SCALING NEXT-BRAIN-TOKEN PREDICTION FOR MEG

Richard Csaky
Foresight Institute
richard.csaky@gmail.com

ABSTRACT

We present a large autoregressive model for source-space MEG that extends token-stream "next-X" prediction to brain activity, positioning MEG as an additional modality for multimodal foundation models. We scale next-brain-token prediction to long context across datasets and scanners, handling a corpus of over 500 hours and thousands of sessions across the three largest MEG datasets. A modified SEANet-style vector-quantizer reduces multichannel MEG into a flattened token stream on which we train a Qwen2.5-VL backbone from scratch to predict the next brain token and to recursively generate minutes of MEG by continuing a session from up to a minute of within-session context (a prefix from the same MEG recording). To evaluate long-horizon generation, we introduce task-matched stress tests for (i) on-manifold stability via generated-only drift compared to the time-resolved distribution of real sliding windows, and (ii) conditional specificity via correct context versus prompt-swap controls using a neurophysiologically grounded metric set. We train on CamCAN and Omega and run all analyses on held-out MOUS, establishing cross-dataset generalization. Across metrics, generations remain relatively stable over long rollouts and are closer to the correct continuation than swapped controls. Code available at: https://github.com/ricsinaruto/brain-gen.

1 INTRODUCTION

Predicting the future—the next sensory input, the next state of the world, the next action—is central to both natural and artificial intelligence. In neuroscience, predictive coding and the free-energy principle frame perception and cognition as continual prediction and correction through prediction errors (Rao & Ballard, 1999; Friston, 2010). In machine learning, recent progress has revealed a similar unification across domains: many successful systems can be viewed as solving one and the same causal modeling problem, *given the past, what happens next?*

Across domains, the most successful instantiation of this recipe has been scaling flexible sequence models (Vaswani et al., 2017). Whether through diffusion, normalizing flows, or discrete tokens (Ho et al., 2020; Kingma & Dhariwal, 2018; van den Oord et al., 2017), large Transformer models can predict the next word, the next image patch, the next video frame or audio chunk, and the next action, increasingly exhibiting representations that behave like implicit world models (Vafa et al., 2024). This raises a natural question: if implicit models of the world can emerge from predicting observations of the world (video, audio) or of human behavior (language), what kind of models might emerge from predicting the process that produces intelligent behavior itself: brain activity?

Brain recordings provide a privileged view into the internal dynamics that mediate perception, cognition, and action. In the *learning using privileged information* (LUPI) framework, extra signals available at training time can improve generalization by exposing latent variables that are causally upstream of the observed outputs (Vapnik & Vashist, 2009). Neural signals can act as such a privileged training signal: brain-based objectives can regularize vision models toward robust representations (Li et al., 2019), and "brain-tuning" speech language models on fMRI can induce brain-relevant semantics and improve downstream performance (Moussa et al., 2024). We therefore argue that a powerful *generative* model of brain signals will have to internalize reusable structure about these dynamics, and indirectly, about the world models and future-state predictions the brain itself implements. Such a model could be useful both scientifically (simulation, data augmentation, probing) and as a source of grounding or "privileged teaching" for AI systems trained primarily on observations of behavior.

This motivates our quest here as a first step in this direction: *apply the causal prediction and scaling paradigm to brain data*.

We focus on magnetoencephalography (MEG), a unique and comparatively under-explored non-invasive modality which provides millisecond temporal resolution and high information density relative to EEG. MEG differs sharply from standard generative domains: it is a multi-channel time series of continuous values with low SNR and weak human interpretability, making both modeling and evaluation challenging. Still, the same scaling paradigm that has worked for language, audio, and video should apply to MEG, provided we can represent it as a token sequence that a modern generative backbone can model efficiently.

We are especially interested in pushing this recipe in terms of context length and *conditional specificity*. We do not only want to generate plausible brain activity; we want to generate brain activity that remains specific to the observed context. In this work, the prompt is a within-session conditioning prefix (the first 60 seconds of the same MEG session), and generation is open-loop continuation of that session rather than instruction-following in the LLM sense. The analogy to LLM prompting is therefore only about sensitivity to conditioning context: a large brain model should produce *long-range on-manifold signals that are conditionally specific to the session, subject, and task implied purely by the observed prefix* (without auxiliary embeddings or metadata). This framing also makes brain models compatible with "token-stream" multimodal backbones: in the long run, brain tokens could be interleaved with language, vision, and action tokens in a single causal sequence.

Token-stream modeling is becoming a dominant design pattern in multimodal generative systems: high-bandwidth modalities are first mapped to compact sequences of tokens, and a single decoder-only backbone is trained over interleaved multimodal sequences. Emu3 and Emu3.5 show that a native multimodal Transformer trained solely with next-token prediction over unified vision-language tokens can support both perception and high-fidelity generation, including video synthesis (Wang et al., 2024b; Cui et al., 2025). Qwen2.5-VL and Qwen3-VL extend this token-stream interface to high-resolution inputs and long-video understanding (Bai et al., 2025; Qwen Team, 2024).

To summarize, a good generative MEG model should have: 1. Token-based AR without auxiliary information. 2. Conditional specificity to the input context (conditioning prefix). 3. The ability to ingest long context. 4. Stable and on-manifold long-horizon generation. 5. An efficient and scalable architecture.

To achieve these goals, we propose applying the causal prediction paradigm to MEG by improving the BrainOmni tokenizer (Xiao et al., 2025) and training a Qwen-2.5-VL-style decoder-only backbone from scratch for next-brain-token prediction, without auxiliary task/dataset information. We scale this to the three largest publicly accessible MEG datasets: CamCAN, Omega, and MOUS (Taylor et al., 2017; Niso et al., 2016; Schoffelen et al., 2019), with a combined size of over 500 hours across rest and many diverse tasks. We train on CamCAN and OMEGA and report all results on MOUS, a fully held-out dataset with out-of-distribution tasks.

To address the signal interpretation issue, we propose an extensive evaluation framework comparing neurophysiologically grounded metrics across multiple minutes of free-running recursive generation. Our protocol is designed to evaluate long-horizon on-manifold stability, conditional specificity (via prompt-swap controls), and variability calibration (via a task-matched real-real baseline). Since MEG signals are not directly interpretable to humans, evaluation frameworks that mimic long-range stress tests used for LLMs are especially important.

Contributions. 1. `BrainTokMix`, a causal channel-mixing RVQ tokenizer for source-space MEG that provides a discrete token interface compatible with token-stream multimodal Transformers. 2. `FlatGPT`, a decoder-only Transformer trained from scratch on `BrainTokMix` tokens using standard next-token cross-entropy, enabling multi-minute prefix-and-generate MEG rollouts (within-session continuation) without auxiliary metadata. 3. A *cross-dataset* evaluation protocol that stress-tests long-horizon stability and context dependence using neurophysiological metrics and prompt-swap controls.

2 Related Work

Neural foundation modeling has rapidly adopted self-supervised pretraining and discrete tokenization for heterogeneous EEG/MEG. LaBraM pretrains Transformers over quantized EEG patches via masked prediction (Jiang et al., 2024b); BrainOmni introduces a sensor-aware tokenizer and unified EEG/MEG pretraining (Xiao et al., 2025); and NeuroRVQ studies multi-scale RVQ codebooks for MEG tokenization (Barmpas et al., 2025). Generative models for electrophysiology include autoregressive code models (e.g., MEG-GPT, which focuses on sub-second contexts) and diffusion-style approaches (Lim & Kuo, 2024; Huang et al., 2025). Compared to these efforts, `FlatGPT` emphasizes (i) purely next-token objective over discrete MEG tokens through an efficient and scalable paradigm (ii) long-context conditioning through prompting rather than labels, and (iii) stress-testing open-loop generations for stability and specificity using an out-of-distribution evaluation. Due to space constraints we provide an extended related-work discussion in Appendix Section A.1.

3 Methods

MEG poses an unusual combination of challenges for modern generative modeling: high sampling rate (here 100 Hz), long temporal horizons (tens of seconds to minutes), and multichannel structure ($C = 68$ source-space regions in our main setup). The following inductive biases summarize the constraints that guided our method design: 1. Minute-scale context. 2. Spatiotemporal tokens: A token should represent a temporally and spatially reduced patch of MEG. 3. Flatten into a single sequence: Serialize temporal and spatial axes into one token stream to enable full attention mixing under a standard causal mask. 4. Prefix-only: Condition only on the observed MEG prefix, without auxiliary embeddings or metadata. 5. Pure next-token objective.

We therefore build `FlatGPT` around a simple but scalable decomposition that mirrors frontier LLM/VLM pipelines: (i) learn a causal discrete tokenizer that compresses multichannel MEG into a grid of discrete code indices, and (ii) train a decoder-only Transformer with only teacher-forced next-token prediction (cross-entropy) in token space. This choice is motivated by both scalability and interoperability: once MEG becomes a token stream, we can directly leverage the same decoder-only architectures used for language, audio, and video token streams, and later even interleave with these modalities in a unified token sequence. Appendix Table 3 summarizes where our proposed method sits compared to prior art.

3.1 Problem formulation

A preprocessed MEG recording is a multichannel time series $x \in \mathbb{R}^{C\times T}$ here sampled at $f_s = 100\,\text{Hz}$. Given a context $x_{:t}$, we aim to model the conditional distribution over future activity, $p(x_{t+1:t+H} \mid x_{:t})$, and to generate realistic continuations for long horizons H. We work in a discrete latent space. Let $\mathcal{E}_\psi$ and $\mathcal{D}_\psi$ denote a tokenizer encoder and decoder. For an input segment x we compute discrete codes and decode back to the signal domain:

$$y = \mathcal{E}_\psi(x) \in \{0, \ldots, K-1\}^L, \qquad \hat{x} = \mathcal{D}_\psi(y), \tag{1}$$

where K is the codebook size and L is the flattened token length. We then train an autoregressive model $p_\theta(y)$ with next-token prediction and generate in token space before decoding.

