213 sits below the floor's 1.298. A constant injection weight, by contrast, overshoots amplitude to 2.45 times the real signal on ISRUC; the geometric decay repairs this to 1.87 while retaining 59–100% of the constant-weight spectral gain on the 9 datasets where injection helps (Appendix M). ISRUC is the one dataset where injection is uninformative and pure sampling is the better mode (psd_corr 0.721 vs. 0.699)—the framework degrades gracefully rather than catastrophically, and we report this exception as is.

4 Conclusion

We have presented GenBrain, a generative autoregressive model of EEG that reframes the modeling of brain activity: instead of reconstructing masked segments, it predicts neuro-mechanistically parameterized distributions of the next time step and generates by sampling. Rayleigh amplitudes, von Mises phases, a $1/f$ background, and phase-amplitude coupling make every predicted parameter interpretable; the block-causal architecture preserves strict temporal causality while modeling cross-channel structure; and neurophysiological priors keep predictions and generations physiologically plausible. A zero-initialized reconstruction head captures the out-of-family transients at no cost to distributional quality, and the same residue lifts autoregressive generation through decaying transient injection. Evaluated zero-shot on 10 held-out datasets, GenBrain improves one-step NLL over all baselines on every dataset, and its generation tracks the real future spectrum with amplitude calibration at or below the real–real floor on 8 of 10. Heavy-tailed families, conditional generation, and invasive and real-time closed-loop settings define future work (Appendix O).

AI USE STATEMENT

In this work, we used generative AI tools for none of the research tasks with required disclosure. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, interpreting results, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, or supporting qualitative and thematic data analysis. We used generative AI tools for writing-related tasks only: drafting parts of the paper, editing the text to improve readability, suggesting the paper structure, proposing the title, and formatting references. We have reviewed all AI-assisted work. All AI-assisted text was carefully reviewed and revised by the authors, and all research content—including the model, the experiments, and every number reported—was conducted and produced solely by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

anonymous. LIGHT hypnotherapy, 2025.

Bruno Aristimunha, Raphael Yokoingawa de Camargo, Sylvain Chevallier, Adam G Thomas, Oeslle Lucena, Jorge Cardoso, Walter Hugo Lopez Pinaya, and Jessica Dafflon. Synthetic sleep EEG signal generation using latent diffusion models. In *DGM4H 2023-1st Workshop on Deep Generative Models for Health at NeurIPS 2023*, 2023.

Michael Breakspear. Dynamic models of large-scale brain activity. *Nature neuroscience*, 20(3): 340–352, 2017.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. BCI competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16:1–6, 2008.

György Buzsáki. *Rhythms of the Brain*. Oxford university press, 2006.

György Buzsáki. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, 68(3):362–385, 2010.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, et al. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1):178, 2022.

Ryan T. Canolty and Robert T. Knight. The functional role of cross-frequency coupling. *Trends in Cognitive Sciences*, 14(11):506–515, 2010.

Ryan T Canolty, Erik Edwards, Sarang S Dalal, Maryam Soltani, Srikantan S Nagarajan, Heidi E Kirsch, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, 313(5793):1626–1628, 2006.

Alexander J Casson, David C Yates, Shelagh JM Smith, John S Duncan, and Esther Rodriguez-Villegas. Wearable electroencephalography. *IEEE engineering in medicine and biology magazine*, 29(3):44–56, 2010.

Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, 12(9):513–525, 2016.

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.

Computational Neuroimaging and Neuroengineering Lab ar the University of Oklahoma. Resting EEG, 2025.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-GPT: Towards a foundation model for EEG. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 7480–7512, 2023.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. Domain-specific denoising diffusion probabilistic models for brain dynamics. *arXiv preprint arXiv:2305.04200*, 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. *arXiv preprint arXiv:2510.21585*, 2025.

Seif Eldawlatly. On the role of generative artificial intelligence in the development of brain–computer interfaces. *BMC Biomedical Engineering*, 6:4, 2024.

Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. NeurIPT: Foundation model for neural interfaces. *arXiv preprint arXiv:2510.16548*, 2025.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs. *arXiv preprint arXiv:2402.17772*, 2024.

Catalina M Galván, Rubén D Spies, Diego H Milone, and Victoria Peterson. Neurophysiologically meaningful motor imagery EEG simulation with applications to data augmentation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32, 2024.

Sanaz Ghaffari, Arian Yavari, Sara Bonyadian, Arsalan Ghofrani, and Russell Butler. Visual EEG study (BrainVision $\rightarrow$ BIDS), 2025.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the PhysioNet/Computing in Cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pp. 1–4. IEEE, 2018.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

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.

Roberta Grech, Tracey Cassar, Joseph Muscat, Kenneth P Camilleri, Simon G Fabri, Michalis Zervakis, Petros Xanthopoulos, Vangelis Sakkalis, and Bart Vanrumste. Review on solving the inverse problem in EEG source analysis. *Journal of neuroengineering and rehabilitation*, 5(1):25, 2008.

Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. *arXiv preprint arXiv:1806.01875*, 2018.

Biyu J. He, John M. Zempel, Abraham Z. Snyder, and Marcus E. Raichle. The temporal structures and functional significance of scale-free brain activity. *Neuron*, 66(3):353–369, 2010.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. SEED-VII: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 2024a.

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

Netiwit Kaongoen, Jaehoon Choi, Jin Woo Choi, Haram Kwon, Chaeeun Hwang, Guebin Hwang, Byung Hyung Kim, and Sungho Jo. The future of wearable EEG: A review of ear-EEG technology and its applications. *Journal of neural engineering*, 20(5):051002, 2023.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2018.

Aigerim Keutayeva and Berdakh Abibullaev. Data constraints and performance optimization for transformer-based models in EEG-based brain–computer interfaces: a survey. *IEEE Access*, 12, 2024.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.

Andy Jeesu Kim, Mara Mather, Santiago Morales, and Joshua Senior. Young adult resting state and auditory oddball task EEG data, 2025a.

Andy Jeesu Kim, Santiago Morales, Joshua Senior, and Mara Mather. HeartBEAM: Older adult resting state and auditory oddball task EEG data, 2025b.

Isaac Kinley and Sue Becker. Cueing the future to reduce temporal discounting, 2025.

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

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Mikhail A Lebedev and Miguel A L Nicolelis. Brain–machine interfaces: from basic science to neuroprostheses and neurorehabilitation. *Physiological Reviews*, 97(2):767–837, 2017.

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*, 14(2):715–729, 2022.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10-year update. *Journal of Neural Engineering*, 15(3):031005, 2018.

Yun Luo, Li-Zhen Zhu, Zi-Yu Wan, and Bao-Liang Lu. Data augmentation for enhancing EEG-based emotion recognition with deep generative models. *Journal of Neural Engineering*, 17(5): 056021, 2020.

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. *arXiv preprint arXiv:2506.09110*, 2025.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, 2000.

María Melcón, Enrique Stern, Lydia Arana, and Almudena Capilla. EEG data during three near-threshold visual detection tasks: a no-cue task, a noninformative cue task (50% validity), and an informative cue task (100% validity), 2025.

Kai J. Miller, Larry B. Sorensen, Jeffrey G. Ojemann, and Michel den Nijs. Power-law scaling in the brain surface electric potential. *PLoS Computational Biology*, 5(12):e1000609, 2009.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

Paul L Nunez and Ramesh Srinivasan. *Electric fields of the brain: the neurophysics of EEG*. Oxford university press, 2006.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in neuroscience*, 10:196, 2016.

OpenNeuro. OpenNeuro: A free and open platform for validating and sharing BIDS-compliant MRI, PET, MEG, EEG, and iEEG data, 2025. URL `https://openneuro.org`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

Stephen O. Rice. Mathematical analysis of random noise. *Bell System Technical Journal*, 24(1): 46–156, 1945.

Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

Simanto Saha and Mathias Baumert. Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, 13:87, 2020.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2496–2511, 2023.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Kai Shu, Le Wu, Yuchang Zhao, Aiping Liu, Ruobing Qian, and Xun Chen. Data augmentation for seizure prediction with generative diffusion model. *IEEE Transactions on Cognitive and Developmental Systems*, 2024.

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.

Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Chapman and Hall/CRC, 2024.

Chenxi Tian, Yuliang Ma, Jared Cammon, Feng Fang, Yingchun Zhang, and Ming Meng. Dual-encoder VAE-GAN with spatiotemporal features for emotional EEG data augmentation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2018–2027, 2023.

Giulio Tosato, Cesare M Dalbagno, and Francesco Fumagalli. EEG synthetic data generation using probabilistic diffusion models. *arXiv preprint arXiv:2303.06068*, 2023.

Logan Trujillo. Raw EEG Data, 2020. URL https://doi.org/10.18738/T8/SS2NHB.

Bradley Voytek, Mark A. Kramer, John Case, Kyle Q. Lepage, Zechari R. Tempesta, Robert T. Knight, and Adam Gazzaley. Age-related changes in 1/f neural electrophysiological noise. *Journal of Neuroscience*, 35(38):13257–13265, 2015.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, 2022.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=lvS2b8CjG5.

Jiquan Wang, Sha Zhao, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Generalizable sleep staging via multi-level domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 265–273, 2024b.

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

Jiquan Wang, Sha Zhao, Yangxuan Zhou, Yiming Kang, Shijian Li, and Gang Pan. DeeperBrain: A neuro-grounded EEG foundation model towards universal BCI. *arXiv preprint arXiv:2601.06134*, 2026.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.

Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

X. Memory reactivation levels remain unaffected by anticipated interference experiment 2 dataset, 2025.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. *arXiv preprint arXiv:2505.18185*, 2025.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10524–10533, 2020.

xx. Memory reactivation levels remain unaffected by anticipated interference, 2025.

Ryuichi Yamamoto, Eunwoo Song, and Hirokazu Kameoka. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, 2020.

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

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems*, 2023.

Zihan Zhang, Yu Bao, Tianyi Jiang, Xiao Ding, Xia Liang, Juntong Du, Yi Zhao, Kai Xiong, Bing Qin, and Ting Liu. Chisco-2.0, 2025.

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

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

EEG foundation models. Prior EEG foundation models adopt discriminative pretraining objectives: masked reconstruction (Chien et al., 2022; Foumani et al., 2024; Jiang et al., 2024b; Wang et al., 2025; 2024a; 2022; Zhang et al., 2023), contrastive learning (Kostas et al., 2021), or prediction of signal statistics (Yang et al., 2023; Wang et al., 2026). Their common limitation is that the objective is point-based or discriminative: the model reconstructs a masked signal, aligns positive pairs, or predicts a scalar statistic, but it neither models the probability distribution of the signal nor generates it autoregressively. Recent extensions along axes such as mixture-of-experts (Fang et al., 2025), unified EEG/MEG pretraining (Xiao et al., 2025), and discrete tokenization (Ma et al., 2025) remain within this discriminative paradigm. GenBrain departs from this line by explicitly defining and predicting the distribution of the next time step, so generation becomes a native capability of the pretraining objective rather than a downstream add-on.

Autoregressive generative models. Decoder-only autoregressive models predict the next-token distribution for both generation and representation (Achiam et al., 2023; Min et al., 2023), and the same principle has been extended to continuous modalities such as video (Brooks et al., 2024). Neuro-GPT (Cui et al., 2024) applies next-step prediction directly to EEG but targets the waveform value rather than an explicit continuous distribution, inheriting the noise-chasing problem of pointwise regression. GenBrain adopts the next-step distribution-prediction paradigm but applies it to a continuous signal governed by neurophysiological constraints, using a parameterized continuous distribution rather than discrete codebook tokens or raw samples as the prediction target.

EEG generative modeling. Existing EEG generation relies on diffusion models or GANs (Hartmann et al., 2018; Aristimunha et al., 2023; Tosato et al., 2023; Duan et al., 2023), which operate directly on waveforms in a purely data-driven manner, lack neurophysiological constraints, and are trained for specific tasks rather than as foundation models. Generative augmentation has been explored for emotion and seizure settings (Luo et al., 2020; Tian et al., 2023; Shu et al., 2024), and a recent line injects neurophysiological structure into motor-imagery EEG synthesis (Galván et al., 2024), but these remain task-specific generators rather than pretrained foundation models with a representation path. The case for generative AI in BCI development has been articulated at the application level (Eldawlatly, 2024); GenBrain fills this gap by making distribution prediction the pretraining objective itself.

Neurophysiology-inspired EEG modeling. Computational neuroscience models EEG as narrowband oscillations on a $1/f$ background (Buzsáki, 2006; 2010). The envelope of a narrowband Gaussian process follows a Rayleigh distribution (Rice, 1945); the phase of a coherent oscillation follows a von Mises distribution (Mardia & Jupp, 2000); the background spectrum decays as $1/f^{\alpha}$ with α reflecting cortical excitation-inhibition balance (He et al., 2010; Miller et al., 2009; Voytek et al., 2015); and phase-amplitude coupling is a core mechanism of neural coordination (Canolty et al., 2006; Canolty & Knight, 2010). These properties are typically used as post-hoc descriptors or handcrafted features. GenBrain instead builds them into the model itself: the Rayleigh, von Mises, and $1/f$ components define the parameterization of the next-step distribution that the model predicts, so neurophysiological priors shape the generative process directly.

B Datasets and Preprocessing Details

GenBrain is pretrained on 14 public EEG corpora and evaluated on 10 held-out datasets; all 24 pass through the unified preprocessing pipeline of Appendix C, resampled to a common 200 Hz so that one time step spans exactly one second ($W = 200$) for every source. Mastoid-referenced bipolar recordings without a full scalp montage (PhysioNet 2018, CHB-MIT) retain their native reference and skip CAR/REST; FACED ships pre-lowpassed by its authors, so the band-pass is applied only for uniformity and out-of-band bins are automatically masked at scoring time.

Pretraining corpora. Table 4 lists the 14 corpora. The largest, TUEG, is a large-scale clinical EEG archive; we use its full release across three montages (01_tcp_ar, 02_tcp_le, 03_tcp_ar_a) unified to a common 19-channel set. PhysioNet 2018 provides whole-night sleep EEG from 993 subjects (the challenge training set), annotated for sleep staging. The Raw EEG archive contains cognitive-task recordings (information integration and rule-based categorization). Siena collects scalp EEG from epilepsy patients, originally recorded with 31 channels and restricted here to the standard 10–20

Table 4: Pretraining corpora after unified preprocessing (all resampled to 200 Hz; dataset citations in Section 3.1). Subject counts, segments, and hours are measured on the processed LMDB stores (subjects deduplicated from key prefixes). TUEG contributes its full release across three montages unified to a 19-channel set; ds006317 provides dense within-subject recording (8 sessions); Raw archive covers 48 subjects in two task conditions.

