ZUNA: FLEXIBLE EEG SUPERRESOLUTION WITH POSITION-AWARE DIFFUSION AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce ZUNA, a 380M-parameter diffusion autoencoder with four-dimensional rotary positional embeddings for flexible EEG signal reconstruction and superresolution. We train two variants at scale: ZUNA1, which uses a fixed, whole-channel dropout scheme, and ZUNA1.1, which introduces a new set of masking strategies designed to reflect realistic EEG corruption and missing-data patterns beyond fixed patch reconstruction. ZUNA1.1 is capable of reconstructing variable length sequences of up to 30 seconds, with an arbitrary number of EEG channels at arbitrary scalp locations, and can reconstruct arbitrary temporal intervals within channels in addition to reconstructing entire channels. Among open-source models, ZUNA shows state-of-the-art reconstruction results across a range of dropout tasks.

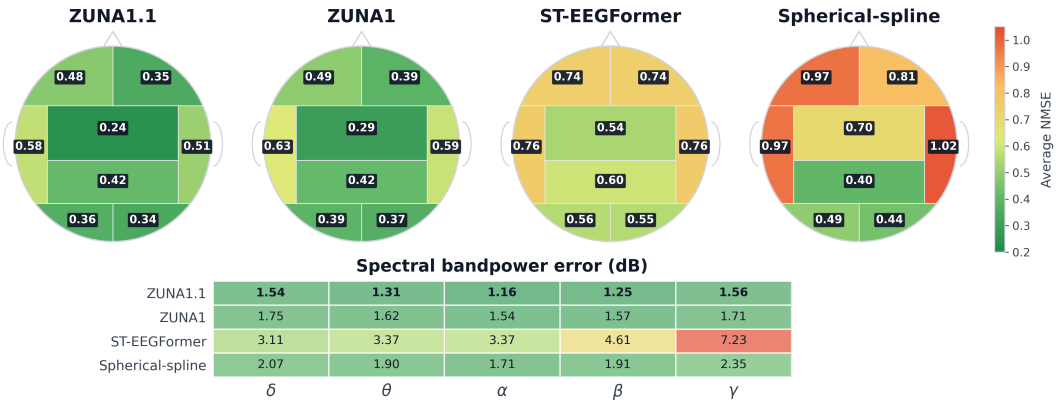
1 INTRODUCTION

Electroencephalography (EEG) is a non-invasive modality for measuring brain activity, valued in clinical, research, and consumer settings for its simple setup and low cost. Although the skull blurs signal at each electrode over an average of a billion neurons, scalp recordings still carry substantial decodable information. Prior work has classified sleep stages (Supratak et al., 2017; Li et al., 2022), decoded emotional states (Liang et al., 2019; Duan et al., 2020), and tracked attentional focus (Su et al., 2022; Kaushik et al., 2022). This success has driven adoption beyond the lab. A growing range of consumer-grade headsets now makes EEG affordable and easy to deploy outside research settings. However, consumer and inexpensive lab-grade headsets often have a much smaller number of channels and poorer data quality than full lab-grade devices. Moreover, all EEG data collection methods often suffer from occasional corruption, artifacts, and general low quality recordings. For instance, electrodes can lose contact with the skin or motion artifacts can disrupt entire groups of channels at a time.

The denoising and superresolution task presents several core challenges. First, anticipating the myriad cases of corruption requires task-specific knowledge of how EEG is deployed in the real world. For example, while patch-level reconstruction is common in EEG masked reconstruction, there are often time segments where the entire headset may be momentarily corrupted. These design decisons entail real tradeoffs for both denoising and use of latent representations for downstream tasks. Many previous methods have attempted to balance both tasks simultaneously, while we show through ZUNA that a surprising tension can arise between the two tasks.

A number of other EEG foundation models (Kuruppu et al., 2025) have been developed in recent years. Contrastive models such as BENDR (Kostas et al., 2021) or SleepFM (Thapa et al., 2024) and a large family of masked-reconstruction models including BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024b), CBraMod (Wang et al., 2024b), Brant (Zhang et al., 2023), BrainWave (Yuan et al., 2024), LUNA (Doner et al., 2025), and REVE (El Ouahidi et al., 2025) are the dominate families. Autoregressive (Jiang et al., 2024a; Cui et al., 2024; Wang et al., 2024a) and latent-predictive (JEPA) (Panchavati et al., 2026) approaches are also increasingly common. All of these models are optimized to learn a general-purpose latent space which can then be probed for downstream *classification*, whereas ZUNA's training objective is the signal *reconstruction* itself. This makes ZUNA suited to infilling missing or corrupted EEG, the problem we target, rather than producing features that a separate model must then decode. While some models like Spatio-Temporal Adaptive Diffusion (STAD) (Wang et al., 2025) also target superresolution, open-source code is suprisingly

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.



Spectral bandpower error (dB)

	δ	θ	α	β	γ
ZUNA1.1	**1.54**	**1.31**	**1.16**	**1.25**	**1.56**
ZUNA1	1.75	1.62	1.54	1.57	1.71
ST-EEGFormer	3.11	3.37	3.37	4.61	7.23
Spherical-spline	2.07	1.90	1.71	1.91	2.35

Figure 1. **Region-occlusion reconstruction metrics.** Top: average NMSE by scalp region for ZUNA1.1, ZUNA1, ST-EEGFormer (Yang et al., 2026), and spherical-spline interpolation, averaged over four held-out datasets. All electrodes in one region are deleted and reconstructed from the remaining seven regions. Bottom: mean absolute spectral bandpower reconstruction error in dB, averaged across datasets from Table 5. Lower is better for both panels.

hard to find for this problem. In practice, researchers often default to MNE-Python's (Gramfort et al., 2013) implementation of spherical-spline interpolation (Perrin et al., 1989) out of convenience and flexibility.

Of the methods that do release their code (Kalita et al., 2024; Saba-Sadiya et al., 2020), almost all operate on fixed-channel layouts, limiting hardware compatibility. Moreover, consumer headsets impose specific sparse layouts of channels that tend to fail in spatially or temporally correlated bursts.

To address these limitations, we introduce ZUNA, a 380M-parameter masked diffusion autoencoder that repairs corrupted recordings. We train two variants at scale: ZUNA1, which has a typical, more simple masked-channel dropout scheme, and ZUNA1.1, which has a far more flexible, generalized set of dropout actions that capture realistic, difficult situations where interpolation is necessary. We do not claim that the dropout schemes are solely responsible for any performance gains, however, as ZUNA1.1 introduces many other engineering choices to improve results. Generalized preprocessing allows us to increase our training corpus from 2 to 3.5 million channel-hours. Our main contributions are:

- *Hardware-agnostic denoising and superresolution*: ZUNA tokenizes multichannel recordings into short temporal windows and injects spatiotemporal structure via 4D rotary positional encoding over (x, y, z, t), enabling inference on arbitrary channel subsets and positions, making it a rare hardware-agnostic foundation model for interpolation and superresolution.
- *Diverse dropout mixture*: training uses a mixture of eight different dropout schemes spanning uncorrelated, contiguous, anatomical, montage-density, and consumer-headset patterns, broadening the range of realistic infilling scenarios the model handles without retraining.

2 Architecture

Prior works have applied deep learning to signal interpolation and superresolution. Wang et al. (2025) use a diffusion autoencoder with a convolutional encoder and transformer decoder. Saba-Sadiya et al. (2020) use a U-Net architecture with a fixed 10-20 montage channel array. Li et al. (2025) use a masked transformer autoencoder with a fixed-masking strategy.

These works have assumed a fixed channel layout, masking strategy, and even a specific fixed upsampling ratio (e.g. 24 to 48 channels with the same channels upsampled each time). Instead, we aim to interpolate any number and location of dropped-out channels. We utilize a novel 4-dimensional rotary positional embedding (RoPE) scheme where we represented the spatial and coarse-time channel positions (Heo et al., 2024). This allows ZUNA to generalize across channel montages and even interpolate entirely novel channel locations which were not recorded for a given signal.

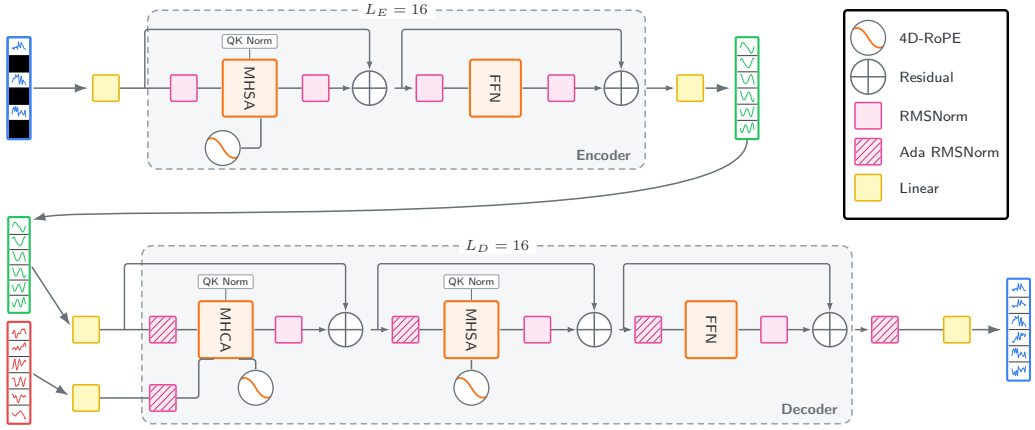


Figure 2. **ZUNA1.1 architecture.** A transformer-based diffusion autoencoder with multi-head self-attention (MHSA) and cross-attention (MHCA) between noisy signal and the latents. Masked data is reconstructed in the decoder by cross-attention between the latent representation and noisy EEG signal. ZUNA has approximately 380M parameters.