3.2 Tokenization: BrainTokMix

A good MEG tokenizer must trade off three conflicting goals: (i) high compression along *time* and *channels*, (ii) low reconstruction error, and (iii) as few discrete symbols as possible (small vocabulary). In our setting, the tokenizer defines the interface between a high-bandwidth continuous signal and a scalable Transformer, the better the tokenizer, the more "language-like" the downstream modeling becomes. A subtle but important point is that many vision-language models use continuous "tokens" (latent patches) only as conditioning for language generation. In contrast, our downstream model is trained purely by next-token cross-entropy in the token space, which requires a discrete vocabulary.

Why not treat MEG as audio or video directly? MEG has native shape $T \times C$ (time $\times$ channels), unlike audio (T) or video ($T \times H \times W$). Applying an audio codec independently to each channel

would postpone cross-channel mixing until the Transformer and yields redundant tokens due to strong spatial correlations. Rasterizing sensors into an image and treating MEG as a long low-resolution "video" is appealing because video tokenizers mix space and time jointly (Tang et al., 2024; Cui et al., 2025; Agarwal et al., 2025), but we found the resulting representation sparse and the tokenizer slower and less accurate for our setting. These observations motivate a domain-specific tokenizer that (i) mixes channels early and (ii) compresses both time and spatial axes aggressively while remaining causal.

From BrainOmni to `BrainTokMix`. BrainOmni (Xiao et al., 2025) introduced a powerful sensor-aware neural tokenizer for EEG and MEG: it applies a SEANet-style codec (Défossez et al., 2022b) to each channel and then uses a dedicated sensor module to mix across sensors before quantization. We adopt two core ingredients from this line of work: (i) a causal SEANet encoder-decoder backbone and (ii) their reconstruction and frequency-domain objectives (Eq. 2). In our MEG source-space regime, we can simplify the tokenizer and move mixing into the convolutional backbone. `BrainTokMix` removes BrainOmni's sensor encoder and per-window sensor attention, sets sensor embeddings to zero, and performs spatiotemporal mixing via multichannel causal convolutions. This yields an end-to-end causal codec that is easier to train efficiently (no batching over channels, no lstm over time, no attention over the C-sensor axis, and no metadata path) and produces discrete tokens that summarize joint spatiotemporal structure.

3.2.1 CHANNEL-MIXING SEANET BACKBONE

Given a windowed multichannel recording $x \in \mathbb{R}^{C \times L_w}$, we use a strictly causal SEANet encoder-decoder (Défossez et al., 2022b). Concretely, the encoder consists of an initial causal convolution, two strided downsampling blocks (ratios $(2, 2)$; overall hop length $r = 4$), residual blocks with two residual layers. This maps each window to a latent sequence $y \in \mathbb{R}^{T_w \times n_{\text{dim}}}$ with $T_w = L_w/r$ and $n_{\text{dim}} = 4096$. We then reshape the latent dimension into n_{neuro} streams: $y_t \in \mathbb{R}^{n_{\text{neuro}} \times d}$ where $n_{\text{neuro}} = 4$ and $d = n_{\text{dim}}/n_{\text{neuro}} = 1024$. Intuitively, the model first performs spatiotemporal compression in a latent space, and the split exposes a small latent "spatial" axis that the downstream Transformer can attend over. Each latent vector $z_{h,t} \in \mathbb{R}^d$ (stream h, time t) is discretized using a Q-stage residual vector quantizer (RVQ) (Défossez et al., 2022b). The decoder inverts the encoder by concatenating the (summed) quantized streams back into a 4096-D latent per timestep and applying a causal SEANet decoder to reconstruct x.

3.2.2 TOKENIZER TRAINING

The tokenizer is trained end-to-end to reconstruct the input while encouraging informative discrete codes. Given reconstruction $\hat{x}$, we minimize the same loss designed for the BrainOmni tokenizer:

$$\mathcal{L} = \|x - \hat{x}\|_1 + \exp(-\text{pcc}(x, \hat{x})) + \mathcal{L}_{\text{com}} + \mathcal{L}_{\text{amp}} + \tfrac{1}{2}\mathcal{L}_{\text{phi}}, \tag{2}$$

where pcc is the (channel-averaged) Pearson correlation coefficient, $\mathcal{L}_{\text{com}}$ is the RVQ commitment penalty, and $\mathcal{L}_{\text{amp}}$/$\mathcal{L}_{\text{phi}}$ compute the L1 loss between input and reconstruction FFT magnitudes and phases, respectively. We train the tokenizer on 10.24 s windows and then freeze it for autoregressive Transformer training. For a segment of length T (divisible by L_w), tokenization produces RVQ indices

$$c_{t,h,q} \in \{0, \ldots, K-1\}, \tag{3}$$

where $t \in \{1, \ldots, T'\}$, $h \in \{1, \ldots, n_{\text{neuro}}\}$, and $q \in \{1, \ldots, Q\}$. $T' = T/r$ is the downsampled time length. The flattened token length is $L = T' n_{\text{neuro}} Q$, corresponding to a token rate

$$\text{tokens/s} = f_s \cdot \frac{n_{\text{neuro}} Q}{r} = 100 \cdot \frac{4 \cdot 4}{4} = 400. \tag{4}$$

This compression is what makes minute-scale contexts feasible for Transformers. Compared to flattening amplitude-quantized tokens at the full temporal and spatial dimensions this achieves a 17x compression ratio, and it is only 4x higher compared to folding the full spatial dimension into the batch or embedding, i.e. having $f_s = 100$ tokens/s.

3.3 AUTOREGRESSIVE MODELING: FLATGPT

A practical modeling question is how to represent a spatiotemporal signal in a decoder-only Transformer, which expects inputs shaped as $(\text{batch}, \text{length}, \text{embedding})$. There are three natural options:

(1) put channels in the batch (yielding channel-independent models), (2) put channels in the embedding (forcing the model to predict all spatial tokens for a time step jointly, without attention over them), or (3) serialize/flatten spatiotemporal axes into the sequence. We follow option (3), consistent with token-stream video models (Cui et al., 2025; Agarwal et al., 2025): flattening permits full attention across both time and latent spatial streams under a standard causal mask. While this imposes an arbitrary order over the latent spatial axis, (i) the axis is small ($n_{\text{neuro}} = 4$), and (ii) we preserve its identity through axis-aware positional encodings (Section 3.3.1).

We serialize the token grid by iterating RVQ level q fastest: $i = \big((t-1)n_{\text{neuro}} + (h-1)\big)Q + q$, with $y_i \equiv c_{t,h,q}$ and $L = T' n_{\text{neuro}} Q$. We then train a causal Transformer to model $p_\theta(y)$ with the standard teacher-forced cross-entropy loss.

3.3.1 Transformer backbone and MRoPE

Once MEG is represented as a 3D token grid $(T', H', W') = (T/r,\ n_{\text{neuro}},\ Q)$, we can reuse video-capable Transformers. We instantiate a Qwen-2.5-VL-style text Transformer (Qwen Team, 2024; Bai et al., 2025) because it supports multimodal rotary position embeddings (MRoPE) used for flattened video tokens. For each serialized token corresponding to (t, h, q) we provide a 3-tuple position id $\mathbf{p}_i = (p_i^{(t)}, p_i^{(h)}, p_i^{(q)}) = (t,\ h,\ q)$, stacked as $\mathbf{p} \in \mathbb{N}^{3\times B\times L}$. MRoPE applies rotary embeddings to axis-specific subspaces of each attention head, allowing the model to reason about time, space, and even residual code levels distinctly while still using full attention.

3.3.2 FlatGPT with RVQ-aware embeddings

Our `FlatGPT` implementation is a thin wrapper that composes an arbitrary tokenizer with an arbitrary HuggingFace[1] decoder-only Transformer. In our main configuration, we handle RVQ levels explicitly: we use Q separate embedding tables $\{E^{(q)}\}_{q=1}^{Q}$ so that the token embedding depends on the RVQ level, $\text{emb}(y_i) = E_{y_i}^{(q)} \in \mathbb{R}^{d_{\text{model}}}$. The output head is tied to the embedding weights with a one-step cyclic shift across RVQ levels, i.e. the head processing input from RVQ level 0 is tied to the embedding of RVQ level 1. This matches the fixed ordering of RVQ indices within each (t, h) group and keeps parameters minimal. Total vocab size is $Q \times K$. This per-level vocabulary was crucial for good generation.

3.3.3 Generation with a sliding KV cache

To generate long continuations, we encode the provided context into tokens, autoregressively sample future tokens from p_θ, and decode the generated tokens with $\mathcal{D}_\psi$. Because rollouts can exceed the model's nominal context length, we use KV-cached decoding with a sliding-window approach: at generation time we keep a maximum of N context tokens (varies by experiment), and once this is reached we slide at the rate of the tokenizer encoding window (4096 tokens, 10.24 s), refill the KV cache, then generate with caching up to N again, supporting multi-minute conditional generation efficiently. We align the stride to full tokenizer windows to avoid shifting RoPE position embeddings for partial windows; window boundaries can still induce subtle boundary effects, but we found this does not affect generation quality.