Dataset	Subj.	Rate	C	Seg. (s)	Segments	Hours
TUEG	14,955	256	19	30	2,955,340	24,627.8
PhysioNet 2018	993	200	6	30	872,089	7,267.4
Siena	14	256	27	10	50,749	141.0
Raw archive	48	256	64	5	30,978	43.0
ds006171	36	1024	128	5	29,397	40.8
ds006317	2	1000	109	5	44,708	62.1
ds006367	52	1000	28	8	29,929	66.5
ds006370	56	1000	28	5	67,953	94.4
ds006437	9	256	64	5	12,069	16.8
ds006446	29	2048	64	5	12,970	18.0
ds006466	66	1000	64	5	94,392	131.1
ds006480	68	1000	64	5	51,572	71.6
ds006525	34	250	128	5	4,080	5.7
ds006547	31	500	63	5	28,222	39.2
Total	**16,393**				**4,284,448**	**32,625.4**

subset of 27 channels. The ten OpenNeuro BIDS datasets span contemporary cognitive and clinical studies: near-threshold visual detection with predictive cues (Melcón et al., 2025), imagined-speech BCI (Chisco-2.0) (Zhang et al., 2025), memory-reactivation experiments (xx, 2025; X, 2025), hypnotherapy (anonymous, 2025), cueing to reduce temporal discounting (Kinley & Becker, 2025), resting-state and auditory-oddball recordings from younger and older adults (HeartBEAM) (Kim et al., 2025a;b), resting-state EEG (Computational Neuroimaging and Neuroengineering Lab ar the University of Oklahoma, 2025), and a visual study (Ghaffari et al., 2025); they vary widely in channel count (28–128) and sampling rate (250–2048 Hz), reflecting modern high-density recording practices. After preprocessing, the corpora total 4,284,448 segments (≈32,625.4 hours) from 16,393 subjects, spanning 6–128 channels and 200–2048 Hz native sampling rates. Electrode layouts mix four systems—ten corpora use standard 10–05 positions (or 10–20 subsets, which 10–05 subsumes), two use GSN-HydroCel-65, one GSN-HydroCel-128, and PhysioNet 2018 a mastoid-referenced bipolar set—variation in electrode count, sampling rate, and montage that the format-agnostic tokenizer (Appendix D) absorbs into a single pretrained model.

Evaluation datasets. Table 5 details the 10 held-out datasets. The paradigms cover emotion recognition from film clips, imagined limb movements, whole-night sleep scoring, seizure detection in long-term scalp EEG from pediatric epilepsy patients, clinical depression diagnosis (MODMA), and cognitive workload estimation under sustained mental arithmetic (MentalArithmetic). Splits are subject-level (disjoint subject groups across train/val/test) for seven datasets; SEED-V and SEED-VII split at the trial level (all subjects appear in every split, within sessions and subjects respectively); MODDA splits by diagnosis group (train 15 MDD + 18 NC subjects, val 4 + 5, test 5 + 6). ISRUC stores each record as 20 stacked 30 s epochs and is evaluated as 30 s sequences; SEED-V's 1 s windows are grouped into recording blocks (48 sessions from 16 subjects $\times$ 5 test trials $=$ 240 blocks, cut to sequences of at most 30 s).

Zero-shot evaluation usage. All results use only the test splits of the evaluation datasets. The exp1 statistical baselines are fitted on the first half of each dataset's test records and all forecasters are scored on the second half (Section 3.2); exp2 and exp3 run on the full test sequences. Rollout geometries (context $\rightarrow$ generated seconds) per dataset: BCIC-IV-2a 2$\rightarrow$2, PhysioNet-MI 2$\rightarrow$2, SHU-MI 2$\rightarrow$2, MentalArithmetic 3$\rightarrow$2, SEED-VII 3$\rightarrow$2, CHB-MIT 7$\rightarrow$3, FACED 7$\rightarrow$3, MODMA 10$\rightarrow$5, SEED-V 20$\rightarrow$10, ISRUC 20$\rightarrow$10.

Table 5: Evaluation datasets after unified preprocessing (200 Hz). Counts are measured on the processed stores. Class counts per paradigm are listed in the group headers in dataset order. Len.: sample duration (ISRUC stacks 20 epochs of 30 s per record). Subj.: number of subjects (SEED-V: 48 sessions from 16 subjects; MODDA: 24 MDD + 29 normal controls).

Dataset	C	Subj.	Len. (s)	Samples	Split
Emotion recognition — 9/5/7 classes					
FACED (Chen et al., 2023)	32	123	10	10,332	S1–80/81–100/101–123
SEED-V (Liu et al., 2022)	62	48	1	117,744	trials 5:5:5/session
SEED-VII (Jiang et al., 2024a)	62	20	5	55,561	trials 10:5:5/subject
Motor imagery — 4/4/2 classes					
PhysioNet-MI (Schalk et al., 2004)	64	109	4	9,837	S1–70/71–89/90–109
BCIC-IV-2a (Brunner et al., 2008)	22	9	4	5,184	S1–5/6–7/8–9
SHU-MI (Ma et al., 2022)	32	25	4	11,988	S1–15/16–20/21–25
Sleep staging — 5 classes					
ISRUC (Khalighi et al., 2016)	6	100	30×20	4,462	S1–80/81–90/91–100
Seizure detection — 2 classes					
CHB-MIT (Shoeb, 2009)	16	23	10	326,991	S1–19/20–21/22–23
Clinical psychiatry — 2 classes					
MODDA (Cai et al., 2022)	128	53	15	1,066	diagnosis groups
Cognitive workload — 2 classes					
MentalArithmetic (Zyma et al., 2019)	20	36	5	1,707	S1–28/29–32/33–36

C Feasibility and Implementation Analysis

This appendix justifies the feasibility of GenBrain from two angles: the data principles behind the distribution family, and the engineering implementability of every component.

C.1 Data-Principle Feasibility

Theoretical basis of the generative model. The "oscillation + $1/f$ noise" account of EEG is a classical result of computational neuroscience: narrowband oscillatory envelopes follow a Rayleigh distribution as the envelope of a narrowband Gaussian process (Rice, 1945), phases in coherent regimes follow a von Mises distribution (Mardia & Jupp, 2000), and the $1/f^{\alpha}$ background is widely validated (Buzsáki, 2006; Miller et al., 2009). The distribution family is built on physically grounded distributions with theorem-level support, not on empirical curve fitting.

Autoregressive predictability premise. Autoregressive modeling presumes temporal autocorrelation of the predicted parameters $\{\sigma_b, \kappa_b, \alpha, \rho_{b_1,b_2}\}$. The neurophysiological timescales of these parameters supply it: band power (σ_b) fluctuates on second-to-minute scales (high predictability), phase synchrony (κ_b) varies slowly (medium-to-high), the aperiodic slope (α) is comparatively stable (high), and coupling (ρ_{b_1,b_2}) is stimulus-driven (medium). Autoregressive prediction is justified as long as the parameter autocorrelation remains significantly non-zero within 1–5 step lags.

Frequency-bin sufficiency. A $W = 1$ s step at $f_s = 200$ Hz yields 1 Hz resolution and per-band bin counts of approximately $\delta \approx 4$, $\theta \approx 4$, $\alpha \approx 5$, $\beta \approx 17$, and $\gamma \approx 15$. Every band retains at least four bins, which suffices to define and evaluate the band-level Rayleigh and von Mises likelihoods; low-frequency bands can be strengthened by lengthening the step to 2 s or by band-aggregated likelihoods.

Non-triviality of the family. The family is not reducible to a Gaussian: Rayleigh amplitudes are one-sided and asymmetric, von Mises phases are circular, and $1/f$ noise is colored. Their composition cannot be approximated by any single Gaussian, which is precisely why the parameterization is necessary.

C.2 Engineering Implementability

All core computations are natively differentiable in modern tensor frameworks: the Rayleigh log-pdf is elementwise arithmetic; the von Mises log-pdf uses a scaled Bessel implementation; the $1/f$ fit is a differentiable least-squares regression on the log-log spectrum; rFFT/irFFT, band grouping, and attention masks are standard primitives; and Rayleigh sampling admits a closed-form inverse-CDF reparameterization.