ZUNA is a transformer diffusion autoencoder for masked channel reconstruction of EEG signals. The model operates on a tokenized representation of multichannel EEG 2D structure (channel by time) while remaining compatible with standard 1D transformer processing. ZUNA introduces two architectural innovations:

1. **Rasterized channel–time token ordering.** Each EEG channel is segmented into fixed-length windows (0.125 s, 32 samples) and each window is represented as a continuous-valued token. Tokens are computed from the raw time series through a token-encoder MLP. Tokens are then serialized in a raster-scan order: all channels for coarse time step 1, then all channels for coarse time step 2, etc. A 5 second sample with C channels contains $40C$ tokens.
2. **4D rotary positional encoding (4D RoPE).** Each token is associated with a 4D coordinate consisting of the channel's 3D scalp location and the token's coarse-time index. We inject this structure using a multi-dimensional RoPE variant following (Heo et al., 2024).

Let the continuous EEG be $X \in \mathbb{R}^{C\times T}$, where C is the number of channels. We segment each channel into non-overlapping (or fixed-stride) windows of length $\tau = 0.125$ s, yielding M coarse windows per channel. Each channel–window pair is converted into a feature vector

$$\mathbf{s}_{c,m} = \text{TokenEncoder}(X_{c,m}) \tag{1}$$

We then serialize the 2D grid (c, m) into a single 1D sequence of length $L = C \cdot M$ using a raster ordering:

$$\big[\mathbf{s}_{1,1}, \mathbf{s}_{2,1}, \ldots, \mathbf{s}_{C,1},\ \mathbf{s}_{1,2}, \ldots, \mathbf{s}_{C,2},\ \ldots,\ \mathbf{s}_{C,M}\big]. \tag{2}$$

This ordering is chosen so that (i) adjacent tokens share the same coarse-time index but different channels, and (ii) tokens at a fixed channel recur periodically every C positions.

Each token i corresponds to a channel $c(i)$ and coarse time index $m(i)$. For positional embeddings, we associate a 4D discrete coordinate

$$\mathbf{p}_i = \big(x_{c(i)}, y_{c(i)}, z_{c(i)}, m(i)\big), \tag{3}$$

where (x_c, y_c, z_c) is the channel position on the head and m is the coarse-time (segment) index. In ZUNA1, spatial coordinates were discretized into 50 bins; ZUNA1.1 retains this 4D-RoPE construction and increases spatial discretization to 100 bins per axis. We then apply rotary positional embeddings by passing the 4D index $\mathbf{p}_i$ into the rotary embedding module.

For the encoder and decoder models we use 16 transformer layers consisting of the standard self-attention plus MLP. Attention is bidirectional and we diffuse the entire set of channels at once during

generation. To inject the encoder's latent information into the decoder we use adaptive-RMS norm, conditioned on the diffusion time step (Crowson et al., 2024).

The training objective asks the decoder to reconstruct masked channel-time EEG tokens from noise using a rectified-flow loss (Liu et al., 2022) on the decoder outputs, while the encoder latent is regularized with an auxiliary MMD loss (Dziugaite et al., 2015). ZUNA1 used adaptive loss-weighting (Geng et al., 2024) on the decoder outputs; in ZUNA1.1, the stability improvements below allowed us to remove this term without reintroducing training instabilities.

Within the encoder latent, following (Darcet et al., 2024; Jebraeeli et al., 2025), we interleaved learnable register tokens which were inserted at a fixed stride. For a downsampling factor d, the token sequence is partitioned into groups of length d, and a learned register vector r is prepended to each group:

$$\big[\mathbf{r}, \mathbf{h}_1, \ldots, \mathbf{h}_d,\ \mathbf{r}, \mathbf{h}_{d+1}, \ldots, \mathbf{h}_{2d},\ \ldots\big]. \tag{4}$$

This produces an interleaved sequence of length $L' \approx L + L/d$. In this work, we used a downsampling factor $d = 1$. We set the sliding-window mask used in FlexAttention (Dong et al., 2024) to 65536 tokens, a value large enough to include all tokens from a 5 second sample with 256 channels even after interleaving registers with $d = 1$.

To improve training stability from ZUNA1 to ZUNA1.1, we added sandwich normalization (Ding et al., 2021) and Query-Key normalization (Henry et al., 2020) to each layer in the encoder and decoder, shown in Figure 2. These architectural improvements allowed us to remove adaptive loss weighting (Geng et al., 2024) from the decoder RF-loss without reintroducing training instabilities.

3 Method

3.1 Variable-length sequence training

In ZUNA1, every training sample was a fixed five-second segment. ZUNA1.1 instead samples a window length w uniformly at random per training example, snapped to a multiple of the $\tau = 0.125\,\text{s}$ coarse-time unit. Window lengths are drawn from a piecewise-uniform distribution over four bins: very short ($0.5 - 1$ seconds), short ($1 - 5$ seconds), medium ($5 - 10$ seconds), and long ($10 - 30$ seconds) with weights 0.2, 0.3, 0.3, and 0.2, respectively.

This distribution focuses on the 1–10 s range, which we found to be the most common operating regime in downstream applications, while preserving meaningful coverage of extrema.

A consequence of variable-length training is that the per-sample token count varies dramatically. Random-length windows are packed contiguously into a batch and padded to fit a target packed sequence length. FlexAttention is then used to compute a sample-aware attention mask that prevents tokens from one sample attending to tokens of another. This approach allows a compiled model to efficiently process a large batch of tokens from multiple samples in parallel, filling GPU memory and maximizing utilization.

3.2 Quality-aware preprocessing

ZUNA1's preprocessing pipeline is common to the EEG foundation model literature: channels which exceeded noise thresholds for any portion of a recording was zeroed for the whole recording, and epochs with too many bad channels were discarded entirely. However, ZUNA1.1's pipeline stores the full continuous recording on disk and computes a per-channel, per-second *quality matrix* $Q \in [0, 1]^{C \times S}$, where C is the number of channels and S is the number of one-second segments in the recording. The quality score combines three markers, each comparing a segment against the statistics of its own channel (or, for flatness q^{flat}, of the whole recording). Let $\sigma_{c,s}$ denote the standard deviation of channel c in segment s, $\tilde{\sigma}_c$ the median of $\sigma_{c,s}$ over segments, and $\rho_{c,s}$ the segment's

peak-to-peak amplitude relative to its channel-median peak-to-peak. The three scores are

$$q_{c,s}^{\text{var}} = \exp\left(-\tfrac{1}{2}\left(\frac{\log(\sigma_{c,s}/\tilde{\sigma}_c)}{0.7}\right)^2\right),$$
$$q_{c,s}^{\text{ptp}} = \left(1 + e^{\,2(\rho_{c,s}-3.5)}\right)^{-1},$$
$$q_{c,s}^{\text{flat}} = \mathbf{1}[\sigma_{c,s} \geq 0.05 \cdot \text{median}_c\, \tilde{\sigma}_c],$$

and the final quality is their element-wise minimum, $Q_{c,s} = \min(q_{c,s}^{\text{var}}, q_{c,s}^{\text{ptp}}, q_{c,s}^{\text{flat}})$. The variance score penalizes both noise bursts and flat stretches (deviations of the segment's variance from the channel's typical variance, measured in log-space); the peak-to-peak score falls off steeply for large transient artifacts; and the flatness score hard-zeroes dead channels, including reference channels that are flat by construction. We include more details in Appendix B.1 that describe how the quality threshold is tuned and how the data is loaded during training.

3.3 Multiple filter variants

EEG analyses use a wide variety of upstream filtering conventions, and there is no universally agreed default. To allow ZUNA1.1 to interoperate with users' existing pipelines, we precompute and save two filter variants of every recording:

- **Notch variant**: A 0.01 Hz highpass plus zero-phase notch filters at powerline frequencies and their harmonics (restricted to frequencies below Nyquist). Since our corpus mixes recording sites on 50 Hz and 60 Hz mains power, both families are removed. This variant preserves the broadband signal, including slow drifts and high-frequency content, apart from line noise.
- **Bandpass variant**: A 0.1–45 Hz bandpass filter applied independently to the raw recording. Suitable for users whose downstream analyses operate in the conventional clinical EEG band.

Both variants are sampled with equal probability during training. Users can select a variant at inference time to match their preferred preprocessing convention. Filtering is applied with EEGLAB-compatible boundary handling so that annotations from re-concatenated cleaned epochs do not introduce filter ringing across discontinuities.

3.4 Diverse channel dropout mixture

Table 1. Channel-dropout schemes and their sampling weights across the three training stages. Stage 1 (0–250k steps) samples all eight schemes uniformly; Stage 2 (250–440k steps) samples all eight schemes with the (non-uniform) weights shown; Stage 3 (440–580k steps) drops the four *layout* schemes (bottom four, whose loss had plateaued) and samples the four *structure* schemes (top four) uniformly. A schematic of the four structure schemes appears in Appendix B, Figure 6.