3.4 Evaluation of long-horizon rollouts

We sample full held-out sessions and form a context of 61.44 s (start of session), followed by a continuation, with total evaluation segments of 296.96 s (4.95 min). Since many resting-state sessions are 5-minutes long this ensures we can include all sessions in our evaluations. We generate one rollout per context. In a single analysis contexts are always drawn from a single task type from MOUS, i.e. rest, visual, or auditory, which makes our swapped controls rigorous.

We compute feature summaries on 30 s windows with 5 s stride for both generated and real continuations. In our main analyses we show: $1/f$ exponent, channel-covariance eigenvalue entropy, Welch PSD centroid, and an α-bandpower ratio; with additional band-specific spectral metrics, long-range autocorrelation statistics (DFA/Hurst), and cross-channel connectivity summaries (covariance/coherence) in the Appendix. To obtain an interpretable scalar curve we report an *out-of-envelope*

[1] https://huggingface.co/docs/transformers/en/index

rate (OER): for each metric and window, we compute the 5–95% envelope of real continuations and measure the fraction of generated runs outside it.

Let a test segment i be split into a context $c_i \in \mathbb{R}^{C\times T_c}$ and its ground-truth continuation $y_i \in \mathbb{R}^{C\times T_y}$. In our main setting, c_i is the first 60 s of the same recording session and serves only as a within-session conditioning prefix (prompt); the model performs open-loop session continuation, not instruction-conditioned generation. Conditioned on c_i, the model samples an open-loop continuation $x_i \sim p_\theta(\cdot\,|c_i)$. For a prefix time $\tau \in (0, T_y]$ we write $x_i^{\leq\tau}$ and $y_i^{\leq\tau}$ for the first τ seconds of the *continuation* (excluding the context). We embed each prefix into a feature space $\phi(\cdot)$ (e.g., spectral and long-range statistics) and evaluate a distance $d(\cdot,\cdot)$, producing a prefix-divergence curve.

To disentangle conditional specificity from unconditional realism, we pair each context i with a task-matched partner index $j = \pi(i) \neq i$ (i.e. different test session) and define the following per-context controls:

$$D_i^{\text{CORRECT}}(\tau) = d\Big(\phi\Big(x_i^{\leq\tau}\Big)\,,\,\phi\Big(y_i^{\leq\tau}\Big)\Big)\,, \qquad D_i^{\text{PROMPT-SWAP}}(\tau) = d\Big(\phi\Big(x_j^{\leq\tau}\Big)\,,\,\phi\Big(y_i^{\leq\tau}\Big)\Big)\,, \quad (5)$$

$$D_i^{\text{REAL-REAL}}(\tau) = d\Big(\phi\Big(y_j^{\leq\tau}\Big)\,,\,\phi\Big(y_i^{\leq\tau}\Big)\Big)\,, \qquad D_i^{\text{TARGET-SWAP}}(\tau) = d\Big(\phi\Big(x_i^{\leq\tau}\Big)\,,\,\phi\Big(y_j^{\leq\tau}\Big)\Big) \quad (6)$$

This isolates whether generations are (i) closer to the correct continuation than swapped baselines and (ii) calibrated relative to intrinsic real-data variability (variability calibration), rather than being merely on-manifold. We summarize paired effects at the context level using bootstrap confidence intervals (5000 resamples) and Wilcoxon signed-rank tests.

4 Experimental Setup

4.1 Datasets and preprocessing

We train and evaluate on three public MEG datasets that differ in acquisition hardware and protocol: CamCAN (Taylor et al., 2017), OMEGA (Niso et al., 2016), and MOUS (Schoffelen et al., 2019). All recordings are converted to a common representation of $C = 68$ source-space regions (Desikan–Killiany parcels (Desikan et al., 2006)) sampled at $f_s = 100$ Hz, yielding a fixed channel set across datasets, enabling direct cross-dataset training and evaluation. Our preprocessing pipeline includes minimal filtering and outlier removal. Due to space constraints we provide full details in Appendix Section A.3. We train *both* the tokenizer and Transformer on CamCAN+OMEGA and hold out MOUS entirely for validation and testing. MOUS subjects are split 50/50 into val/test with a fixed random seed. After cleaning, this yields 2684 training sessions from CamCAN and 1719 from OMEGA (420 hours, 6×10^8 tokens), and 198/191 MOUS sessions for validation/testing (roughly 70 hours each). See Appendix Table 4 for a summary table.

4.2 BrainTokMix and FlatGPT setup

`BrainTokMix` uses a causal SEANet with window length $L_w = 1024$ samples (10.24 s), downsampling ratios $(2, 2)$, $n_{\text{filters}} = 1024$, $n_{\text{dim}} = 4096$, and $n_{\text{neuro}} = 4$ streams (token width $d = 1024$). For the RVQ we use $Q = 4$ codebooks, codebook size $K = 16384$, and code dimension 1024. Full model size is 294M parameters. After training, we freeze the tokenizer weights for all autoregressive experiments. We instantiate `FlatGPT` with a Qwen2.5-VL-style decoder-only Transformer backbone[2] and train it from scratch on `BrainTokMix` tokens. The backbone has 12 layers, hidden size 1200, 10 attention heads (2 KV heads), head dimension 120, and MLP width 4560, in total 336M parameters.

We train `BrainTokMix` with the objective in Eq. 2 using 10.24 s examples. We use AdamW (Loshchilov & Hutter, 2017) (lr 5×10^{-5}, weight decay 10^{-2}), linear warmup over 300 steps, and gradient clipping of 1.0. The batch size is 480 windows, i.e., $480 \times 10.24s \approx 82$ minutes of MEG per optimization step. We train for 20 epochs, which takes about 5 hours on a B200 GPU. VQVAEs are hard to overfit even without regularization and we simply stop training when improvement over one epoch is marginal. In additional runs, increasing tokenizer capacity improved reconstruction, but

[2]https://huggingface.co/docs/transformers/en/index

Table 1: **Conditional specificity at 235.5 s.** We report paired median improvements Δ (control $-$ correct) with 95% bootstrap CIs on the MOUS test set. *Prompt-swap* tests dependence on the correct conditioning prefix. *Real-real* compares to a task-matched baseline distance between two real continuations (variability calibration). Larger Δ means the correct generation is closer to the target than the control.

Task	n	Control	Covariance distance	PSD JSD	Coherence distance
Auditory	41	Prompt-swap	0.130 [0.075,0.235]	0.048 [0.024,0.073]	0.0065 [0.0052,0.0086]
		Real-real	0.097 [0.062,0.181]	0.047 [0.025,0.054]	0.0059 [0.0030,0.0077]
Visual	34	Prompt-swap	0.096 [0.027,0.129]	0.027 [0.011,0.052]	0.0099 [0.0083,0.0112]
		Real-real	0.076 [0.001,0.149]	0.046 [0.024,0.067]	0.0065 [0.0055,0.0122]
Rest	71	Prompt-swap	0.088 [0.063,0.173]	0.062 [0.042,0.077]	0.0096 [0.0080,0.0115]
		Real-real	0.098 [0.046,0.135]	0.074 [0.050,0.080]	0.0074 [0.0053,0.0096]

we found the gains modest relative to the extra compute and therefore use this 294M configuration as a practical trade-off (see Appendix A.7).

We train `FlatGPT` with AdamW with learning rate 2×10^{-4}, weight decay 0.1, linear warmup over 2000 steps, and gradient clipping at 1.0. With a token rate of 400 tokens/s, a 61.44 s example contains 24,576 tokens. At batch size 8 this corresponds to 196,608 tokens per optimization step. We use early-stopping on the MOUS validation sessions, resulting in 8 epochs. 1 epoch takes about 1.7 hours on a B200 GPU. Both the tokenizer and backbone are trained with BF16 mixed precision and `torch.compile`. For the Qwen backbone we found the cuDNN sdpa backend the fastest[3].

4.3 Generation and evaluation protocol

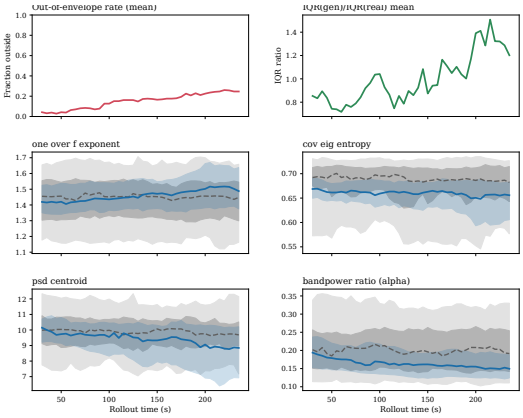
All results are reported on the MOUS **test** split. For each evaluation run and task type we sample all available session segments and use the first 61.44 s as context and a total length of 296.96 s, i.e., 235.52 s of open-loop continuation. We generate 1 rollout (94k tokens) per context with temperature $= 1.0$ and top-$p = 1.0$ for sampling, i.e., pure multinomial sampling; alternative sampling heuristics (e.g., lower top-p or per-RVQ-level temperature schedules) generally worsened rollouts. We evaluate MOUS task types independently: auditory/listening ($n = 41$), visual/reading ($n = 34$), and resting-state ($n = 71$). The sum of these is less than the number of test sessions due to not all sessions having a contiguous 5-minute beginning after the session cleaning. We compute *prefix divergence curves* at prefix times $\tau \in \{20, 40, 60, 80, 100, 150, 200, 250\}$ s, plus the max prefix. In our main analyses we use the following distances and features: 1. normalized L2 distance between channel-covariance matrices, 2. Jensen-Shannon divergence between PSD distributions, 3. normalized L2 distance between broadband coherence matrices, with many additional metrics in the Appendix.