Unified preprocessing pipeline. All 14 pretraining corpora and 10 evaluation datasets pass through one minimal pipeline that preserves the signals' native statistics while guaranteeing comparability: channel selection $\rightarrow$ CAR $\rightarrow$ resampling to 200 Hz $\rightarrow$ 0.5–75 Hz bandpass $\rightarrow$ adaptive power-line notch $\rightarrow$ REST re-referencing $\rightarrow$ $\pm 500\,\mu$V clip $\rightarrow$ dead-channel repair $\rightarrow$ float32/μV $\rightarrow$ LMDB. CAR and REST unify the reference semantics across datasets; the 0.5–75 Hz passband covers the δ–γ range while suppressing line drift and high-frequency muscle artifacts; the notch frequency is auto-detected (50/60 Hz and harmonics), and the spectrum-fit sinusoid subtraction is iterated until the line-frequency residual falls below ten times the local spectral flank, with a ± 1 Hz narrowband FIR fallback when drifting line noise stalls the fit (on MODMA, whose $\sim$200 μV drifting 50 Hz contamination survives CAR, this drives the residual line peak from 2839 to 0.07 times the flank floor); trials shorter than the FIR length switch to a zero-phase IIR filter to avoid edge artifacts, and trials are never concatenated, since concatenation would introduce artificial continuity; the $\pm 500\,\mu$V clip removes only extreme artifacts. We avoid aggressive artifact rejection (ASR, ICA): such dataset-dependent processing harms the distributional consistency required for cross-corpus pretraining, and a model trained at scale should learn the statistics of both physiology and artifacts from the data itself.

Numerical stability. The Bessel term $I_0(\kappa)$ overflows float32 at moderate κ. We therefore evaluate $\log I_0(\kappa)$ through the exponentially scaled variant, $\log I_0(\kappa) = \log(\mathrm{i0e}(\kappa)) + \kappa$, and switch to the asymptotic expansion $\log I_0(\kappa) \approx \kappa - \frac{1}{2}\log(2\pi\kappa)$ for $\kappa > 500$. The Rayleigh $\log A$ is ϵ-smoothed as $A \rightarrow 0$, and the $1/f$ fit is protected against zero-power bins. Amplitudes are normalized by 10^{-2} and gradient norms are clipped; the zero-initialized reconstruction head starts as a no-op and its loss weight ramps in from zero (Appendix I), so training opens from the distribution-prediction objective alone.

Computational cost. The NLL involves only per-bin elementwise log-densities, $O(CW)$, the same order as one forward pass; sampling is dominated by a single inverse FFT per channel, $O(CW \log W)$, which is real-time feasible; no $O(W^2)$ pairwise computation is introduced.

C.3 Risks and Mitigations

- *Insufficient low-band bins* may destabilize distribution estimation: mitigate by lengthening steps to 2 s or using band-aggregated likelihoods.
- *Conditional-independence approximation*: retained as a first-order baseline; within-band correlation modeling is a natural extension.
- *Autoregressive error accumulation* in long-horizon generation: mitigated during training by teacher forcing and the prior constraints, and at inference by the geometrically decaying transient injection (Equation 12); each rollout is anchored on a real context segment (Section 3.4).
- *von Mises sampling has no exact native reparameterization*: use the normal approximation in the large-κ regime or Best–Fisher circular sampling for small κ.

D Format-Agnostic Patch Tokenizer

Design goal. Pretraining corpora arrive with heterogeneous channel counts, montages, and sampling rates; the tokenizer maps all of them into one shared token space so that a single pretrained model absorbs every source. It consists of a sampling-rate normalization performed upstream, a channel-shared convolutional patch embedding, and position encodings.

Sampling-rate normalization. All recordings are resampled to a common base rate $f_s = 200$ Hz in the unified preprocessing pipeline (Appendix C), so one time step of $W = 1$ s spans exactly $f_s \cdot$

$1\,\mathrm{s} = 200$ samples for every source. The tokenizer therefore always sees the same patch geometry, and the frequency-bin layout of the predicted distribution is identical across corpora.

Channel-shared convolutional patch embedding. Each token corresponds to one (channel, time-step) pair: a patch $x_{t,c} \in \mathbb{R}^W$ is embedded by a temporal-convolutional stack whose weights are shared by all $C \times N$ patches of a recording, so the parameter count is independent of C and any channel count is accepted without modification. The stack comprises (1) a strided convolution with 25 kernels of length 49 samples ($\approx 0.25\,\mathrm{s}$) and stride 25, reducing the 200-sample patch to 8 temporal positions (implicit boundary padding at the patch edges makes the position count exactly $200/25$), and (2) two convolutions with 25 kernels of length 3 and unit stride that refine local temporal structure. Every convolution is bias-free and followed by GELU and a per-token group normalization (five groups); the resulting 25×8 feature map is flattened into a token of dimension $d = 200$.

Per-token isolation. All kernels act along time within a patch and never mix information across the token axis, and the group normalization computes its statistics from each token's own features alone. A token's embedding is therefore a function of only that token's one-second signal—independent of C, N, and the other tokens present. This invariance is what makes KV-cached autoregressive decoding valid: a newly generated block embeds identically whether processed alone or inside the full sequence, so previously cached keys and values remain valid.

Position encodings. Each token is augmented with a spatial and a temporal position encoding; both follow DeeperBrain (Wang et al., 2026), and we refer to it for the exact constructions. In brief, the spatial encoding injects the channel's 3-D electrode coordinates smoothed by a learnable distance kernel, so any montage is identified by physical position rather than by channel index; the temporal encoding is a fixed-basis function of absolute time combining slow sinusoidal and exponential-decay components, keeping full-sequence training and incremental decoding on one shared time index.

Training-time channel reorganization. To reinforce the inductive bias that the channel set is unordered and identified by its coordinates rather than by position, each training batch undergoes a random channel permutation followed by a random channel-subset selection before the forward pass. Together with the coordinate-based spatial encoding, this prevents the model from relying on any fixed channel order or count and underwrites transfer to arbitrary montages.

Table 6: Candidate architectures for organizing $C \times N$ patch tokens in an autoregressive EEG model. GenBrain adopts Option B.

Option	Sequence organization	Attention pattern	Trade-off
A. Per-step aggregation	One token per step, length N	Standard causal	Simplest causally; loses per-channel structure
B. Block-causal (ours)	$C \times N$ tokens in time-first blocks	Within-block full attention, across-block strictly causal	Per-channel prediction, joint channel modeling, strict temporal causality
C. Factorized space-time	Two passes: intra-channel bidirectional, then cross-step causal	Two separated attentions	Efficient $O(NC^2 + CN^2)$; loses joint spatiotemporal modeling
D. Fully causal	Lexicographic lower-triangular	Full lower-triangular	Asymmetric ordering of simultaneous channels; not physical

E Candidate Architecture Comparison

We consider four candidate organizations of the $C \times N$ patch tokens, summarized in Table 6. Option A aggregates channels into a single token per step, which is simplest causally but discards per-channel structure. Option C factorizes space and time into two separate attention passes, gaining efficiency at the cost of joint spatiotemporal modeling. Option D uses a fully causal lower-triangular mask under a lexicographic ordering, which asymmetrically orders simultaneously recorded channels and is not physically meaningful. We adopt Option B, the block-causal mask.

Table 7: Block-causal attention mask for $C = 3$ channels and $N = 3$ time steps. Rows are queries, columns are keys. 1 indicates visible, 0 indicates masked.