Scheme	Pattern	Stage 1	Stage 2	Stage 3	Description
`full-channel-random`	spatial, uncorrelated	0.125	0.275	0.25	Drops a random subset of channels for the entire sample window.
`random-uniform`	unstructured	0.125	0.075	0.25	Drops individual tokens independently.
`full-time-pt-random`	temporal	0.125	0.125	0.25	Drops disjoint contiguous blocks of coarse-time steps across all channels
`correlated-channel-time`	spatio-temporal	0.125	0.125	0.25	Each dropped time contiguous block is removed from a random subset of channels: spatially and temporally clustered dropout.
`standard-montage`	montage density	0.125	0.125	—	Keeps only channels nearest to a standard 8/16/32/64-channel 10–20 montage.
`random-montage`	montage density	0.125	0.125	—	Greedily removes one channel of the closest remaining electrode pair down to a random count (8/16/32/64): sparse, uniform coverage.
`brain-region`	anatomical	0.125	0.125	—	Partitions channels into eight scalp regions by 3D coordinates and drops all channels in a random subset of regions.
`consumer-eeg`	device layout	0.125	0.025	—	Keeps only channels nearest the electrode positions of a randomly selected consumer headset template.

ZUNA1, like most infilling models, is trained exclusively with a single dropout scheme, randomly removing whole channels with uniform probability. However, patterns of degradation are more complex in practice. For ZUNA1.1, we introduce a diverse mixture of eight dropout schemes which fall into two families. Four *structure* schemes control the pattern of dropout, from fully unstructured token dropout to spatio-temporally correlated dropout that mimics motion artifacts and transient electrode failures. Four *layout* schemes control which channels are retained. The montage and consumer-headset schemes select channels by matching target electrode positions to the nearest recorded electrode, while the brain-region scheme partitions the recorded electrodes by scalp coordinates. Together, the eight schemes expose the model to a wide range of infilling scenarios we target at deployment: upsampling a sparse consumer headset to a dense research montage, repairing an anatomically-clustered block of bad channels, and filling temporal gaps from transmission dropouts. Because the layout schemes proved hardest and appeared to quickly saturate at a high loss, we concentrated the final training stretch on the four structure schemes which were amenable to further progress.

Driven by the observation that different dropout schemes have different difficulties, as shown in Appendix B, Figure 7, we used a three-stage curriculum. For the first half of training, we sampled all eight schemes using uniform weights. From 250k - 440k steps, observing that evaluation metrics had plateaued for many of the dropout schemes because they were too hard or already learned, we changed the parameters of individual dropout schemes to adjust their difficulty level and adjusted the weight of each scheme as reflected in Stage 2. This included increasing overall dropout probability from 0.9 to 0.99 because the model had perfectly learned the easy task of reconstructing non-dropped tokens. Finally, in Stage 3, from 440k - 580k steps, we retained only the 4 structure dropout schemes at uniform weighting.

We found the four layout schemes were substantially more challenging for the model to learn than the structural schemes. For instance, the consumer-headset scheme can require reconstructing a full 256-channel montage from as few as four electrodes.

4 Results

4.1 Reconstruction

We evaluate our models and dropout schemes across an array of reconstruction tasks. Some models like ST-EEGFormer can handle full channel or patch-level reconstruction. Other models like CBraMod or LUNA only operate at a patch-level and are thus omitted from full channel dropout tasks.

For patch-level reconstruction, we chose four datasets that represent an array of tasks and channel counts and evaluate under random-uniform dropout, the training scheme common to all models. ZUNA1.1 is the most faithful to both amplitude and spectral signatures, achieving the lowest average NMSE and band-pass error across all datasets.

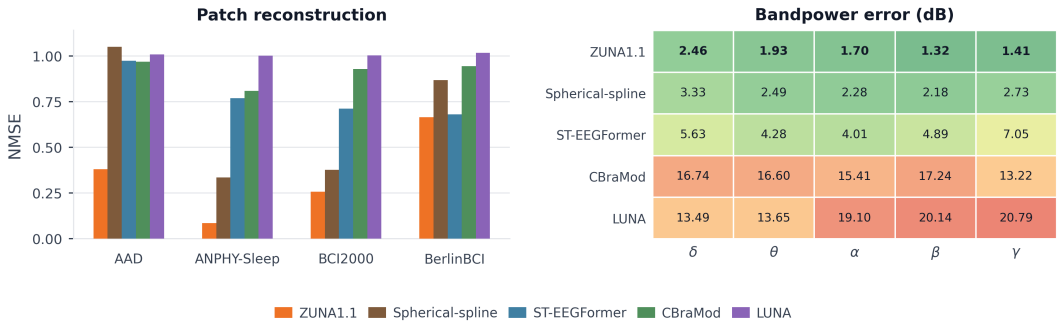


Figure 3. **Patch-level reconstruction benchmark.** Left: average NMSE under 50% random-uniform patch dropout across randomly sampled subjects. Right: mean absolute spectral bandpower error in dB, averaged across the four datasets. Lower is better for both panels.

We also tested the sensitivity of the reconstruction methods with respect to the proportion of deleted channels, progressively deleting more channels and giving the models less information for reconstruction. In Figure 4, deleted channels are chosen uniformly at random, and are deleted for all time in the sample. ST-EEGFormer is competitive on BerlinBCI where the number of channels is small (32 channels), but has higher error on datasets like AAD (256 channels) with more electrodes, reflecting the lack of montage diversity in its training corpus.

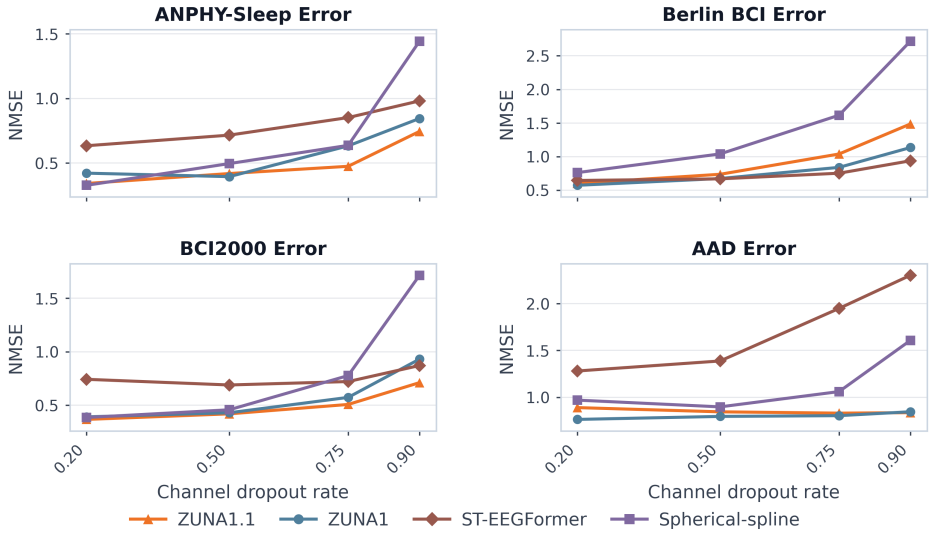


Figure 4. **Reconstruction accuracy as channels drop.** Reconstruction NMSE versus channel-dropout rate on four held-out datasets; lower is better. ZUNA1.1 remains comparable to ZUNA1 on this original channel-dropout benchmark despite its wider range of dropout schemes, while ST-EEGFormer is competitive on some datasets but less stable across the full evaluation set as channel counts increase. Evaluation restricted to 5 s samples for comparability with ZUNA1.

Since spherical spline interpolation heavily leverages proximal channels to predict missing signals, the method can maintain reasonable performance predicting channels when signal is retained in nearby channels. To stress test the channel prediction methods, we stratified channels into eight regions and dropped all channels from each region, predicting them from the remaining seven regions. Figure 1 illustrates that ZUNA1.1 achieves meaningful, experimentally-realistic improvements over established methods under region occlusion. Figure 5 provides a random sample of a masked BCI2000 token.

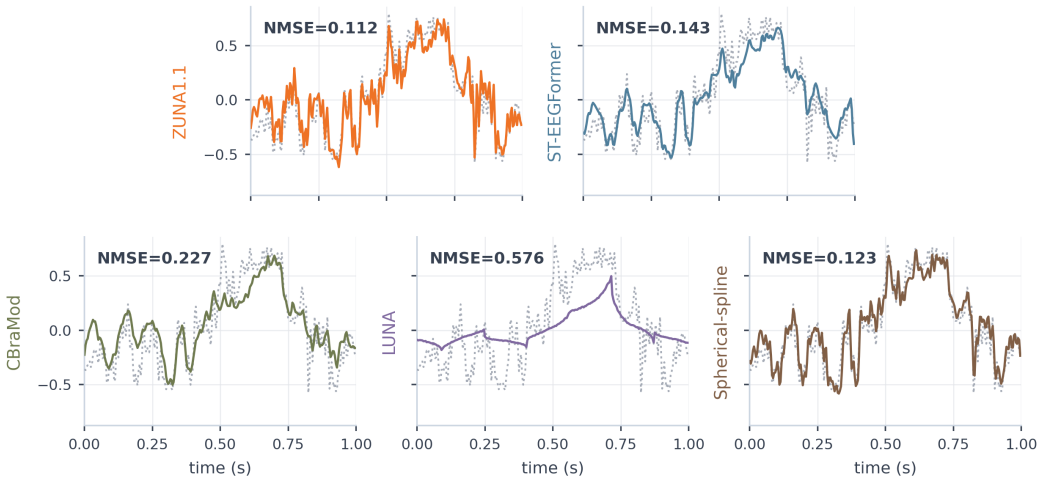


Figure 5. **Example masked-channel reconstruction.** A random BCI2000 patch segment under 50% dropout with the exact same physical time intervals removed for every model.