5 Results

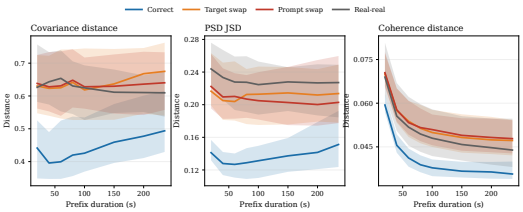
5.1 BrainTokMix reconstruction fidelity

Because `FlatGPT` operates purely in `BrainTokMix` token space, tokenizer reconstruction bounds downstream signal fidelity. On held-out MOUS, `BrainTokMix` achieves low reconstruction error ($\mathrm{MAE} = 0.2$, $\mathrm{PCC} = 0.944$) with high channel-wise correlation and near-uniform code usage (Appendix Table 5). Note that this is much better than what was reported in Xiao et al. (2025), likely due to improved model expressivity and scaling of model size. A small but consistent attenuation of high-frequency power is visible in the reconstructed PSD (Appendix Figure 2); this likely contributes to, slightly reduced gamma-band power in long-horizon generations. In additional experiments, near-perfect gamma-band reconstruction can be achieved at the expense of doubling the number of tokens. Pushing either temporal or spatial reduction further (i.e. $2\times$ our current setup) resulted in worse reconstruction quality, and a maximum PCC of 0.9. Increasing the number of RVQ levels mitigates this, but then the actual tokens/s is not reduced (due to our flattening approach). Therefore our current setup is quite close to optimality in terms of the reduction–reconstruction trade-off.

[3] https://docs.pytorch.org/docs/stable/backends.html



(a) **On-manifold stability.** Gray bands show the real 5–95% and 25–75% envelopes; blue shows the generated distribution across contexts. Top 2: mean OER and IQR ratio. Bottom 4: feature stability.



(b) **Conditional specificity.** Prefix divergence over increasing generated duration τ for the correct pairing (blue) versus prompt-swap (red) target-swap (orange) controls and a real-real baseline (gray). Shaded regions show interquartile ranges across contexts.

Figure 1: **Main analysess on resting-state rollouts (MOUS test).**

5.2 On-manifold stability and conditional specificity

We first test whether open-loop generation drifts off-manifold using the *out-of-envelope rate* (OER; Section 4.3). Across all three tasks (rest shown in Figure 1a), generated windows largely remain within the distributional envelope of real windows for key neurophysiological summaries, with drift accumulating gradually over the 4 min continuation. Full stability plots for each task (including band-specific spectral metrics, DFA/Hurst exponents capturing long-range autocorrelation) are provided in Appendix Figures 7 to 9.

We next test whether generations are *conditionally specific* to the correct prompt and continuation using prefix divergence curves with task-matched swap controls (Eq. 6). Figure 1b shows our main metrics for rest only, with all metrics and task types in Appendix Figures 10 to 12. Correct generations are consistently closer to the true continuation than task-matched controls and the real-real baseline, but the gap does decrease with increased rollout horizon. Table 1 quantifies these gaps at the end of the rollout, supporting both conditional specificity (prompt-swap) and variability calibration against natural real variability (real-real). Prompt dependence persists far beyond the conditioning window: at 235.5 s generated, correct continuations reduce covariance distance by 0.088–0.130 relative to prompt-swap controls, and remain closer than the real-real baseline (Table 1; target-swap shown in Appendix Table 6). Qualitatively, long-horizon generations preserve global structure: average covariance heatmaps and PSDs closely match ground-truth across tasks (Appendix Figures 4 to 6). Representative time-series and STFT rollouts qualitatively resemble their targets without obvious artifacts (Appendix Figures 13 and 14).

Shorter model and conditioning contexts degrade both stability and conditional specificity. Reducing the context from 61.44 s to 30.72 s increases mean OER on nearly all stability metrics, and shrinks the correct-vs-swap gaps (Appendix Section A.12). Table 2 quantifies how prompt-swap separation at 235.5 s weakens with a 30 s context. Teacher-forced loss also decreases slightly with context length (Appendix Figure 3).

6 Discussion

We introduced `BrainTokMix`, a causal spatiotemporal RVQ tokenizer for fixed-channel-order MEG, and `FlatGPT`, a decoder-only Transformer trained on the resulting flattened token stream. Training the tokenizer *and* Transformer backbone on CamCAN+OMEGA and evaluating solely on held-out MOUS, `FlatGPT` can condition on 1 minute of context and generate at least 4 minutes of open-loop continuation while (i) largely staying within the real-data envelope of neurophysiological summaries, and (ii) remaining measurably dependent on the specific conditioning prefix (within-session context).

Table 2: **Prompt-swap separation at 235.5 s: 60 s vs. 30 s context.** We report paired median improvements Δ (control $-$ correct) for the prompt-swap control and the corresponding p-value for the 30 s context (paired Wilcoxon).

Task	Metric	Δ (60 s)	Δ (30 s)	p (30 s)
Auditory	covariance	0.130	0.087	5.4e-3
Auditory	PSD-JSD	0.048	0.045	7.9e-4
Auditory	coherence	0.0065	0.0048	1.9e-10
Visual	covariance	0.096	0.050	0.041
Visual	PSD-JSD	0.027	0.009	0.37
Visual	coherence	0.0099	0.0052	1.0e-8
Rest	covariance	0.088	0.086	9.0e-6
Rest	PSD-JSD	0.062	0.044	1.3e-6
Rest	coherence	0.0096	0.0062	3.6e-13

The tokenizer sets sequence length and determines whether the downstream autoregressive distribution is learnable. In our experiments, BrainOmni-style tokenization achieved comparable reconstruction quality (to `BrainTokMix`) at the same reductions but was roughly $\sim 3\times$ slower to train, and a VidTok (Tang et al., 2024) baseline was substantially slower and reached only 0.90 PCC. In additional experiments (not shown), we found diminishing returns from simply increasing codebook size, whereas adding RVQ levels can improve fidelity but seems to make later levels harder to predict and destabilize long rollouts. Transformer context scaling also had diminishing returns: a long-context curriculum (progressively increasing context length/RoPE parameters up to 160 s) did not improve rollout metrics. More practical takeaways and lessons learned are discussed in Appendix A.7

Due to lack of good baselines we omitted baseline sweeps. To our knowledge, the recent MEG-GPT (Huang et al., 2025) is the only prior multi-channel brain foundation model demonstrated for open-loop generation, but it reports training and generation with an 800 ms context and, in our setup, took roughly $10\times$ longer to train than `FlatGPT`; extending it to minute-scale contexts would be computationally prohibitive, so an apples-to-apples comparison is currently impractical. Classical AR/VAR baselines can match coarse PSD statistics, but fail to reproduce cross-channel covariance and transient events (see our time-series/STFT plots), making them weak comparators for conditional long-horizon generation (Csaky et al., 2024).

Our experiments focus on *brain-to-brain* continuation, but the underlying design is intentionally compatible with multimodal learning. `BrainTokMix` plays the role of a modality tokenizer (analogous to audio/video codecs), while `FlatGPT` is a standard token-stream decoder with multimodal rotary position embeddings (MRoPE) that can attend over spatiotemporal token grids. This makes it straightforward to extend the same interface to multimodal transformations, e.g., conditioning MEG generation on tokenized video/audio/text stimuli, or using MEG tokens as a privileged input when training vision-language models.

ACKNOWLEDGEMENTS

This research was fully funded by an AI Safety Grant from the Foresight Institute. Data collection and sharing for this project was provided by the Cambridge Centre for Ageing and Neuroscience (CamCAN). CamCAN funding was provided by the UK Biotechnology and Biological Sciences Research Council (grant number BB/H008217/1), together with support from the UK Medical Research Council and University of Cambridge, UK. Data were provided (in part; MOUS dataset) by the Radboud University, Nijmegen, The Netherlands.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. URL `https://arxiv.org/abs/2501.03575`.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns LLMs to the human brain. *arXiv preprint arXiv:2312.00575*, 2023. URL https://arxiv.org/abs/2312.00575.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Konstantinos Barmpas, Na Lee, Alexandros Koliousis, Yannis Panagakis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Neurorvq: Multi-scale eeg tokenization for generative large brainwave models, 2025. URL https://arxiv.org/abs/2510.13068.

Paul J Besl and Neil D McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791. URL https://ieeexplore.ieee.org/document/121791.

Richard Csaky, Mats WJ van Es, Oiwi Parker Jones, and Mark Woolrich. Foundational gpt model for meg. *arXiv preprint arXiv:2404.09256*, 2024. URL https://arxiv.org/abs/2404.09256.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025. URL https://arxiv.org/abs/2510.26583.

Anders M Dale, Arthur K Liu, Bruce R Fischl, Randy L Buckner, John W Belliveau, John D Lewine, and Eric Halgren. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26(1):55–67, 2000. doi: 10.1016/S0896-6273(00)81138-1. URL https://pubmed.ncbi.nlm.nih.gov/10798392/.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2024. URL https://arxiv.org/abs/2310.10688.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech from non-invasive brain recordings. *arXiv preprint arXiv:2208.12266*, 2022a. URL https://arxiv.org/abs/2208.12266.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022b. URL https://arxiv.org/abs/2210.13438.

Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Patricia Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006. doi: 10.1016/j.neuroimage.2006.01.021.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787. URL https://doi.org/10.1038/nrn2787.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. Increasing alignment of large language models with language processing in the human brain. *Nature Computational Science*, 5(11):1080–1090, 2025. doi: 10.1038/s43588-025-00863-0. URL https://www.nature.com/articles/s43588-025-00863-0.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1. *arXiv preprint arXiv:2310.03589*, 2023. URL https://arxiv.org/abs/2310.03589.

Chetan Gohil, Rukuang Huang, Evan Roberts, Mats WJ van Es, Andrew J Quinn, Diego Vidaurre, and Mark W Woolrich. osl-dynamics: A toolbox for modelling fast dynamic brain activity. *bioRxiv*, pp. 2023–08, 2023. doi: 10.1101/2023.08.07.549346. URL https://doi.org/10.1101/2023.08.07.549346.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Thomas Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013. doi: 10.3389/fnins.2013.00267.

Matti S H"am"al"ainen and Risto J Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, 32(1):35–42, 1994. doi: 10.1007/BF02512476. URL https://pubmed.ncbi.nlm.nih.gov/8182960/.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Rukuang Huang, Sungjun Cho, Chetan Gohil, Oiwi Parker Jones, and Mark Woolrich. Meggpt: A transformer-based foundation model for magnetoencephalography data, 2025. URL https://arxiv.org/abs/2510.18080.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals, 2024a. URL https://arxiv.org/abs/2409.00101.

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

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1807.03039.

Zhe Li, Wieland Brendel, Edgar Y. Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian H. Sinz, Xaq Pitkow, and Andreas S. Tolias. Learning from brains how to regularize machines. *arXiv preprint arXiv:1911.05072*, 2019. URL https://arxiv.org/abs/1911.05072.

Jia-He Lim and Po-Chih Kuo. Eegtrans: Transformer-driven generative models for eeg synthesis, 2024. URL https://openreview.net/forum?id=ydw2l8zgUB. Submitted to ICLR 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. URL https://arxiv.org/abs/1711.05101.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning. *arXiv preprint arXiv:2410.09230*, 2024. URL https://arxiv.org/abs/2410.09230.

Guillaume Niso, Catherine Rogers, Justin T. Moreau, Li-Yuan Chen, Carole Madjar, Samir Das, Eva Bock, Francois Tadel, Alan C. Evans, Pierre Jolicoeur, and Sylvain Baillet. Omega: The open meg archive. *NeuroImage*, 124:1182–1187, 2016. doi: 10.1016/j.neuroimage.2015.04.028. URL https://doi.org/10.1016/j.neuroimage.2015.04.028.

Fabian Pedregosa, Ga"el Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://arxiv.org/abs/1201.0490.

Qwen Team. Qwen2.5: A party of foundation models. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Rajesh P N Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580. URL https://doi.org/10.1038/4580.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023. URL https://arxiv.org/abs/2310.08278.

Jan-Mathijs Schoffelen, Robert Oostenveld, Nietzsche H. L. Lam, Julia Udden, Annika Hulten, Peter Hagoort, et al. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, 6(17), 2019. doi: 10.1038/s41597-019-0020-y. URL https://doi.org/10.1038/s41597-019-0020-y.

Anni Tang, Tianyu He, Junliang Guo, Xinle Cheng, Li Song, and Jiang Bian. Vidtok: A versatile and open-source video tokenizer. *arXiv preprint arXiv:2412.13061*, 2024. URL https://arxiv.org/abs/2412.13061.

Samu Taulu and Juha Simola. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and Biology*, 51(7):1759–1768, 2006. doi: 10.1088/0031-9155/51/7/008.

Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Richard N. Henson, and Cam-CAN. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 2017. doi: 10.1016/j.neuroimage.2015.09.018. URL https://doi.org/10.1016/j.neuroimage.2015.09.018.

Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024. URL https://arxiv.org/abs/2406.03689.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. URL https://arxiv.org/abs/1711.00937.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042. URL https://doi.org/10.1016/j.neunet.2009.06.042.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Julius Vetter, Jakob H Macke, and Richard Gao. Generating realistic neurophysiological time series with denoising diffusion probabilistic models. *Patterns*, 5(9):101047, 2024. doi: 10.1016/j.patter.2024.101047. URL https://doi.org/10.1016/j.patter.2024.101047.

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 *Advances in Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=lvS2b8CjG5.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b. URL https://arxiv.org/abs/2409.18869.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D. Calhoun. fmri-lm: Towards a universal foundation model for language-aligned fmri understanding, 2025. URL https://arxiv.org/abs/2511.21760.

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, 2025. URL https://arxiv.org/abs/2505.18185.

Yiqian Yang, Yiqun Duan, Hyejeong Jo, Qiang Zhang, Renjing Xu, Oiwi Parker Jones, Xuming Hu, Chin-teng Lin, and Hui Xiong. Neugpt: Unified multi-modal neural gpt, 2024. URL https://arxiv.org/abs/2410.20916.

A Appendix

A.1 Extended Related Work

Generative modeling and forecasting of electrophysiology. Beyond representation learning, there is growing interest in generative models that can synthesize realistic neural signals. EEGTrans uses a quantized autoencoder together with an autoregressive Transformer decoder to generate discrete EEG code sequences for data synthesis (Lim & Kuo, 2024). MEG-GPT trains an autoregressive Transformer with next-step prediction on tokenized MEG region time courses, showing that generated signals match spatio-spectral properties and can improve downstream decoding (Huang et al., 2025). In parallel, diffusion models and other continuous generative approaches have been explored for time-series generation and forecasting. Compared to these efforts, `FlatGPT` emphasizes (i) purely next-token objective over discrete MEG tokens through an efficient and scalable paradigm (ii) long-context conditioning through prompting rather than task labels, and (iii) stress-testing open-loop generations for stability and context specificity across datasets.

Time-series foundation models. Outside neuroscience, recent work has started to build general-purpose *time-series foundation models* (TSFMs) by pretraining large Transformers on large corpora of heterogeneous time series and evaluating them in zero-/few-shot forecasting settings. Representative examples include decoder-only pretrained forecasters such as TimesFM (Das et al., 2024) and TimeGPT (Garza et al., 2023), probabilistic TSFMs such as Lag-Llama (Rasul et al., 2023), and approaches that explicitly discretize values and apply language-model training, such as Chronos (Ansari et al., 2024). While the primary goal of TSFMs is typically accurate and transferable forecasting for generic (often low-dimensional) time series, `FlatGPT` targets a different axis: building a generative prior over high-bandwidth multichannel MEG and stress-testing *open-loop* rollouts for long-horizon stability and context dependence.

Brain-language alignment and multimodal neural models. Generative models of neural signals are also motivated by downstream decoding tasks, such as reconstructing stimuli or behavior. For example, MEG can be used to decode continuous speech from non-invasive recordings (Défossez et al., 2022a). More recently, several works treat brain activity as a "foreign language" by learning neural tokenizers and coupling them to LLM backbones. NeuroLM learns a text-aligned EEG tokenizer and uses instruction tuning for multi-task EEG inference (Jiang et al., 2024a). NeuGPT and fMRI-LM similarly aim to jointly model neural tokens and text to enable language-conditioned understanding from neural recordings (Yang et al., 2024; Wei et al., 2025). Orthogonally, work in cognitive NLP studies representational alignment between LLMs and neural responses, including the effect of instruction tuning (Aw et al., 2023) and evidence that model scaling and training choices can systematically increase alignment (Gao et al., 2025). Related "brain-tuning" approaches fine-tune speech/language models directly on fMRI to induce brain-relevant semantics (Moussa et al., 2024). These approaches typically rely on curated neural-text alignment or task supervision; `FlatGPT` is complementary in targeting an unsupervised generative prior over MEG dynamics, which could serve as a backbone for future multimodal conditioning or decoding.

Evaluating generative neural models. Unlike text or images, the realism of generated MEG cannot be judged visually, and models can match simple marginal statistics while failing to respect the conditioning prompt or long-range dynamics. Most prior work reports token reconstruction, masked prediction accuracy, or downstream decoding performance (Wang et al., 2024a; Jiang et al., 2024b; Xiao et al., 2025; Huang et al., 2025). To evaluate open-loop generation, we introduce metrics and controls that probe (i) distributional drift over long rollouts and (ii) context specificity via prompt swapping and permutation-style controls. This evaluation perspective mirrors how generative models are stress-tested in other modalities, but is adapted to the unique challenges of electrophysiology.

A.2 Positioning relative to prior work

A.3 Preprocessing details

Stage 1: preprocessing and source projection. We use an OSL (Gohil et al., 2023)/MNE-Python (Gramfort et al., 2013) preprocessing pipeline. For CamCAN we apply Maxwell filtering (Taulu & Simola, 2006), for MOUS and OMEGA we apply gradient compensation (grade 3). Then we run a minimal pipeline for each dataset consisting of a causal notch filter at the line noise frequency, then a causal bandpass filter between 1 and 50 Hz, and causal resampling to 100 Hz. Bad channel detection is run and, when metadata exist, bad channels are interpolated. We project sensor data to the `fsaverage` template and extract ROI time courses, yielding a consistent 68-channel source-space signal per session. While this does not give

Table 3: High-level positioning of `FlatGPT` relative to closely related work. Entries for prior work summarize the primary setting emphasized in each paper. (Huang et al., 2025; Xiao et al., 2025; Jiang et al., 2024b; Vetter et al., 2024)

	`FlatGPT` (ours)	**MEG-GPT**	**BrainOmni**	**LaBraM**	**NTD**
Space	source	source	sensors	sensors	sensors
Tokenizer	RVQ	lossless	RVQ	VQ	none
Training	AR	AR	masked	masked	denoise
Context	60s	0.32s	30s	4-8s	1-4s
Generation	240s	60s	N/A	N/A	1-4s
Data	multi-dataset	single-dataset	multi-modality	multi-dataset	single-dataset
Eval	held-out dataset	within-subject	held-out dataset	held-out dataset	within-subject

Table 4: **Cleaned dataset breakdown.** Session counts are after cleaning; hours/tokens are totals per split. H refers to number of hours.