	T_1C_1	T_1C_2	T_1C_3	T_2C_1	T_2C_2	T_2C_3	T_3C_1	T_3C_2	T_3C_3
T_1C_1	1	1	1	0	0	0	0	0	0
T_1C_2	1	1	1	0	0	0	0	0	0
T_1C_3	1	1	1	0	0	0	0	0	0
T_2C_1	1	1	1	1	1	1	0	0	0
T_2C_2	1	1	1	1	1	1	0	0	0
T_2C_3	1	1	1	1	1	1	0	0	0
T_3C_1	1	1	1	1	1	1	1	1	1
T_3C_2	1	1	1	1	1	1	1	1	1
T_3C_3	1	1	1	1	1	1	1	1	1

F Block-Causal Attention Mask Visualization

Table 7 illustrates the 9×9 block-causal mask for $C = 3$ channels and $N = 3$ time steps, where 1 denotes a visible entry and 0 a masked entry. Block t sees all blocks $1, \ldots, t$ entirely, and no block beyond t.

Because attention is full within each block, the result is invariant to the permutation of channels within a block, as long as the spatial encoding correctly identifies each channel.

Table 8: Physical and neurophysiological meaning of the distribution parameters θ.

Parameter	Distribution	Physical meaning	Neurophysiological mechanism
σ_b	Rayleigh scale	Oscillation strength in band b	Band-limited rhythm energy
μ_b	von Mises mean direction	Dominant phase of band b	Phase-locking direction
κ_b	von Mises concentration	Phase synchrony of band b	Neural synchrony (high κ = strong locking)
α	$1/f$ spectral slope	Background noise slope	Cortical excitation-inhibition balance ($\alpha \in [1, 4]$)
ρ_{b_1,b_2}	PAC coupling strength	Cross-frequency coupling	Low-frequency phase modulates high-frequency amplitude

G Physical Meaning of Distribution Parameters

Table 8 summarizes the physical and neurophysiological interpretation of each parameter in θ, with band subscripts b ranging over $\mathcal{B} = \{\delta, \theta, \alpha, \beta, \gamma\}$ (the θ and α in these band names are labels, distinct from the parameter vector θ and the spectral slope α). Empirically, every predicted parameter indeed carries its reading: the predicted $1/f$ slope tracks the held-out fitted slope at correlations of 0.68–0.73 on the two large emotion corpora (SEED-V, SEED-VII), and the parameters separate brain states in line with the literature (flatter slopes under clinical loads than in healthy emotion tasks).

H Neurophysiological Prior Constraints

The hard transforms applied by the distribution-prediction head (Equation 20) restrict individual parameters to valid ranges but do not enforce relations among parameters or across time and channels. We impose five differentiable soft regularization terms.

$1/f$ **spectral slope constraint.** Empirical studies bound α to a physiological range $[\alpha_{\min}, \alpha_{\max}]$ (we use $[1, 4]$) (Miller et al., 2009; Voytek et al., 2015), and rhythmic activity does not switch abruptly across consecutive time steps:

$$\mathcal{L}_{1/f} = \mathrm{Huber}(\alpha,\, [\alpha_{\min}, \alpha_{\max}]) + \lambda_\alpha \left\|\alpha_{t+1} - \alpha_t\right\|^2. \tag{13}$$

Band power ratio constraint. Because σ_b^2 is the mean power of band b, we constrain pairwise power ratios to lie within prespecified bounds $[r^{\min}_{b_1,b_2}, r^{\max}_{b_1,b_2}]$:

$$\mathcal{L}_{\mathrm{ratio}} = \sum_{(b_1,b_2)} \mathrm{ReLU}\left(\frac{\sigma^2_{b_1}}{\sigma^2_{b_2}} - r^{\max}_{b_1,b_2}\right) + \mathrm{ReLU}\left(r^{\min}_{b_1,b_2} - \frac{\sigma^2_{b_1}}{\sigma^2_{b_2}}\right). \tag{14}$$

Phase synchrony constraint. Excessively large κ_b corresponds to unphysiological over-locking of phase. We impose an upper bound $\kappa_{\max}$:

$$\mathcal{L}_{\mathrm{sync}} = \sum_{b\in\mathcal{B}} \mathrm{ReLU}(\kappa_b - \kappa_{\max})\,. \tag{15}$$

Cross-frequency coupling constraint. We penalize PAC strengths exceeding a physiological ceiling $\rho_{\max}$:

$$\mathcal{L}_{\mathrm{pac}} = \sum_{(b_1,b_2)} \mathrm{ReLU}(\rho_{b_1,b_2} - \rho_{\max})\,. \tag{16}$$

Spatial consistency constraint. Under volume conduction, nearby electrodes record similar potentials (Nunez & Srinivasan, 2006; Grech et al., 2008). We construct a distance kernel $K_{c,c'} = \exp(-d_{c,c'}/\ell)$ from 3D electrode coordinates and penalize parameter divergence between spatial neighbors:

$$\mathcal{L}_{\mathrm{spatial}} = \sum_{c} \sum_{c'\in\mathcal{N}(c)} K_{c,c'} \left\|\theta_{t,c} - \theta_{t,c'}\right\|^2. \tag{17}$$

All constraint terms use hinge (ReLU) or Huber forms, so they vanish at the likelihood optimum whenever the predicted parameters already satisfy physiological bounds, biasing the solution only where physiology would otherwise be violated.

I Training and Likelihood Details

Intra-band relative power mask. Upstream low-pass filtering in some corpora (e.g., a 45 Hz cutoff) attenuates stop-band bins by more than 30 dB, and counting these spurious bins would systematically distort σ_b. We therefore retain only bins whose power exceeds a threshold relative to the within-band maximum,

$$\mathcal{V}_b = \left\{k \in b \;:\; P_k > P_{b,\max}\,\tau_b\right\}, \tag{18}$$

with τ_b at -30 dB for the $\theta/\alpha/\beta/\gamma$ bands and -60 dB for δ, which is unaffected by low-pass filtering. The retained-bin set of Equation 9 is $\mathcal{V} = \bigcup_{b\in\mathcal{B}} \mathcal{V}_b$, and the NLL is normalized by the number of retained bins $|\mathcal{V}|$.

Explicit log-densities. The log-densities entering Equation 9 are

$$\log p(A;\,\sigma) = \log\left(\frac{A}{\sigma^2}\right) - \frac{A^2}{2\sigma^2}, \qquad \log p(\varphi;\,\mu,\kappa) = \kappa\cos(\varphi-\mu) - \log(2\pi I_0(\kappa))\,, \tag{19}$$

where I_0 is the modified Bessel function of the first kind, computed via a numerically stable exponentially scaled implementation with an asymptotic expansion for large κ (Appendix C).

Hard transforms of the distribution-prediction head. The distribution-prediction head maps its raw outputs to valid parameter ranges via

$$\begin{aligned} &\sigma_b = \mathrm{softplus}(\cdot) + \epsilon, \;\; \mu_b = \tanh(\cdot)\cdot\pi, \;\; \kappa_b = \mathrm{softplus}(\cdot), \\ &\alpha = 1 + 3\,\mathrm{sigmoid}(\cdot), \;\; \rho_{b_1,b_2} = \mathrm{sigmoid}(\cdot). \end{aligned} \tag{20}$$

The reconstruction head is a two-layer MLP (Linear $\rightarrow$ GELU $\rightarrow$ Linear, $\approx$0.12M parameters) operating on amplitude-normalized inputs.

MR-STFT loss details. The multi-resolution STFT term of Equation 10 uses FFTs of size 64/128/256 and combines spectral convergence with a log-magnitude L_1 distance at each resolution (Yamamoto et al., 2020).

Reconstruction-weight schedules. Three schedules for w_{rec} in Equation 11. *Fixed*: a constant λ_{rec}, used as a baseline. *Ramp* (default): w_{rec} rises linearly from 0 to λ_{rec} over the first 20% of training steps—the Rayleigh NLL contains a steep $A^2/2\sigma^2$ cliff, and early optimization benefits from stability while the learning rate warms up, so reconstruction pressure is introduced gradually. *Uncertainty* (Kendall et al., 2018): homoscedastic task uncertainties s_i are learned jointly, $\mathcal{L} = \sum_i \frac{1}{2s_i^2}\mathcal{L}_i + \log s_i$, automatically balancing losses whose relative magnitudes are hard to predict.