4.2 Downstream Tasks

We probed ZUNA1.1's latent representation with a linear decoder that averages over tokens and finetunes the encoder using NeuralBench (Banville et al., 2026). Frozen backbones with linear heads performed significantly worse than finetuned models for ZUNA1.1, consistent with NeuroAtlas (Kontras et al., 2026) and EEG-FM-Bench (Xiong et al., 2026). Many of our frozen results were slightly above random chance, adding more evidence that EEG foundation model features are either typically nonlinear or otherwise hard to disentagle.

In particular, ZUNA seems to support the idea of a tension between masked autoencoders and downstream representations. El Ouahidi et al. (2025) argues that the final layer of the encoder in a masked autoencoder is prone to overfitting to the reconstruction task, harming representations for downstream adaptation in EEG. However, they provide no ablations for their regularization that attempts to mitigate this issue. We provide experiments to show how this tension manifests in classification and reconstruction evaluations.

We find a counterintuitive relationship between classification accuracy and reconstruction error in ZUNA1.1: as the number of training steps increases, reconstruction accuracy typically *increases* while classification accuracy typically *decreases*. The full ablation tables in Appendix G, Tables 8 and 9, show a uniform trend on the NeuralBench classification datasets and across different regularization schemes. None of the top-performing downstream task checkpoints occur after 50k steps, while none of the top-performing reconstruction task checkpoints occur before 400k steps. Figure 7 of Appendix B shows slight learning in certain dropout schemes after 400K steps, potentially indicating very brittle downstream features. If REVE were trained longer with more data, would its classification accuracy also decrease given its objective is entirely reconstructive? Without these ablations, the answer is inconclusive. However, this seems to be the case for ZUNA1.1.

On the NeuralBench datasets evaluated using the best performing checkpoint step, ZUNA1.1 achieves second place by mean accuracy and third place by mean normalized rank. However, this occurs early in training, as our final checkpoint drops to fifth and seventh place by each metric, respectively. This degradation casts serious doubts on our model as a generalized representation. ZUNA is specifically designed and trained for interpolation and superresolution.

Furthermore, we also find no improvement with model size on classification performance when scaling ZUNA1.1 up to 1.2B parameters, which performed worse than the current 380M model on downstream tasks for both finetuning and frozen probes. Our experiments and the NeuralBench results suggest that simply adding parameters is not a panacea to representation learning.

Table 2. Classification accuracies across tasks. The final row reports mean accuracy across the listed datasets, with rank by mean accuracy in parentheses.

Dataset	ZUNA1.1	BENDR	BIOT	CBraMod	LaBraM	LUNA	REVE	Dummy	Handcrafted
Audiovisual	0.400±0.030	0.419±0.020	0.261±0.023	0.311±0.076	0.366±0.095	0.262±0.016	0.505±0.010	0.250±0.000	**0.623±0.000**
CVEP	**0.880±0.090**	0.188±0.240	0.047±0.003	0.871±0.013	0.876±0.005	0.872±0.009	0.791±0.008	0.050±0.000	0.048±0.000
Depression	**0.910±0.010**	0.857±0.019	0.867±0.018	0.873±0.006	0.867±0.038	0.844±0.003	0.870±0.006	0.500±0.000	0.813±0.001
ERN	0.810±0.030	0.768±0.046	0.464±0.036	0.811±0.015	**0.833±0.008**	0.819±0.011	0.814±0.033	0.500±0.001	0.775±0.000
LRP	0.780±0.010	0.705±0.004	0.502±0.004	0.780±0.007	0.789±0.010	0.747±0.006	**0.792±0.006**	0.500±0.001	0.652±0.001
Mental Arithmetic	0.680±0.020	0.681±0.012	0.640±0.023	0.685±0.021	0.687±0.075	0.643±0.069	**0.739±0.027**	0.500±0.000	0.698±0.000
Mental imagery	0.250±0.010	0.209±0.011	**0.275±0.017**	0.202±0.003	0.218±0.008	0.209±0.014	0.256±0.021	0.200±0.000	0.263±0.000
N170	0.720±0.010	0.662±0.009	0.504±0.005	0.707±0.008	0.728±0.003	0.693±0.006	**0.755±0.013**	0.500±0.000	0.663±0.000
N2PC	0.620±0.020	0.533±0.030	0.503±0.003	0.616±0.006	0.631±0.022	0.607±0.015	**0.653±0.003**	0.500±0.000	0.507±0.000
N400	0.610±0.020	0.616±0.003	0.506±0.009	0.611±0.022	0.639±0.007	0.648±0.012	0.645±0.011	0.500±0.000	**0.661±0.000**
Average	0.666 (2)	0.564 (7)	0.457 (8)	0.647 (4)	0.663 (3)	0.634 (5)	**0.682 (1)**	0.400 (9)	0.570 (6)

We also ran a series of layer-wise probes to test if different features were being represented at different layers of our encoder, a common probe in speech and NLP benchmarks such as SUPERB (Yang et al., 2021) and GLUE (Wang et al., 2018). However, these probes (shown in Appendix G, Figure 12) show no obvious trends and require deeper exploration, perhaps across many more models. We do not find uniform evidence that the final layer of the encoder is typically worse for downstream classification.

5 DISCUSSION

ZUNA is a highly flexible, montage-agnostic diffusion autoencoder for interpolation and superresolution. We introduce rasterized channel–time token ordering, 4D rotary positional encodings, and new dropout training schemes to EEG recordings that significantly expand range and ability to process and reconstruct realistic "in-the-wild" EEG data from a variety of montages and tasks. ZUNA's ability to handle variable length sequences (up to 30 seconds) as well as reconstruct arbitrary time windows within a channel are relatively rare for neural reconstruction methods.

While ZUNA1.1 improves over ZUNA1 in its range and versatility, the two models are roughly comparable on specific tasks for which both models can operate. Since ZUNA1.1 is trained for substantially longer on a much greater quantity of data, this is somewhat surprising. One possibility is that the ZUNA1.1 model is generally superior but that due to having to handle a much wider range of data, the model has to spread its competence over a much wider distribution, and hence necessarily dilute its capabilities in the specific kind of task which the original ZUNA1 model has specialized in. This implies that the model is primarily limited by capacity. However, preliminary experiments with naively scaling up the model architecture from 380M to approximately 1.2B parameters did not show dramatically improved performance versus the smaller model, which led to ZUNA1.1 being trained at the same size as the original ZUNA1 model. This implies that parameter count alone is not the primary bottleneck for the model's capabilities.

Additional possibilities are that the diffusion encoder-decoder architecture of ZUNA1.1 is itself the bottleneck. It is possible that the architectural scheme of passing the core information through the encoder bottleneck in "registers" imposes a heavy constraint on the ultimate reconstruction capabilities of the model, although expanding this bottleneck naively does not immediately result in better reconstructions. The encoder-decoder architecture itself could be a bottleneck. Another possibility is that the diffusion training scheme and especially the chosen noise schedule reduced the potential SNR of training causing an effective plateau in performance.

Finally, it is of course possible, although we think it is unlikely, that even at 380M parameters, the model has already exhausted all of the signal present in the data which could be used for reconstruction. If this were true, it would mean that the irreducible noise floor in the EEG data itself is relatively high while the information redundancy across channels and time is fairly low. Some evidence that indirectly supports this is that we observed a similarly unexpected phenomenon with our initial set of dropout schemes, where a few of them appeared extremely challenging to learn such that the model made effectively little progress over the course of training. These were the intuitively harder dropout schemes such as the consumer or standard montage channel dropout. The lack of improvement on these schemes seems to imply that good reconstructions in these cases may be largely impossible in that the dropout scheme irreversibly removes the information required to make good predictions.

A further point is that during development of ZUNA, we observed a growing dissociation between reconstructive accuracy and the utility of the learned representations for downstream tasks. We discuss these explorations in Section 4.2 and Appendix G. While uncertain, we hypothesize that a substantial fraction of this dissociation may be due to the diffusion objective and architecture, which in other modalities is known to produce less clearly generalizable representations than more standard autoregressive objectives (He et al., 2021). More generally, for the goal of constructing powerful representations for downstream classifiers, pure reconstruction may not be the optimal objective in general, and latent-space objectives such as contrastive (Oord et al., 2018; He et al., 2019; Chen et al., 2020; Radford et al., 2021) and JEPA-like (LeCun et al., 2022; Assran et al., 2023) losses may be preferred. In future work, we aim to focus more on developing general encoder models with representations well-suited to downstream tasks. This is ultimately a necessary step to establish the latent space which can then be joined with pretrained language or image models to train effective multimodal models. This will likely require moving beyond regularized reconstruction as an objective, and developing novel architectures to move beyond ZUNA's diffusion autoencoder base.

AI USE STATEMENT

AI was used in the LaTeX formatting and also for design of the architecture figure. The figure was created with `tikz`, and was not directly generated by an image model. We also used AI to convert OpenNeuro dataset indices to citations for our training corpus, which we then manually verified.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Hubert Banville, Stéphane d'Ascoli, Simon Dahan, Jérémy Rapin, Marlène Careil, Yohann Benchetrit, Jarod Lévy, Saarang Panchavati, Antoine Ratouchniak, Elisa Cascardi, et al. Neuralbench: A unifying framework to benchmark neuroai models. *arXiv preprint arXiv:2605.08495*, 2026.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z. Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers, 2024. URL https://arxiv.org/abs/2401.11605.

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.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. URL https://arxiv.org/abs/2105.13290.

Doner et al. Luna: A latent unified network architecture for topology-agnostic eeg modeling, 2025.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels, 2024. URL https://arxiv.org/abs/2412.05496.