Dataset	Split	Sessions	H.	Tokens
CamCAN	train	2684	420	6×10^8
OMEGA	train	1719		
MOUS	val	198	70	1×10^8
MOUS	test	191	70	1×10^8

the most accurate source localization, we did not have access to subject MRIs for each dataset; for the purpose of cross-dataset generative modeling, a consistent and conservative projection is preferable to dataset-specific pipelines.

We perform MRI-less coregistration to the `fsaverage` template using digitized fiducials and head-shape points (conservative ICP; MNE defaults) (Besl & McKay, 1992; Gramfort et al., 2013). We then compute an `ico5` forward model (BEM; `mindist`=3 mm) and obtain dSPM minimum-norm source estimates (`snr`=3, `loose`=0.2, `depth`=0.8, ad-hoc noise covariance) with fixed normal orientation (H"am"al"ainen & Ilmoniemi, 1994; Dale et al., 2000). Finally, we extract Desikan–Killiany ROI time courses (mode=`mean`; MNE default) and linearly detrend each ROI, yielding a consistent 68-channel source-space signal per session.

Stage 2: session cleaning. We apply robust normalization per session and channel using scikit-learn's `RobustScaler` (median/IQR; defaults) (Pedregosa et al., 2011). We split the signal into fixed windows, drop windows whose standard deviation exceeds a threshold, and discard sessions with too many bad windows. In our runs we use 5 s windows, a standard-deviation threshold of 1.5, and discard sessions with more than 20% bad windows. We clip remaining samples to $[-10, 10]$ (in normalized units) and save contiguous "good" segments that are at least 60 s long, discarding any shorter segments.

Table 4 summarizes the cleaned dataset sizes used in our experiments. Hours refer to the total duration of contiguous "good" segments retained after Stage 1–2 preprocessing (Section 4.1). Token counts are obtained by multiplying hours by the tokenizer rate (400 tokens/s).

A.4 Tokenizer diagnostics

Table 5: **`BrainTokMix` reconstruction metrics on held-out MOUS**. Maximum codebook usage perplexity is 16,384.

Metric	Value
MAE ↓	0.203
PCC ↑	0.944
FFT amplitude error ↓	0.0835
FFT angle error ↓	0.806
Commit loss ↓	2.67×10^{-4}
Codebook perplexity ↑	15,518

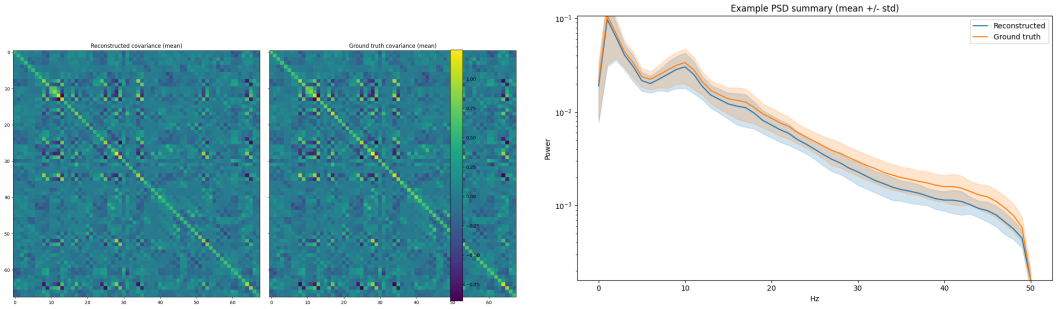
A.5 Target-swap statistics

A.6 Token-level loss vs. context length

These teacher-forced summaries (Figure 3) quantify how next-token prediction improves as more real context is available. The periodic "sawtooth" structure in bits-per-token and perplexity is due to the tokenizer window length (10.24 s), which induces window-aligned shifts in token statistics.

A.7 Extended Discussion

We use *scaling* primarily to mean scaling in data: making a single model work across hundreds of hours and thousands of sessions drawn from multiple datasets and scanners, and then evaluating out-of-distribution on a fully held-out dataset. This is hard: MEG variability across sessions (subjects, tasks, hardware) is large. In exploratory baselines, several channel-mixing sequence models that worked on a handful of sessions collapsed



(a) Covariance structure averaged over held-out windows.

(b) Power spectra averaged over held-out windows and channels.

Figure 2: **`BrainTokMix` reconstruction preserves spatial and spectral statistics.** Reconstructions closely match target covariance and PSD across held-out MOUS windows, with mild attenuation at higher frequencies (likely contributing to slightly reduced gamma-band power downstream).

Table 6: **Target-swap control at 235.5 s generated** (paired median Δ with 95% bootstrap CI; MOUS test). Δ is reported as (target-swap − correct), so larger is better.

Task	Covariance distance	PSD-JSD	Coherence distance
Auditory	0.119 [0.057,0.181]	0.042 [0.031,0.056]	0.0096 [0.0065,0.0118]
Visual	0.081 [0.058,0.115]	0.037 [0.021,0.050]	0.0085 [0.0042,0.0122]
Rest	0.135 [0.062,0.185]	0.056 [0.049,0.074]	0.0101 [0.0077,0.0116]

when trained even on a full single-dataset corpus, suggesting that implicit robustness to variability is a key bottleneck.

Tokenizer lessons: reconstruction, compression, and predictability must be balanced. We observed that window length is a real modeling constraint: longer windows help reconstruction by allowing the codec to use past context within the window, but windowing can introduce periodic effects in token statistics (reflected in the sawtooth token-loss curves; Appendix Figure 3). Importantly, in OMEGA-only trials we obtained similar downstream generation results with a much shorter tokenizer window (1.28 s), suggesting the model cannot "cheat" by relying on windowing; if anything, boundary effects make the autoregressive task harder. While overlap-add decoding can reduce boundary artifacts for reconstruction, it is not available for open-loop autoregressive generation because overlapping regions would require future tokens.

We also tried several other modifications of BrainOmni, including interleaving temporal and spatial reductions, but training proved difficult. Compared to the original BrainOmni setup, removing channel-masking, denoising, and normalization improved reconstruction quality substantially. Since the tokenizer is not able to overfit anyway (due to the RVQ bottleneck and large reductions) we believe there is no need for these further regularization techniques.

Practical scaling notes for long-context MEG. A recurring theme in this work is that "LLM-style simplicity" is a feature: `FlatGPT` uses the standard next-token objective, standard decoder-only training, and standard KV-cached sampling with a sliding context window. In our experience, scaling is most constrained by data heterogeneity rather than architectural novelty: getting a single model to train stably across hundreds of hours (420 after cleaning) and thousands of sessions spanning multiple scanners and tasks is challenging.

Sliding-window attention masks provided only modest training speedups and slightly degraded generation quality; we suspect this is because the flattened token stream interleaves channels/streams and benefits from full causal coupling to maintain covariance structure. Transformer scale exhibited the expected compute trade-off: smaller backbones could reach comparable performance, but typically required more epochs to do so, reducing the effective compute savings. Backbone choice also mattered: in our OMEGA-only trials, Qwen2.5 training was more stable than some alternative bases (including a Qwen3 variant, which produced noisier rollouts).

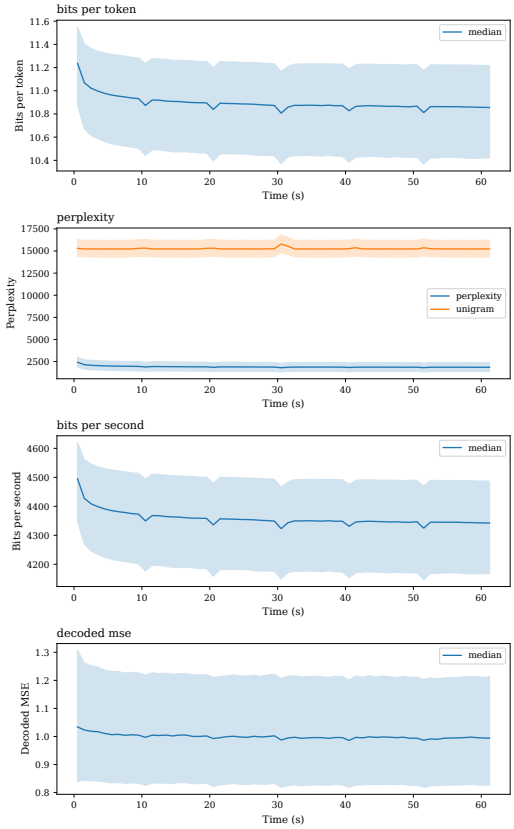
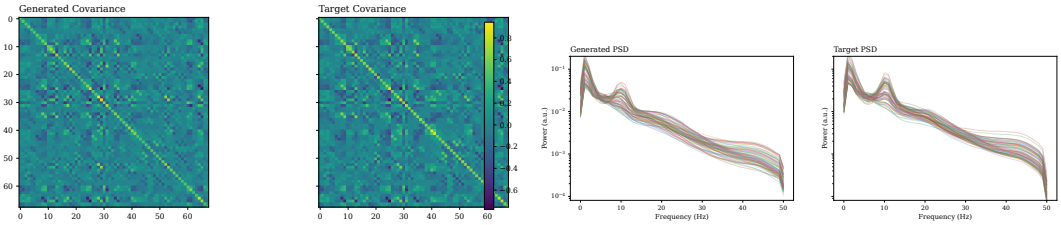


Figure 3: **Token-level prediction vs. available context on test data.**

We have tried `FlatGPT` variants where the RVQ levels are folded (concatenating embeddings) into the hidden dimension of the Transformer to reduce sequence length and be predicted jointly at each step. While this does improve training speed substantially long rollouts were less stable, leading to early degeneration.