Optimization details. All parameters (backbone and both heads) are optimized jointly in a single stage from random initialization, under a linear warmup followed by cosine decay, whose low-learning-rate tail provides fine convergence without a separate fine-tuning phase; both objectives are active from step zero, so there is no "pollute-then-clean-up" staging. The training loop is the standard GPT-style shifted teacher-forcing procedure (Williams & Zipser, 1989): the model ingests $x_{\leq t}$, both heads emit predictions for step $t+1$ at every position, and the losses are accumulated over $t = 1, \ldots, N-1$. Amplitude normalization ($\times 10^{-2}$) and gradient-norm clipping ensure numerical stability.

J Pretraining Hyperparameters

Table 9 lists the complete pretraining configuration of GenBrain, matching the released code. The single-head ablations (*dist-only*, *signal-only*) reuse the identical architecture, data, budget, and schedule; only the active heads differ.

Table 9: Pretraining configuration of GenBrain.

Architecture		*Optimization*	
Step length W	1 s (200 samples)	Optimizer	AdamW
Context length	30 time steps	Peak LR	10^{-4}
Hidden dim.	200	Weight decay	5×10^{-2}
Attention heads	8	LR schedule	1-epoch warmup, cosine to 1%
Layers	12	Batch size	32
FFN dim.	800	Epochs	10
Dropout	0.1	Gradient clip	1.0
Recon. head	2-layer MLP (0.12M)	Precision	fp32, single GPU

Loss weights (Equation 11): $w_{\mathrm{nll}} = 1$; w_{rec}: ramp $0 \rightarrow 1$ over first 20% of steps; $\mathcal{L}_{\mathrm{rec}}$: Huber weight 1, MR-STFT weight 1; regularizers $\lambda_i = 0.1$ for all five constraints.

K Differentiable Sampling Procedure

Once θ_{t+1} is predicted, the next time step is synthesized by differentiable sampling from $p(\,\cdot\,; \theta_{t+1})$ in the frequency domain. For each frequency bin k belonging to band $b = b(k)$:

$$A_k = \sqrt{-2\sigma_b^2 \log u}, \qquad u \sim \mathcal{U}(0,1), \tag{21}$$

$$\varphi_k \sim \mathrm{vonMises}(\mu_b, \kappa_b). \tag{22}$$

Equation 21 is the inverse-CDF reparameterization of the Rayleigh distribution. The von Mises draw can be implemented by a reparameterized normal approximation in the large-κ regime or by Best–Fisher circular sampling for small κ (Appendix C). We then assemble the complex spectrum $X_k = A_k\, e^{i\varphi_k}$, scale the high-frequency amplitudes by the PAC modulation of Equation 8—that is, by the clipped factor defining σ_b^{eff}, evaluated at the band-level phase $\psi_{b_{\mathrm{low}}}$ of the modulating low band (sampled here at generation time; observed from the data in the training likelihood)—add the

$1/f^{\alpha}$ colored noise floor, and apply the inverse real FFT:

$$x_{t+1} = \mathrm{irFFT}(X). \tag{23}$$

The entire pipeline from θ_{t+1} to x_{t+1} is differentiable, supporting end-to-end training when the sampled signal is used downstream. At inference time, gradients are disabled and generation proceeds by feeding the sampled x_{t+1} back into the model as a new block.

L Evaluation Metrics and Protocols

This appendix details the generative evaluation protocol: the scoring rule behind the one-step experiment, the residual calibration diagnostics, the reconstruction SNR, and the rollout metrics with their reference systems. All metrics operate on the preprocessed 200 Hz signals ($W = 200$ samples per block), are computed per channel and per retained frequency bin, and are aggregated over the full test split of each dataset, with no sampling or subsetting; out-of-band bins are masked at scoring time (Appendix B).

L.1 One-Step Negative Log-Likelihood

The one-step experiment scores each forecaster's predictive distribution over the true next time step under teacher forcing—the context is always the real history—so the score isolates next-step distributional quality from compounding rollout error. The reported NLL is the evaluation-time counterpart of Equation 9 with all training losses removed except the $1/f$ residual, which is retained at unit weight for every forecaster: each scored (channel, time-step) signal contributes its per-bin negative log-likelihood normalized by its own retained-bin count, and the reported value averages these per-signal scores over the test split, in nats per bin.

The log score is strictly proper (Gneiting & Raftery, 2007): among all forecast distributions, only the true conditional distribution minimizes its expected value, so no forecaster can improve its score by hedging, or by sharpening without calibrating. Because every forecaster—GenBrain, its ablations, and the three statistical baselines—is scored under the identical likelihood of Equation 9, tabulated differences reflect distributional quality alone rather than scoring conventions. The baselines are closed-form estimators in the same parameter space, fitted in-domain on held-out records disjoint from the scored records (section 3.2); the AR baseline uses order 3, auto-reduced to $\min(3, N-2)$ on short fit sequences (realized as AR(2) on BCIC-IV-2a, PhysioNet-MI, and SHU-MI). The signal-only ablation, which has no explicit predictive distribution, is scored by moment matching—the Rayleigh and von Mises parameters that match the predicted waveform's amplitude and phase moments define an implicit distribution that is scored under the same likelihood.

L.2 Residual Calibration Diagnostics

Probability-integral-transform (PIT) residuals (Rosenblatt, 1952) diagnose calibration on held-out data without requiring a point-valued ground truth. Writing A_k^{obs} for the observed amplitude of retained bin k and $\hat{\sigma}_{b(k)}$ for the predicted Rayleigh scale of its band, the amplitude PIT is

$$u_k \;=\; F_{\mathrm{Ray}}\big(A_k^{\mathrm{obs}};\, \hat{\sigma}_{b(k)}\big) \;=\; 1 - \exp\!\big(-z_k^2/2\big)\,, \qquad z_k \;=\; \frac{A_k^{\mathrm{obs}}}{\hat{\sigma}_{b(k)}}, \tag{24}$$

so that z_k expresses the observed amplitude in units of the predicted Rayleigh scale. If the predicted distributions were correct, the u_k would be i.i.d. uniform on $[0, 1]$; any systematic departure localizes miscalibration.

KS distance. We summarize the amplitude PIT against uniform with the Kolmogorov–Smirnov statistic

$$D_n \;=\; \sup_{u\in[0,1]} \Big|\hat{F}_n(u) - u\Big|\,, \tag{25}$$

the maximal gap between the empirical CDF of the u_k and the diagonal. D_n is sensitive to the overall shape of the PIT histogram: a U-shaped histogram indicates forecasts that are too narrow, a concentration at the center indicates forecasts that are too diffuse, and skewness indicates a biased level.

Tail exceedance and excess kurtosis. Two statistics probe the tails, where the KS distance is blind—a PIT histogram can be nearly uniform in its bulk yet still carry too little tail mass, which is exactly the failure mode that motivates the transient-injection pathway of Equation 12. The tail exceedance $P(z{>}3)$ reports the empirical mass beyond three predicted scale units; under the Rayleigh family its reference is $\exp(-9/2) \approx 1.11\%$. The excess kurtosis of z (Fisher; Gaussian reference 0) aggregates tail weight, with the Rayleigh family itself sitting at -0.27.

L.3 Reconstruction SNR

The reconstruction SNR measures the reconstruction head under teacher forcing. For the predicted waveform $\hat{x}$ and the true next block x,

$$\mathrm{SNR} \;=\; 10\log_{10}\frac{\sum_j x_j^2}{\sum_j \left(x_j-\hat{x}_j\right)^2} \quad \text{(dB)}, \tag{26}$$

computed per channel and averaged over the test split; the persistence reference reconstructs the next block by copying the current one.