Lijuan Duan, Huifeng Duan, Yuanhua Qiao, Sha Sha, Shunai Qi, Xiaolong Zhang, Juan Huang, Xiaohan Huang, and Changming Wang. Machine learning approaches for mdd detection and emotion decoding using eeg signals. *Frontiers in Human Neuroscience*, 14:284, 2020.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization, 2015. URL https://arxiv.org/abs/1505.03906.

Yassine El Ouahidi et al. Reve: Representation for eeg with versatile embeddings, 2025.

Zhengyang Geng, Ashwini Pokle, William Luo, Justin Lin, and J. Zico Kolter. Consistency models made easy, 2024. URL https://arxiv.org/abs/2406.14548.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013. doi: 10.3389/fnins.2013.00267.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers, 2020. URL https://arxiv.org/abs/2010.04245.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer, 2024. URL https://arxiv.org/abs/2403.13298.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Rongjie Huang, Max W. Y. Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Vahid Jebraeeli, Hamid Krim, and Derya Cansever. Vitcae: Vit-based class-conditioned autoencoder, 2025. URL https://arxiv.org/abs/2509.16554.

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. *arXiv preprint arXiv:2409.00101*, 2024a.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024b.

Bhabesh Kalita, Nabamita Deb, and Daisy Das. AnEEG: leveraging deep learning for effective artifact removal in EEG data. *Scientific Reports*, 14(1):24234, 2024. doi: 10.1038/s41598-024-75091-z. URL https://doi.org/10.1038/s41598-024-75091-z.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Pallavi Kaushik, Amir Moye, Marieke van Vugt, and Partha Pratim Roy. Decoding the cognitive states of attention and distraction in a real-life setting using eeg. *Scientific Reports*, 12(1):20649, 2022.

Konstantinos Kontras, Trui Osselaer, Stylianos G Mouslech, Angeliki-Ilektra Karaiskou, Guido Gagliardi, Thomas Strypsteen, Mohammad Hossein Badiei, Anku Rani, Maarten Vanmarcke, Miguel Bhagubai, et al. Neuroatlas: Benchmarking foundation models for clinical eeg and brain-computer interfaces. *arXiv preprint arXiv:2605.14698*, 2026.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data, 2021.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, Sandipan Pati, Gregory Worrell, and Yogatheesan Varatharajah. Eeg foundation models: A critical review of current progress and future directions. *arXiv preprint arXiv:2507.11783*, 2025.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Chengfan Li, Yueyu Qi, Xuehai Ding, Junjuan Zhao, Tian Sang, and Matthew Lee. A deep learning method approach for sleep stage classification with eeg spectrogram. *International journal of environmental research and public health*, 19(10):6322, 2022.

Dongdong Li, Zhongliang Zeng, Zhe Wang, and Hai Yang. Estformer: Transformer utilising spatiotemporal dependencies for electroencephalogram super-resolution. *Knowledge-Based Systems*, 317:113345, may 2025. ISSN 0950-7051. doi: 10.1016/j.knosys.2025.113345. URL http://dx.doi.org/10.1016/j.knosys.2025.113345.

Zhen Liang, Shigeyuki Oba, and Shin Ishii. An unsupervised eeg decoding system for human emotion recognition. *Neural Networks*, 116:257–268, 2019.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Christopher J. Markiewicz, Krzysztof J. Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O. Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncavles, et al. The openneuro resource for sharing of neuroscience data. *eLife*, 10:e71774, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

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

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

Panchavati et al. Laya: A latent-predictive eeg foundation model, 2026.

F. Perrin, J. Pernier, O. Bertrand, and J.F. Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72(2):184–187, 1989. ISSN 0013-4694. doi: https://doi.org/10.1016/0013-4694(89)90180-6. URL https://www.sciencedirect.com/science/article/pii/0013469489901806.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sari Saba-Sadiya, Tuka Alhanai, Taosheng Liu, and Mohammad M. Ghassemi. Eeg channel interpolation using deep encoder-decoder networks, 2020. URL https://arxiv.org/abs/2009.12244.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Enze Su, Siqi Cai, Longhan Xie, Haizhou Li, and Tanja Schultz. Stanet: A spatiotemporal attention network for decoding auditory spatial attention from eeg. *IEEE Transactions on Biomedical Engineering*, 69(7):2233–2242, 2022.

Akara Supratak, Huaixiu Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

Rahul Thapa et al. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals, 2024.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018. URL https://aclanthology.org/W18-5446/.

Guangyu Wang et al. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *Advances in Neural Information Processing Systems*, 2024a.

Jiquan Wang et al. Cbramod: A criss-cross brain foundation model for eeg decoding, 2024b.

Shuqiang Wang, Tong Zhou, Yanyan Shen, Ye Li, Guoheng Huang, and Yong Hu. Generative ai enables eeg super-resolution via spatio-temporal adaptive diffusion learning, 2025. URL https://arxiv.org/abs/2407.03089.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. Eeg-fm-bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of eeg foundation models, 2026. URL https://arxiv.org/abs/2508.17742.

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

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5Xwm8e6vbh.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal performance benchmark, 2021. URL https://arxiv.org/abs/2105.01051.

Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal foundation model for clinical applications. *arXiv preprint arXiv:2402.10251*, 2024.

Daoze Zhang, Zhizhang Yuan, Yang Yang, et al. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems*, 2023.

A DIFFUSION MODELS

Diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015) are widely used *generative* models for continuous real-valued signals such as images and audio. They work by defining a noising process that gradually steps from clean data to pure noise (Karras et al., 2022; Song et al., 2020). The model is then trained to invert this process, taking a step from noise towards data. Intuitively, diffusion models allow iterative denoising, giving the model many steps to converge to the correct signal, unlike a single step autoregressive model. Where well-behaved noising processes can be defined and inverted, diffusion models perform well, reaching state of the art performance in image and video generation (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Rombach et al., 2022; Ho et al., 2022) as well as performing strongly in audio modeling (Huang et al., 2022). Due to the continuous real-valued nature of EEG signals and the potential of their iterative denoising, for ZUNA we settled on a diffusion approach.

Mathematically diffusion models define a Gaussian forward process that maps data to noise. Let $x_0 \sim p_{\text{data}}(x)$ denote a data point (e.g., a channel of EEG signals $\mathbb{R}^d$). We define the noising process as a Markov chain which adds a small amount of Gaussian noise at each step.

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}) \tag{5}$$

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t;\ \sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\big), \tag{6}$$

where β_t is the noise level of each step. This noise schedule increases from close to 0 to the standard deviation of the data during the noising process. Intuitively, early steps perturb fine details while preserving global structure, and late steps wash out semantics until x_T is nearly isotropic Gaussian noise, making the terminal distribution easy to sample.

Conversely, generation is performed by learning an analogous reverse-time Markov chain

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t) \tag{7}$$

with a simple prior $p(x_T) = \mathcal{N}(0, I)$. Crucially, due to the Gaussian dynamics, the *posterior* $q(x_{t-1} \mid x_t, x_0)$ can be computed analytically. This lets us setup our learning objective as a principled and tractable variational lower bound,

$$\log p_\theta(x_0) \ \geq\ \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log p_\theta(x_{0:T}) - \log q(x_{1:T} \mid x_0)\Big], \tag{8}$$

whose terms decompose across timesteps into KL divergences between Gaussians at each timestep. With an identity covariance the KL divergence simplifies to the standard mean-square-error objective. Classical diffusion models predict the injected noise ε (or equivalently the score) with a neural network $\varepsilon_\theta(x_t, t)$, yielding the simple objective,

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{t\sim\text{Unif}\{1,\ldots,T\},\, x_0}\Big[\|\varepsilon - \varepsilon_\theta(x_t, t)\|_2^2\Big] \tag{9}$$

The key intuition is that learning to denoise at *all* noise levels teaches the model a vector field that points back toward high-density regions of p_{data}. In the continuous-time limit this connects to score-based modeling (Hyvärinen & Dayan, 2005; Vincent, 2011), where the network approximates $\nabla_{x_t} \log q(x_t)$ and sampling follows a learned reverse diffusion (stochastic) or probability flow (deterministic) trajectory. In practice diffusion models are trained by sampling a minibatch of data at different levels of the noise schedule, and minimizing Eqn 9 by gradient descent. While here we present the standard diffusion formulation for intuition, our implementation uses rectified flow (Liu et al., 2022) parameterization for improved stability and sampling efficiency.

Diffusion autoencoders pair a regular diffusion decoder which maps noise to data, with an additional encoder which stores some information about the full signal in a latent bottleneck. This latent bottleneck can then be input into the decoder, typically through cross-attention to aid the denoising process.

B TRAINING HYPERPARAMETERS

B.1 QUALITY-AWARE PREPROCESSING DETAILS

The quality matrix is saved alongside the recording as a separate mmap file, and quality thresholds are applied at *load time*. The data loader takes two thresholds, `min_quality_any` and `min_quality_mean`: a channel is included in the current training window if every one-second segment overlapping the window has quality $\geq$ `min_quality_any` *and* the mean quality across those segments is $\geq$ `min_quality_mean` (set to 0.1 and 0.3 respectively in all reported experiments). This means that a channel which is noisy for part of a recording can still contribute usable training data drawn from the clean portions, and the same on-disk corpus can be re-thresholded for any desired quality stringency without rerunning preprocessing.

The recording is stored as a memory-mapped `float32` array of shape (C, T), where T is the number of samples in the recording at 256 Hz. Pre-epoched recordings (which the original ZUNA1 pipeline could not ingest) are stored as (E, C, T_e) where E is the number of epochs and T_e is the number of samples per epoch.

As in ZUNA1, average referencing is performed after masking, so the missing target cannot leak into the visible channels during z-score normalization.