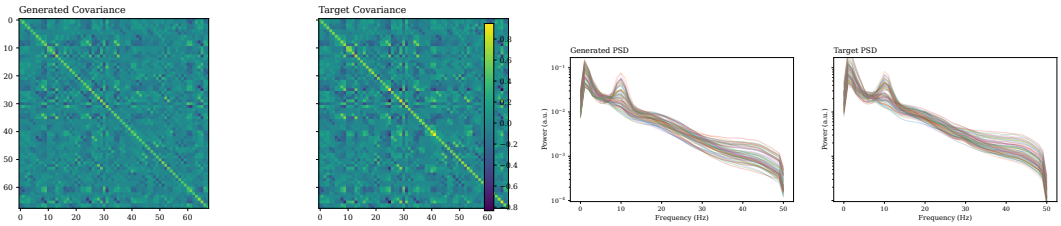
A.8 Global metrics for 60 s-context rollouts

A.9 Full stability metrics for 60 s-context rollouts



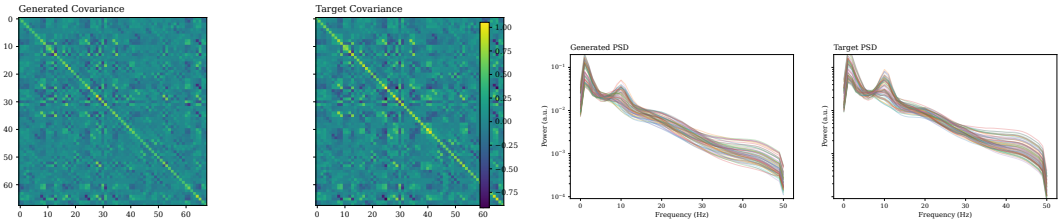
(a) Auditory: covariance. (b) Auditory: PSD.

Figure 4: **Global covariance and PSD for auditory rollouts (60 s context).** Left: covariance heatmaps averaged over generated and target continuations. Right: channel PSDs (0–50 Hz).



(a) Visual: covariance. (b) Visual: PSD.

Figure 5: **Global covariance and PSD for visual reading rollouts (60 s context).**



(a) Rest: covariance. (b) Rest: PSD.

Figure 6: **Global covariance and PSD for resting-state rollouts (60 s context).**

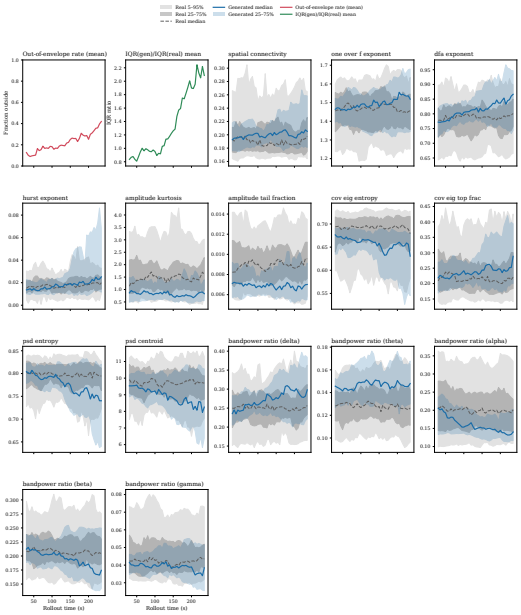


Figure 7: **Auditory (60 s context): full sliding-window stability.**

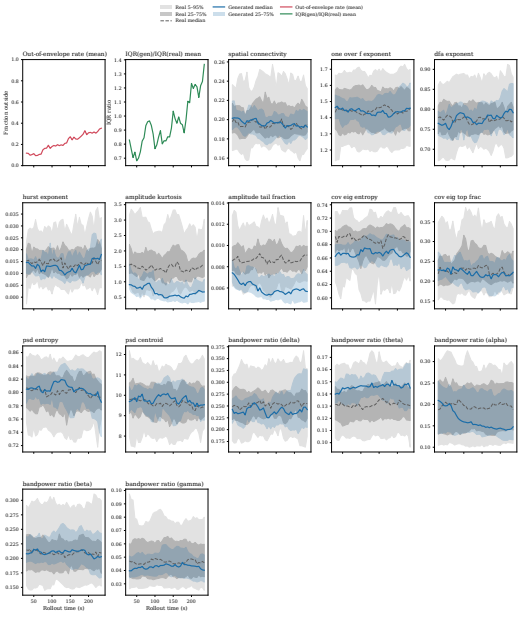


Figure 8: **Visual (60 s context): full sliding-window stability.**

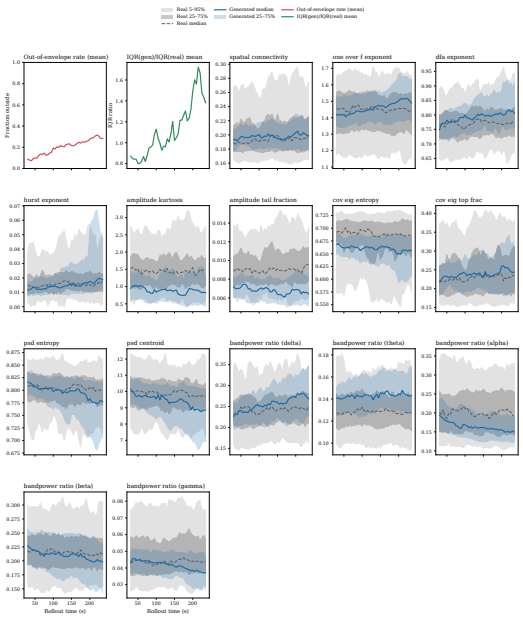


Figure 9: **Rest (60 s context): full sliding-window stability metrics.** Note that correlation and stft/fft angle are expect to have high distance due to phase/dynamics-misalignment between generated and real data.

A.10 FULL PREFIX-DIVERGENCE FOR 60 S-CONTEXT

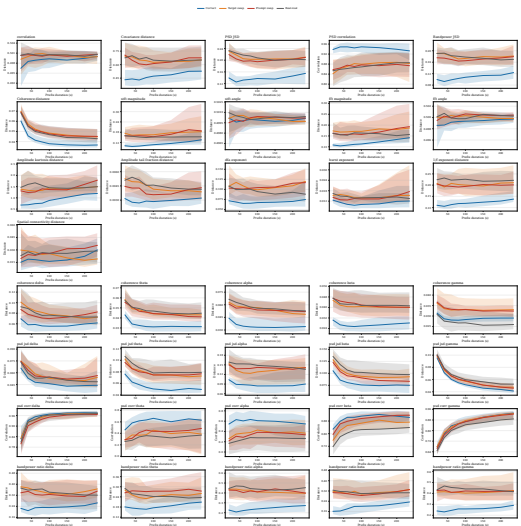


Figure 10: **Auditory (60 s context): full prefix-divergence metrics.**

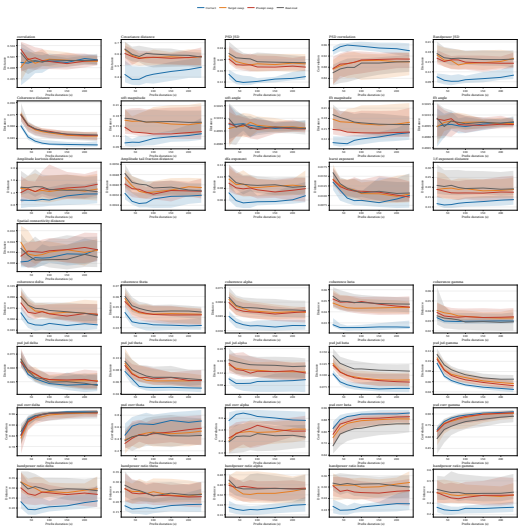


Figure 11: **Visual (60 s context): full prefix-divergence metrics.** Note that correlation and stft/fft angle are expect to have high distance due to phase/dynamics-misalignment between generated and real data.

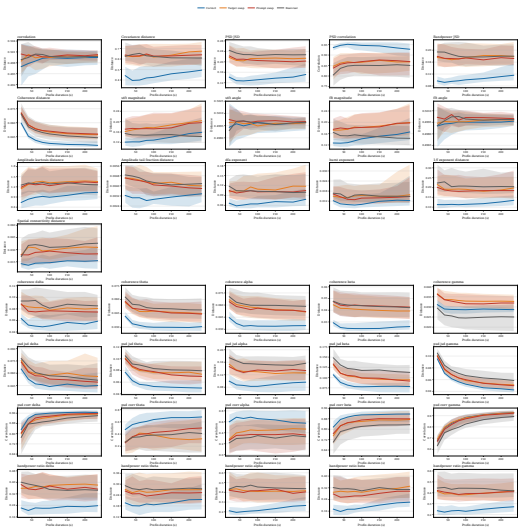
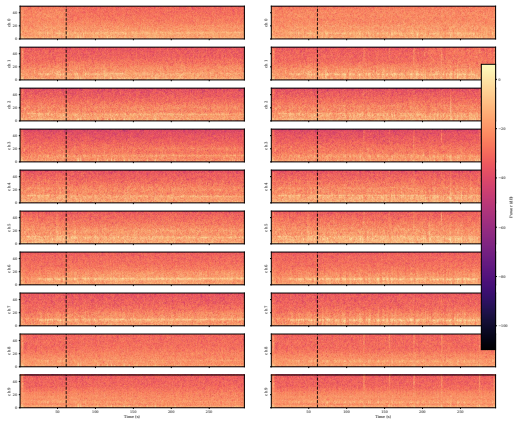


Figure 12: **Rest (60 s context): full prefix-divergence metrics.** Note that correlation and stft/fft angle are expect to have high distance due to phase/dynamics-misalignment between generated and real data.