L.4 Autoregressive Rollout Metrics

In the rollout protocol, each test sequence provides a context of a few seconds to one minute; the model generates the next M seconds autoregressively—each generated block is appended to the context before the next prediction—and the generated segment is compared against the real future segment of the same record.

psd_corr. The power spectra of the generated and real segments (channel-averaged, estimated by Welch's method) are compared in the log-power domain by their Pearson correlation. The logarithm removes the common power level, so psd_corr isolates spectral *shape*—the relative distribution of power across frequency—while absolute level is delegated to rmsR; a value of 1.0 indicates identical spectral shapes.

rmsR. The generated-to-real RMS ratio, computed per channel and averaged, measures absolute amplitude calibration; 1.0 is calibrated. Because cross-record amplitudes are inherently mismatched, rmsR is read against the real–real floor rather than against unity.

Reference systems. Three references bracket the task. *Persistence* repeats the last context patches—a waveform copy, not a generation—and its high psd_corr shows how much of the metric is achievable without any dynamics. *Surrogate* keeps the context amplitude spectrum with random phase—the right spectrum by construction, but no temporal dynamics. The *real–real floor* scores the future segment of a different record of the same dataset: perfect but unrelated EEG, whose metrics mark the level at which a generated segment is statistically indistinguishable from real data. The floor's own rmsR spans 1.05–1.53 on the evaluation suite, which calibrates how far from unity a faithful generator may sit.

M Full Ablation Results

All ablation variants are evaluated on the full test splits of the same 10 datasets as the main results.

Prediction heads. The head ablations (Table 10) compare the dual-head GenBrain against single-head variants pretrained from scratch under identical schedules; rollouts use pure sampling for all three checkpoints, so the table isolates the effect of the heads themselves. The two single-head ablations settle the division of labor between the heads. Dist-only matches GenBrain's NLL within 0.07 nats on every dataset, so adding the reconstruction head leaves next-step distributional quality intact; Table 2 reports the KS and tail-exceedance diagnostics once for the full model rather than per variant, and its per-model excess kurtosis shows the dual head rescuing the most extreme tails while remaining above dist-only on a subset of datasets. The reconstruction head supplies what the distribution head cannot: waveform reconstruction at an SNR comparable to the dedicated signal-only checkpoint, and the transient-injection pathway that lifts spectral fidelity on 9 of 10 datasets. Signal-only reaches the highest psd_corr only by copying the context spectrum while its deterministic rollout collapses in amplitude (rmsR 0.40–0.71); its induced NLL trails GenBrain by 1.6–7.0

Table 10: Head ablations, per dataset. NLL: teacher-forced negative log-likelihood (lower is better); psd_corr: PSD correlation of pure-sampling autoregressive rollout (higher is better); rmsR: generated-to-real amplitude ratio (1.0 calibrated). Dist = distribution-only head, Signal = signal-only head, Dual = GenBrain.

	NLL↓			psd_corr↑			rmsR		
Dataset	Dual	Dist	Signal	Dual	Dist	Signal	Dual	Dist	Signal
BCIC-IV-2a	4.400	4.404	7.393	0.835	0.831	0.875	0.904	0.909	0.630
CHB-MIT	5.523	5.519	10.893	0.679	0.678	0.850	0.654	0.680	0.428
FACED	4.777	4.787	8.040	0.798	0.799	0.910	0.965	0.987	0.552
ISRUC	4.308	4.331	11.325	0.721	0.703	0.921	1.453	1.486	0.428
MODMA	3.971	3.974	6.108	0.863	0.864	0.848	0.801	0.814	0.453
MentalArithmetic	4.511	4.526	9.473	0.767	0.774	0.927	0.865	0.856	0.604
PhysioNet-MI	5.050	5.052	7.727	0.884	0.885	0.866	0.785	0.800	0.497
SEED-V	4.057	4.058	5.759	0.854	0.844	0.833	0.971	1.073	0.395
SEED-VII	4.365	4.359	5.938	0.913	0.914	0.838	0.859	0.866	0.560
SHU-MI	2.932	2.998	4.928	0.859	0.862	0.823	1.865	1.815	0.711
Mean	4.390	4.401	7.758	0.817	0.815	0.869	1.012	1.029	0.526

nats, and it defines no forecast distribution and hence no PIT. The dual-head design is thus the union of the two specialties rather than a compromise between them.

Table 11: Transient-injection schedule ablation under the dual head, per dataset. M: rollout seconds. Const: constant weight w=0.5; Lin: linear decay $w(1-t/M)$; Geo: geometric decay $w\gamma^t$ with decay rate γ=0.85 (here γ denotes the transient-injection decay rate, not the γ rhythm band); Samp: pure sampling (w=0).

		psd_corr↑				rmsR			
Dataset	M	Samp	Const	Lin	Geo	Samp	Const	Lin	Geo
BCIC-IV-2a	2	0.835	0.866	0.856	0.863	0.904	0.980	0.957	0.972
CHB-MIT	3	0.679	0.739	0.721	0.732	0.654	0.745	0.711	0.728
FACED	3	0.798	0.804	0.802	0.804	0.965	1.072	1.026	1.048
ISRUC	10	0.721	0.680	0.695	0.699	1.453	2.454	1.943	1.874
MODMA	5	0.863	0.898	0.881	0.887	0.801	0.966	0.891	0.911
MentalArithmetic	2	0.767	0.789	0.782	0.787	0.865	0.937	0.910	0.925
PhysioNet-MI	2	0.884	0.910	0.908	0.910	0.785	0.848	0.830	0.842
SEED-V	10	0.854	0.912	0.888	0.888	0.971	1.624	1.261	1.213
SEED-VII	2	0.913	0.941	0.937	0.940	0.859	0.924	0.905	0.917
SHU-MI	2	0.859	0.873	0.867	0.871	1.865	1.961	1.929	1.951
Mean		0.817	0.841	0.834	0.838	1.012	1.251	1.136	1.138

Transient-injection schedule. Table 11 compares transient-injection schedules under the dual head: constant weight w, linear decay $w(1-t/M)$, and geometric decay $w\gamma^t$ with γ=0.85, against pure sampling (w=0). A constant weight (w=0.5) lifts psd_corr furthest but overshoots amplitude on the longest feedback horizons: rmsR reaches 2.45 on ISRUC and 1.62 on SEED-V (M=10), versus 1.45 and 0.97 under pure sampling. A linear decay $w(1-t/M)$ halves the overshoot; the geometric decay is uniformly the best compromise, repairing ISRUC to 1.87 and SEED-V to 1.21 while retaining 59–100% of the constant-weight spectral gain on the 9 datasets where injection helps. One schedule is used for all datasets with no per-dataset tuning. Mean rows average the 10 datasets.

N Downstream Fine-Tuned Decoding

Fine-tuning protocol. Starting from the pretrained dual-head checkpoint (the same model as all generative experiments), we fine-tune all parameters with AdamW (learning rate 10^{-4}, weight decay 5×10^{-2}) and label smoothing 0.1, using a classification head over pooled patch representations; the

Table 12: Downstream decoding under end-to-end fine-tuning on the 10 evaluation datasets (balanced accuracy, %, mean ± std; 5 seeds for GenBrain). Bold = best, underline = second best per row. Baselines are evaluated under the identical data splits, preprocessing pipeline, and end-to-end fine-tuning protocol.