B.2 TRAINING DETAILS

ZUNA1.1 was trained for 580k steps on approximately 3.5 million channel-hours of processed and cleaned EEG data. Since predicting the non-dropped tokens is a fairly trivial identity operation, which the model learns extremely quickly, we downweight the loss contribution of the non-dropped tokens to 10% of the weight for dropped tokens, concentrating the objective on the infilling task while still lightly supervising the retained signal.. We used the standard weight decay of 0.1 on all layers except for embedding and norm layers.

ZUNA1.1 is trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) with a weight decay of 0.1 applied to the linear, MLP, and attention-projection weights but not to the embedding or normalization parameters. The learning rate is cosine-annealed from a peak of 5×10^{-4} to 9×10^{-5} over 580k steps, following a $1{,}000$-step linear warmup. The model trains in `bfloat16` with normalization layers and the residual stream kept in `float32` for stability, and is compiled end-to-end with `torch.compile`. We additionally maintain an exponential moving average (EMA) version of the model weights with decay 0.9999 per step, which is used for evaluation and inference. We ran data-parallel training across 6 nodes with a per-rank micro-batch of 22,000 packed tokens and 2 gradient-accumulation steps; giving a global batch size of roughly 2M tokens per optimizer step.

Training at this scale on heterogeneous public EEG surfaced stability issues worth recording. Rare extreme samples produced sudden gradient-norm spikes followed by loss excursions, and in a handful of cases NaN losses. We addressed these with three safeguards. First, optimizer updates are skipped entirely whenever a non-finite loss is detected. Secondly, we add an epsilon to the z-score denominator to protect against zero-variance (flat or clipped) channel segments. Thirdly, we conducted a corpus scan and removed a small set of recordings whose data and quality matrices contained NaNs. The removed data comprised approximately 0.3% of our total dataset, and we traced the nans to conversion artifacts in three sleep datasets. With architectural improvements and these mitigations in place, we were able to train ZUNA1.1 for $4\times$ as many steps as ZUNA1 with no signs of instability.

The consumer-eeg dropout scheme samples headset templates from Muse (4), Emotiv Insight (5), Dreem (5), Neurosity Crown (8), Unicorn (8), OpenBCI Cyton (8), Emotiv EPOC (14), and Flex32 (32).

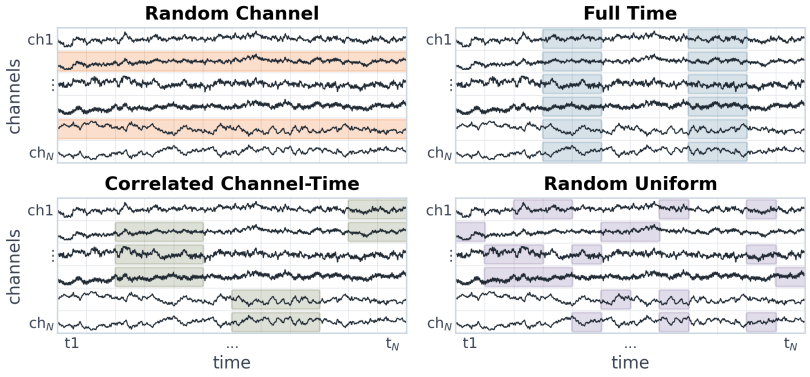


Figure 6. **The four channel-dropout schemes retained throughout training.** Observed tokens are passed to the encoder and the masked spans are set to zero in the encoder. These are the four *structural* dropout schemes retained for the full training run.

Table 3. ZUNA1.1 training configuration.

Hyperparameter	Value
Model dimension	1024
Layers (encoder / decoder)	16 / 16
Head dimension	64
Total parameters	380M
Token size	32 samples (0.125 s at 256 Hz)
RoPE	4D over (x, y, z, t), $\theta = 10^4$
Spatial discretization	100 bins per axis
Max coarse-time steps	240 (30 s)
Optimizer	AdamW
β_1, β_2	0.9, 0.95
Weight decay	0.1
LR schedule	cosine, warmup
Peak learning rate	5e-4
Optimizer steps	580k
Precision	`bf16` (norms, residual in `fp32`)
GPUs	48 (6 nodes $\times$ 8)
Gradient accumulation	2
Packed tokens per rank	22,000
Global batch	$\approx$ 2M tokens / step
Encoder MMD weight	$10^{-2} \longrightarrow 0$
Sample rate	256 Hz
Window lengths	0.5–30 s (Section 3.1)
Quality thresholds (any / mean)	0.1 / 0.3
Filter variants	notch, bandpass (Section 3.3)
Dropout probability p_{drop}	$0.9 \longrightarrow 0.99$
Dropout mixture	Section 3.4
Inference ODE solver	Euler, 50 steps
Initial noise scale	0.1

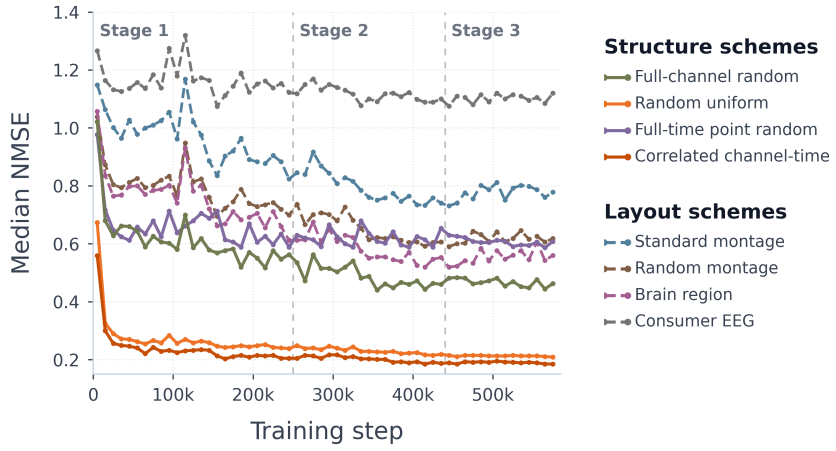


Figure 7. **Reconstruction performance for each dropout scheme throughout training.** Median NMSE (lower is better) on held-out validation OpenNeuro datasets (ds004504 and ds003825, distinct from all test datasets) for each dropout scheme. See Table 1 for details on how the dropout schemes were adjusted during training.

C Reconstruction Performance by Scheme: Sample Duration and Token Dropout Percentage

We evaluate the four structural dropout schemes by: increasing the fraction of dropped tokens at fixed sample duration, and increasing sample duration with token dropout fixed at 15%.

This shows that ZUNA1.1 is fairly robust to sample duration: for three out of the four main structural schemes, average NMSE does not change by more than a few percent.

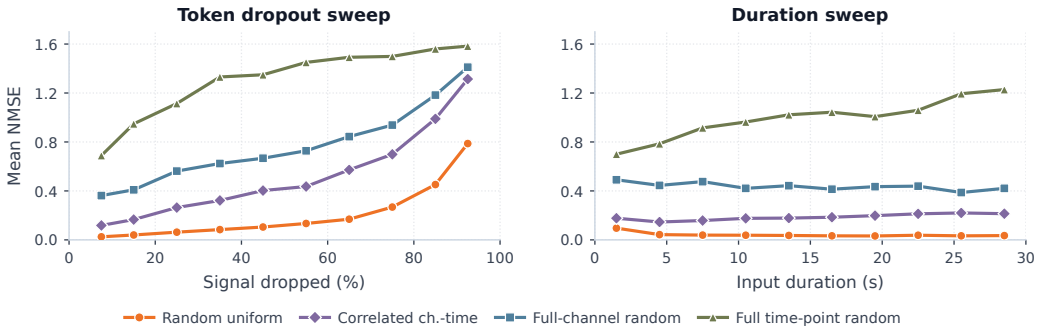


Figure 8. **Reconstruction performance by dropout scheme as token dropout percentage and sample duration vary.** Mean NMSE for each structural dropout scheme. Left: NMSE as a function of dropped-token percentage. Right: NMSE as a function of input duration with token dropout fixed at 15%.

D ORIGINAL ZUNA1 DATA AND TRAINING DETAILS

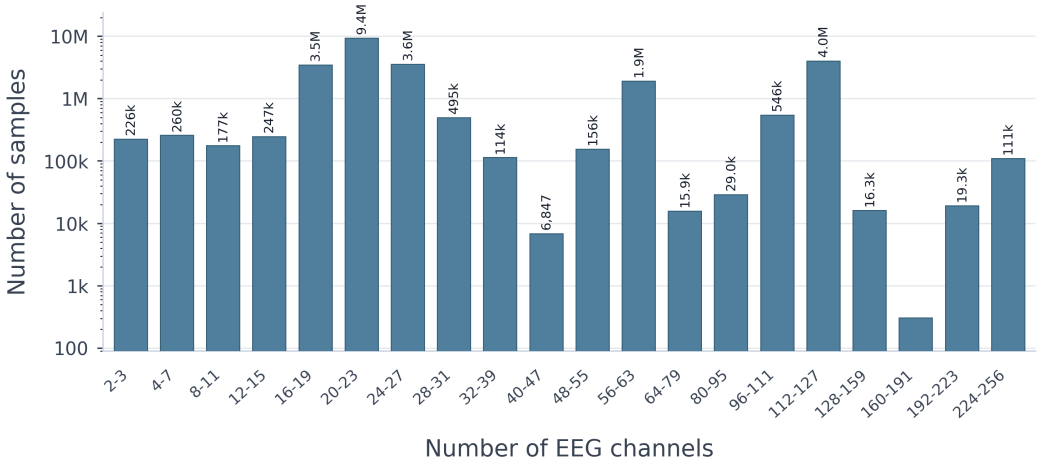


Figure 9. **EEG channel-count distribution in the original ZUNA1 training corpus.** Binned distribution of retained EEG channel counts across all non-overlapping 5-second epochs after preprocessing, comprising approximately 2 million channel-hours with up to 256 channels per sample.