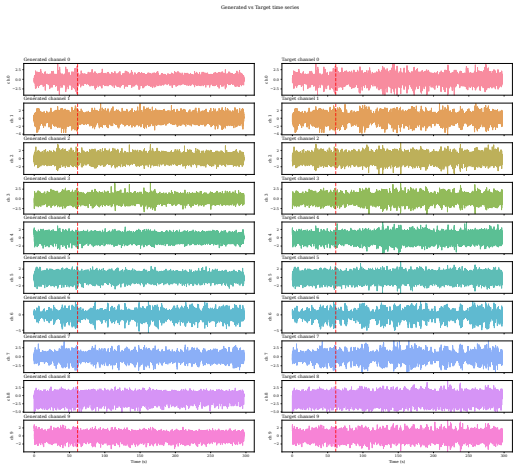
A.11 QUALITATIVE ROLLOUTS

(a) Time series.

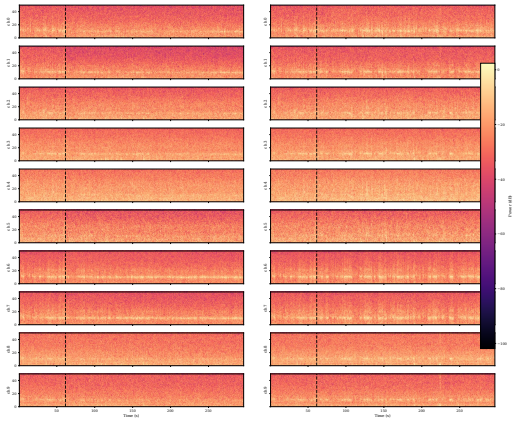


(b) STFT.

Figure 13: **Auditory qualitative rollout.** Dashed lines indicate boundary of context and continuation. 10 random channels are shown due to space constraints.



(a) Time series.



(b) STFT.

Figure 14: **Resting-state qualitative rollout.** Dashed lines indicate boundary of context and continuation. 10 random channels are shown due to space constraints.

A.12 30 S CONTEXT ABLATION

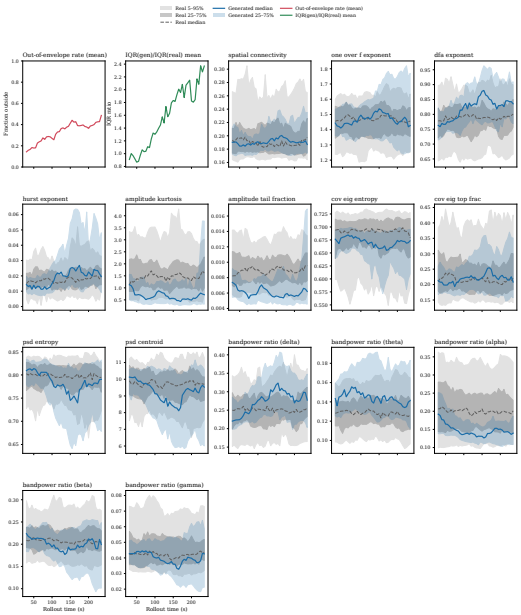


Figure 15: **Auditory (30 s context): sliding-window stability.**

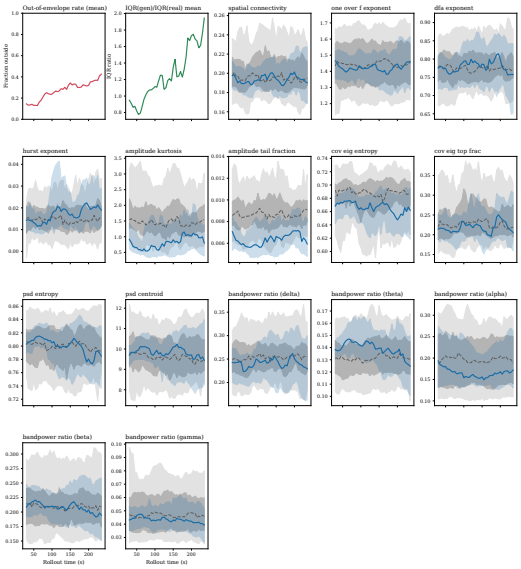


Figure 16: **Visual (30 s context): full sliding-window stability.**

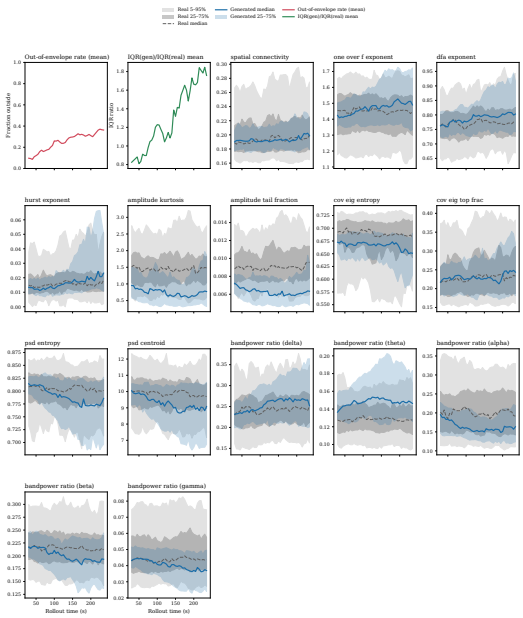


Figure 17: **Rest (30 s context): full sliding-window stability metrics.** Note that correlation and stft/fft angle are expect to have high distance due to phase/dynamics-misalignment between generated and real data.

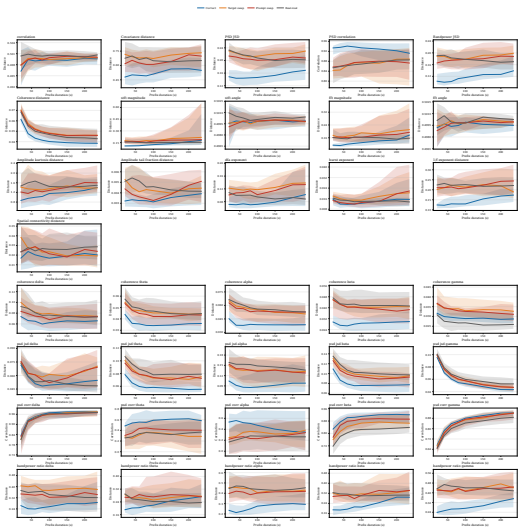


Figure 18: **Auditory (30 s context): full prefix-divergence metrics.** Note that correlation and stft/fft angle are expect to have high distance due to phase/dynamics-misalignment between generated and real data.

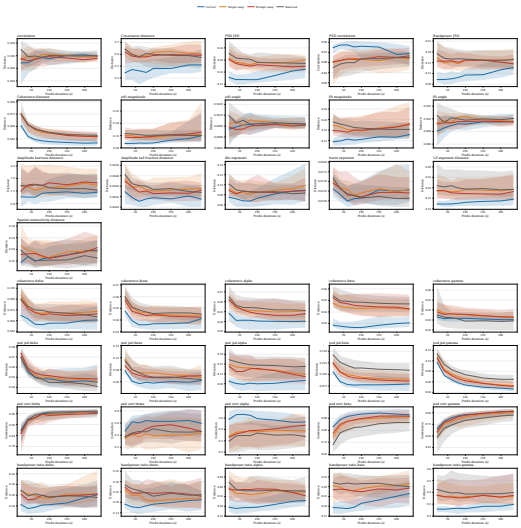


Figure 19: **Visual (30 s context): full prefix-divergence metrics.** Note that correlation and stft/fft angle are expect to have high distance due to phase/dynamics-misalignment between generated and real data.

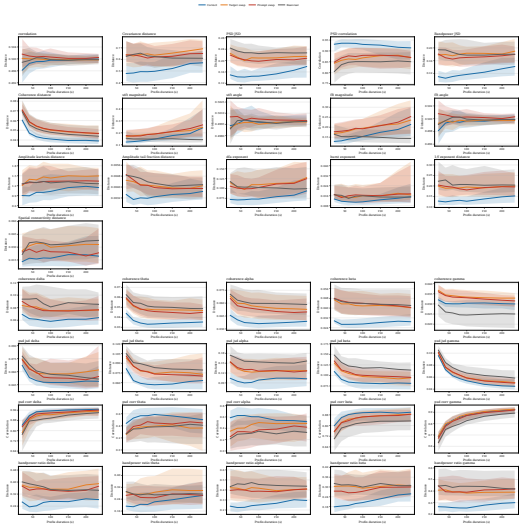


Figure 20: **Rest (30 s context): full prefix-divergence metrics.** Note that correlation and stft/fft angle are expect to have high distance due to phase/dynamics-misalignment between generated and real data.