Dataset	EEGNet	Conformer	LaBraM	CBraMod	CSBrain	REVE	GenBrain
Emotion recognition							
FACED	40.90±1.22	45.59±1.25	52.73±1.07	55.09±0.89	**57.52±0.42**	56.46±1.64	57.02±0.64
SEED-V	29.61±1.02	35.37±1.12	39.76±1.38	40.91±0.97	**41.97±0.33**	40.48±0.23	40.77±0.71
SEED-VII	25.28±0.33	28.75±0.28	**33.28±0.33**	32.47±0.36	32.44±0.38	32.64±0.34	33.05±0.67
Motor imagery							
PhysioNet-MI	58.14±1.25	60.49±1.04	61.73±1.22	64.17±0.91	63.04±0.90	**64.80±1.40**	63.09±0.61
BCIC-IV-2a	44.82±0.94	46.96±1.06	48.69±0.85	51.38±0.66	56.57±0.71	**63.96±0.95**	53.13±3.15
SHU-MI	58.89±1.77	59.00±1.07	61.66±1.92	63.70±1.51	**64.17±0.37**	63.84±0.34	62.60±2.02
Sleep staging							
ISRUC	71.54±1.21	74.00±1.33	76.33±1.02	78.65±1.10	**79.25±0.30**	78.19±0.78	78.44±1.93
Seizure detection							
CHB-MIT	56.58±1.06	59.76±1.41	70.75±3.58	**73.98±2.84**	72.62±1.15	72.02±4.95	73.61±1.21
Clinical psychiatry							
MODMA	61.90±2.40	63.85±2.20	66.81±7.29	70.18±2.34	**72.85±2.60**	72.45±2.10	70.45±2.30
Cognitive workload							
MentalArithmetic	67.70±1.16	68.05±1.23	72.24±2.05	74.39±2.15	75.58±1.06	76.60±3.55	**77.36±3.81**
Mean	51.54	54.18	58.40	60.49	61.60	**62.14**	60.95

reconstruction head is dropped. Model selection runs on the validation split (kappa for multiclass, ROC-AUC for binary) with patience-based early stopping, and we report test metrics of the selected model, averaged over 5 random seeds. The comparison covers two supervised architectures, EEGNet (Lawhern et al., 2018) and EEGConformer (Song et al., 2022), and four discriminative EEG foundation models, LaBraM (Jiang et al., 2024b), CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025), and REVE (El Ouahidi et al., 2025), evaluated under the identical data splits, preprocessing pipeline, and end-to-end fine-tuning protocol; each method uses its own fine-tuning recipe. We report balanced accuracy as the common metric (Table 12).

Results. GenBrain attains a mean balanced accuracy of 61.0%, surpassing the first-generation foundation model LaBraM (58.4%) and clearly above the two architectures trained from scratch (EEGNet 51.5%, EEGConformer 54.2%), while trailing the strongest discriminative foundation models (CSBrain 61.6%, REVE 62.1%) by under 1.2 points—models whose pretraining objectives are optimized precisely for discriminative transfer, whereas GenBrain spends its capacity on the generative objective that the main experiments evaluate. The per-dataset picture shows the same structure with genuine highlights: on MentalArithmetic, GenBrain is the strongest method (77.4%, and the best ROC-AUC at 90.8% and PR-AUC at 80.8%); on CHB-MIT it comes second, within 0.4 points of the best (73.6% vs. 74.0%), with the best ROC-AUC (90.5%); on FACED it trails only CSBrain (57.0% vs. 57.5%) while exceeding LaBraM by over 4 points; and on ISRUC it exceeds LaBraM by 2 points (78.4% vs. 76.3%) and lands within 1 point of the best. The remaining gap concentrates where short trials dominate (BCIC-IV-2a). Together with the generative results of Section 3, these numbers indicate that the same checkpoint that generates neuro-mechanistically faithful activity also retains a usable discriminative representation, without any discriminative component in pretraining.

O Discussion

O.1 Limitations

GenBrain's design involves several simplifying assumptions. First, the frequency-domain distribution treats bins within a band as conditionally independent given the band-level parameters; this first-order approximation ignores cross-bin correlations that may carry additional neurophysiological structure. Second, the parameterized family has limited tail expressiveness: residual calibration analysis empirically locates an amplitude-tail mismatch, and the dual-head design is a mitigation rather than a cure—the distribution head itself still does not model heavy-tailed mechanisms. Third,

the phase pathway currently contributes limited signal: residual analysis indicates that the predicted κ and the observed phase concentration disagree in direction (overestimated in some bands, underestimated in θ), so the von Mises head's learning signal is weak. Fourth, autoregressive generation can accumulate error over long horizons, as each sampled segment is fed back into the context without a corrective signal. Fifth, the five canonical rhythm bands are a modeling assumption rather than a biological invariant, and the physiological ranges of the constraints rely on domain knowledge. Finally, GenBrain currently performs unconditional generation and validation is limited to scalp EEG; the architecture does not yet accept conditioning inputs such as task labels, subject identity, or stimulus, and generalization to invasive recordings (ECoG, stereo-EEG) or other modalities is untested.

O.2 Future Work

Several extensions follow naturally. *Heavy-tailed distribution families* (Weibull, Rice) would address the tail mismatch mechanistically rather than structurally. *Within-band correlation modeling* would relax the conditional-independence approximation. *Conditional generation* can be introduced by injecting task, subject, or context embeddings into the prediction head, enabling targeted synthesis for specific BCI paradigms. *Multi-modal pretraining* could extend the distribution-prediction framework to MEG, fMRI, or invasive recordings, leveraging shared neurophysiological structure across modalities (Xiao et al., 2025). *Real-time closed-loop BCI* requires optimizing autoregressive inference latency; the per-step cost is dominated by a single inverse FFT, but the sequential dependency limits parallelism (Casson et al., 2010; Kaongoen et al., 2023). *Data-driven band discovery* could replace the fixed canonical bands with learned frequency partitions. *Scaling* to larger EEG archives and model sizes may further improve representation quality and generation fidelity, and *clinical validation* through prospective studies is needed before any diagnostic or therapeutic use.

O.3 Potential Applications

A generative model of EEG enables several application directions. *Data augmentation*: synthetic EEG can augment scarce cohorts for rare neurological conditions (e.g., rare seizure subtypes, sleep disorders) where labeled data are insufficient for training discriminative models (Luo et al., 2020; Shu et al., 2024). *BCI decoder training*: generated signals can pre-train or regularize decoders in low-data subject-specific calibration settings. *Biomarker discovery*: the predicted distribution parameters θ (band power, phase synchrony, aperiodic slope, coupling strength) are interpretable neurophysiological features that may serve as biomarkers for neurological or psychiatric conditions. *Simulation and protocol testing*: synthetic EEG enables in-silico evaluation of BCI pipelines, artifact-removal algorithms, and experimental designs before deployment. *Education and training*: physiologically plausible generated signals can support neurophysiology teaching and clinician training.

O.4 Comparison with Related Work

GenBrain differs from prior work along four axes, elaborated in Appendix A. *Generative vs. discriminative*: unlike discriminative EEG foundation models, which learn by reconstructing masked segments or predicting statistics, GenBrain explicitly predicts the next-step probability distribution, making generation a native pretraining capability rather than a downstream add-on. *Distribution vs. point prediction*: unlike autoregressive models that predict the next waveform value, GenBrain predicts a parameterized continuous distribution, avoiding the noise-chasing problem of pointwise regression. *Neurophysiologically constrained vs. data-driven*: unlike diffusion- and GAN-based EEG generators, GenBrain's distribution family is derived from computational neuroscience (Rayleigh envelopes, von Mises phases, $1/f$ background, phase-amplitude coupling), so generated samples inherit interpretable structure rather than mimicking waveforms statistically. *Pretrained vs. task-specific*: unlike task-specific generators, GenBrain is pretrained on large unlabeled corpora and supports both generation and downstream representation transfer within a single model.