D.1 DATA PREPROCESSING PIPELINE

To train a generalizable EEG foundation model across highly heterogeneous datasets, we constructed a unified preprocessing pipeline that standardizes sampling rate, temporal segmentation, normalization, and spatial metadata while deliberately preserving channel diversity across recordings. All preprocessing was implemented using MNE-Python and applied uniformly across datasets.

We aggregated data from two major open EEG repositories: (i) the Temple University Hospital (TUH) EEG Corpus (Obeid & Picone, 2016), and (ii) a large collection of publicly available datasets hosted on OpenNeuro (Markiewicz et al., 2021). Together, these sources span a wide range of experimental paradigms, recording hardware, channel layouts, and data qualities.

Because ZUNA explicitly conditions on electrode location, only datasets and channels for which valid 3D scalp coordinates could be reconstructed were retained. Channel positions were inferred using a multi-stage strategy combining MNE standard montages (e.g., standard_1005, EGI HydroCel layouts), BioSemi cap metadata, and manual dataset-specific mappings. After this spatial filtering step, the final corpus comprised 214 unique datasets for which we could derive reliable 3D scalp coordinates. Channels without resolvable 3D coordinates were excluded, resulting in the removal of an average of 5.82 channels per recording (median = 6). These primarily corresponded to ocular channels (e.g., EOG) and other auxiliary or non-standard sensors lacking well-defined scalp locations. No fixed channel subset or canonical montage was imposed, yielding a highly heterogeneous distribution of channel configurations across recordings as shown in Figure 9.

All recordings were resampled to a common sampling rate of 256 Hz to ensure consistency across datasets. This included upsampling recordings originally acquired at lower sampling rates. Continuous recordings were segmented into non-overlapping 5-second segments (1280 samples per epoch). Recordings shorter than 10 seconds were discarded. To remove slow drifts, signals were high-pass filtered at 0.5 Hz. Line noise was detected adaptively on a per-recording basis by analyzing the power spectral density (Welch method) between 45 Hz and the Nyquist frequency. Narrowband spectral peaks exceeding a data-driven threshold were identified and notch-filtered, typically capturing 50 Hz, 60 Hz, and their harmonics where present. This adaptive approach avoids hard-coding line-noise frequencies and accommodates international datasets with differing power standards, which was necessary given our highly heterogeneous starting set of datasets.

To mitigate extreme artifacts while preserving maximal coverage, we applied a conservative signal-quality procedure at the channel and epoch level, where an epoch here denotes a non-overlapping 5-second segment of continuous EEG. Channels exhibiting near-flat signals (abnormally low variance relative to the recording-wide distribution, using a robust median/MAD criterion) or evidence of

clipping (0.5% of samples within a small tolerance of the channel minimum/maximum) were flagged as bad and subsequently zeroed out during epoch extraction. After high-pass filtering, re-referencing, and adaptive notch filtering, additional noisy channels (per 5s segment) as well as epochs were identified by excessive standardized variability (SD > 3 per channel or epoch respectively) and removed. When more than 50% of channels per epoch were marked as bad, the full epoch was discarded.

Samples were standardized using z-score normalization based on the mean and standard deviation computed across all observed EEG channels (i.e. channels retained after masking) within each recording, yielding approximately zero-mean, unit-variance signals while preserving relative spatial structure. For long recordings, normalization was performed independently on contiguous 10-minute segments; all subsequent analyses operated on non-overlapping 5-second epochs. Crucially, our average-referencing procedure was done after dropping or masking, preventing information leakage across referencing. No value or statistic derived from a masked target is used to construct the encoder input. The same procedure is used during training and in all reconstruction evaluations.

After spatial filtering and preprocessing, the final training corpus comprised 24,823,808 non-overlapping 5-second EEG epochs from a total of 214 datasets and approximately 2 million channel-hours of EEG recordings. The number of retained EEG channels per recording ranged from 2 to 256 (mean = 45.42; median = 22; SD = 41.50). The channel-count distribution was strongly multimodal (Figure 9), with the majority of recordings containing 16–32 channels and additional pronounced peaks at 64, 128, and 256 channels, reflecting common EEG acquisition systems.

After z-score normalization, signals within each recording were standardized. Each training sample corresponded to one 5-second segment sampled at 256Hz, yielding 1280 individual time points per channel.

D.2 Model training

To enable the model to infill signals on noisy, corrupted channels or to upsample to unobserved channels, we trained with a heavy channel dropout scheme. We set the probability of dropout to be 90%. If dropout was selected, we randomly selected between 1 and $C/2$ channels to be dropped with 80% probability or selected between $C/2$ and $C-1$ channels to be dropped with 20% probability. During training, dropped out channels were replaced with zeros in the encoder input and targets (the noising direction) were sent to decoder input. During inference, the encoder receives zero inputs for dropped channels and the decoder only receives the random Gaussian initialization for both dropped and non-dropped channels. When we drop a channel we do not drop the channel's $\{x, y, z, t_c\}$ position and present this information to both encoder and decoder via 4D-RoPE.

We trained with rectified flow loss (Liu et al., 2022) and adaptive loss-weighting (Geng et al., 2024) on the decoder outputs and regularized the encoder latent with an auxiliary MMD loss (Dziugaite et al., 2015). The model was trained for 150000 steps with a learning rate of 1e-4 annealing down to 1e-6 with a cosine schedule. We used the AdamW optimizer with (β_1,β_2) = (0.9, 0.95). For training stability, we found that the dataset scale was incredibly important, and required us to adjust the dataset to have a standard deviation of 0.1 with mean 0, and to adjust the noise variance of the diffusion process appropriately. Similarly, we observed that the global batch size also greatly impacted stability of training and required us to train with larger batches than expected. We trained with 6 gradient accumulation steps, yielding 90000 tokens per microbatch and 2.16M tokens per global batch across 3 nodes with 24 GPUs. Because of the highly variable numbers of channels per sample – see Figure 9 – naively padding the samples up to the maximum length would be extremely inefficient. To train efficiently, we implemented packing with sample-masking using FlexAttention to dynamically compute a special attention mask for each packed sample. Together, these improvements allowed us to train stably with high utilization at a rate of 50 samples per microbatch and 1150 samples per global batch. Training of the final model took approximately a week on three H100 nodes. Since the model size was so small, only data parallelism was used with large microbatch sizes.

The spatiotemporal structure of EEG signals requires some modifications to the standard transformer architecture. Specifically, the transformer is only specifically designed to natively handle 1D sequences of elements while the EEG signal has C channels in parallel. Our novel rotary embedding module means that the channels do not have to be specified in any specific order and enables the

model to represent a variable numbers of channels. By training on heterogeneous datasets, the model instead learns to rely solely on the position embeddings which can handle arbitrary channels.

To validate our design choices, we performed a series of ablations on the model architecture and training scheme. First, we replaced the zeros-vector used to represent dropped out channels with a learnable vector that had norm similar to the data and input that into the encoder. After making the necessary changes to the inference code, we found that it performs comparably and save exploration in that direction for future work. We also performed an ablation on the best way to present the positional information to the model. We tried concatenating the x, y, z position to the encoder input along with the fine-time EEG signal as part of the token while representing temporal position with standard 1D-RoPE, but this performed poorly. Finally, we performed an ablation testing whether sending the noised EEG signal into the decoder input for dropped channel improved performance relative to sending noise without the EEG signal. The relative performance in these two cases were similar.

E Experiment Details

E.1 Metrics

Our two primary evaluation metrics for reconstruction are normalized mean-squared error (NMSE) and spectral bandpower error.

NMSE is computed as follows: (1) select the dropped channels or segments only; (2) include all timepoints from those channels; (3) flatten channel by time into one vector; (4) compute squared-error energy divided by clean-signal energy.

Then the NMSE is the arithmetic mean of these per-sample NMSE values across samples. This captures amplitude differences in signals, but not spectral signatures.

Spectral bandpower error is computed as follows. For each scored signal segment, we compare the spectral power of the clean target and reconstruction in five EEG frequency bands:

- Delta: 1–4 Hz
- Theta: 4–8 Hz
- Alpha: 8–13 Hz
- Beta: 13–30 Hz
- Gamma: 30–45 Hz

We demean each channel individually, apply a Hann window to each channel, and compute a normalized one-sided FFT power spectral density (PSD).

The PSD is integrated across each of the five frequency ranges, which yields total bandpower. The bandpower b is expressed in decibels (dB) by the transformation $10\log(\max(b, \varepsilon))$ where $\varepsilon = 10^{-6}$.

The bandpower error is the absolute (i.e. L_1 norm) difference in decibel bandpower between the reconstruction and the original signal.

We average over channels for each frequency band.

E.2 Competing Models

Unfortunately, many superresolution models do not release their code open-source. Thus, we necessarily had to restrict our competitors.

External neural baselines were frozen pretrained checkpoints with no retraining or adaptation, and all weights were available on Hugging Face.

We note that all of these methods mostly train solely on the Temple corpus, in addition to some extra datasets like the Healthy Brain Network and the Siena Scalp Database.

None of the reconstruction evaluation results expose training data to ZUNA or its competitors.

For LUNA, we applied a bandpass filter between 0.1 Hz and 75 Hz. A notch filter 50Hz is applied to remove power-line interference. All signals are then resampled to 256 Hz, just as in ZUNA.

For CBraMod, recordings are first band-pass filtered from 0.3 to 75 Hz and notch filtered at 60 Hz. The signals are then resampled to 200 Hz. Windows containing amplitudes greater than 100 microvolts are discarded, and the remaining signals are scaled by dividing by 100 microvolts. The data are finally divided into one-second, 200-sample patches for input to the model.

For ST-EEGFormer, recordings are band-pass filtered from 1 to 64 Hz and notch filtered at 50 Hz to remove power-line noise. The data are resampled to 128 Hz, and each observed channel is independently standardized over time to have zero mean and unit variance. Recordings are divided into windows (up to six seconds), electrode names are mapped to the model's channel identifiers, and each signal is divided into 16-sample patches.

During evaluation, the standardization or normalization procedures are inverted and subsequently resampled to a common 256 Hz baseline.

We chose a common reference/filter of 1–45 Hz as this includes the intersection of the filter intervals of all models.

F Reconstruction Performance

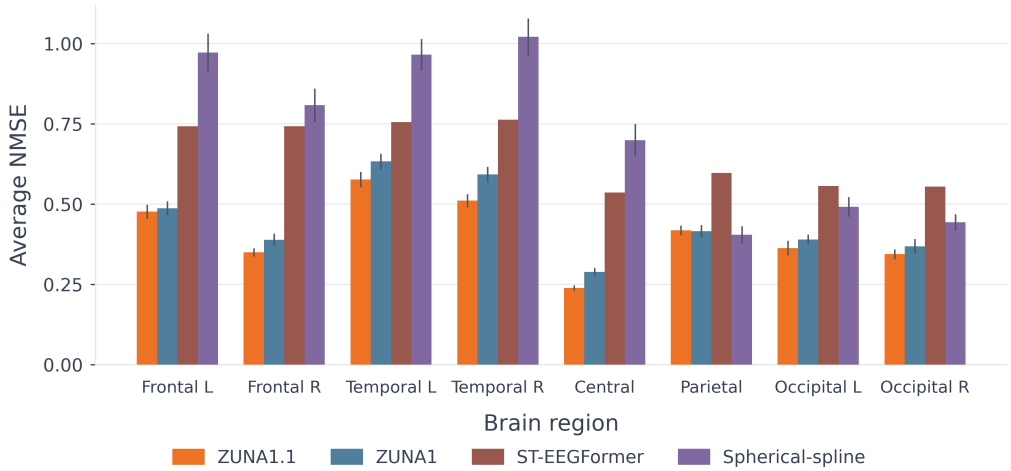


Figure 10. **Per-region reconstruction error.** The region-occlusion evaluation of Figure 1 shown as grouped bars: average NMSE across the four datasets for ZUNA1.1, ZUNA1, ST-EEGFormer, and spherical-spline interpolation. Error bars represent standard deviation.

Table 4. Average NMSE for region occlusion reconstruction across datasets, methods, and target regions.

Dataset	Region	ZUNA1.1	ZUNA1	ST-EEGFormer	Spherical-spline
ANPHY-Sleep	Frontal L	0.241 ± 0.019	**0.207 ± 0.016**	0.581 ± 0.054	0.547 ± 0.035
	Frontal R	**0.210 ± 0.011**	0.234 ± 0.021	0.604 ± 0.049	0.453 ± 0.049
	Temporal L	**0.393 ± 0.021**	0.465 ± 0.036	0.761 ± 0.058	0.718 ± 0.078
	Temporal R	**0.368 ± 0.015**	0.449 ± 0.035	0.753 ± 0.051	0.800 ± 0.087
	Central	0.100 ± 0.009	**0.069 ± 0.005**	0.315 ± 0.027	0.169 ± 0.018
	Parietal	0.111 ± 0.012	0.111 ± 0.007	0.652 ± 0.073	**0.105 ± 0.011**
	Occipital L	**0.091 ± 0.007**	0.136 ± 0.010	0.457 ± 0.032	0.184 ± 0.027
	Occipital R	**0.079 ± 0.007**	0.119 ± 0.008	0.541 ± 0.030	0.212 ± 0.023
BerlinBCI	Frontal L	**0.625 ± 0.048**	0.650 ± 0.050	0.645 ± 0.072	1.709 ± 0.185
	Frontal R	**0.608 ± 0.047**	0.634 ± 0.049	0.649 ± 0.059	1.615 ± 0.175
	Temporal L	**0.511 ± 0.035**	0.565 ± 0.024	0.715 ± 0.064	1.399 ± 0.151
	Temporal R	**0.506 ± 0.049**	0.566 ± 0.051	0.709 ± 0.080	1.334 ± 0.144
	Central	**0.328 ± 0.015**	0.335 ± 0.026	0.431 ± 0.021	0.642 ± 0.080
	Parietal	0.405 ± 0.031	**0.353 ± 0.027**	0.416 ± 0.044	0.513 ± 0.056
	Occipital L	**0.228 ± 0.018**	0.248 ± 0.019	0.520 ± 0.067	0.521 ± 0.021
	Occipital R	0.194 ± 0.015	**0.188 ± 0.014**	0.472 ± 0.039	0.404 ± 0.044
BCI2000	Frontal L	**0.240 ± 0.028**	0.289 ± 0.032	0.781 ± 0.023	0.400 ± 0.053
	Frontal R	**0.249 ± 0.019**	0.360 ± 0.028	0.743 ± 0.045	0.410 ± 0.064
	Temporal L	**0.404 ± 0.041**	0.468 ± 0.036	0.533 ± 0.061	0.714 ± 0.067
	Temporal R	**0.400 ± 0.027**	0.449 ± 0.035	0.596 ± 0.062	0.637 ± 0.069
	Central	**0.216 ± 0.007**	0.223 ± 0.017	0.407 ± 0.034	0.328 ± 0.044
	Parietal	**0.164 ± 0.013**	0.185 ± 0.014	0.308 ± 0.019	0.183 ± 0.030
	Occipital L	**0.132 ± 0.019**	0.165 ± 0.013	0.226 ± 0.009	0.196 ± 0.021
	Occipital R	**0.098 ± 0.008**	0.141 ± 0.009	0.188 ± 0.012	0.138 ± 0.011
AAD	Frontal L	**0.798 ± 0.062**	0.802 ± 0.062	0.963 ± 0.040	1.233 ± 0.134
	Frontal R	0.330 ± 0.006	**0.327 ± 0.045**	0.972 ± 0.058	0.753 ± 0.082
	Temporal L	**0.998 ± 0.077**	1.033 ± 0.079	1.012 ± 0.078	1.030 ± 0.072
	Temporal R	**0.767 ± 0.059**	0.903 ± 0.069	0.992 ± 0.063	1.311 ± 0.142
	Central	**0.311 ± 0.024**	0.528 ± 0.041	0.989 ± 0.055	1.656 ± 0.179
	Parietal	0.994 ± 0.047	1.014 ± 0.068	1.013 ± 0.087	**0.814 ± 0.088**
	Occipital L	**1.001 ± 0.087**	1.009 ± 0.057	1.023 ± 0.091	1.064 ± 0.115
	Occipital R	**1.004 ± 0.058**	1.026 ± 0.089	1.014 ± 0.091	1.017 ± 0.089

Table 5. Spectral bandpower reconstruction error for region occlusion. Values are mean absolute bandpower errors in dB, averaged over target-region median errors from the regional summaries; lower is better. Bold indicates the lowest error within each dataset and band.

Dataset	Method	Delta	Theta	Alpha	Beta	Gamma
ANPHY-Sleep	ZUNA1	1.214	1.175	1.172	1.247	1.346
	ZUNA1.1	**0.835**	**0.701**	**0.752**	**0.848**	**1.110**
	ST-EEGFormer	1.784	2.088	2.282	3.099	5.071
	Spherical-spline	0.917	0.833	0.877	1.064	1.574
BerlinBCI	ZUNA1	1.836	1.485	1.401	1.292	1.373
	ZUNA1.1	**1.381**	**0.935**	**0.783**	**0.818**	**1.040**
	ST-EEGFormer	3.574	3.824	3.548	5.256	8.554
	Spherical-spline	1.835	1.197	1.247	1.446	1.987
BCI2000	ZUNA1	1.398	1.205	1.148	1.087	**1.277**
	ZUNA1.1	**1.376**	**1.152**	**1.040**	**1.038**	1.479
	ST-EEGFormer	3.350	3.670	3.831	5.305	8.876
	Spherical-spline	1.466	1.188	1.121	1.060	1.744
AAD	ZUNA	2.557	2.595	2.428	2.650	2.863
	ZUNA1.1	**2.556**	**2.433**	**2.077**	**2.299**	**2.629**
	ST-EEGFormer	3.724	3.914	3.828	4.786	6.411
	Spherical-spline	4.049	4.364	3.611	4.062	4.099

Table 6. Patch-level reconstruction benchmark under 50% random-uniform dropout across random subjects. Error bars report standard deviation.

Dataset	Method	NMSE
AAD	ZUNA1.1	**0.381 ± 0.163**
	Spherical-spline	1.050 ± 0.584
	ST-EEGFormer	0.974 ± 0.583
	CBraMod	0.968 ± 0.076
	LUNA	1.008 ± 0.207
ANPHY-Sleep	ZUNA1.1	**0.094 ± 0.328**
	Spherical-spline	0.336 ± 0.968
	ST-EEGFormer	0.770 ± 0.962
	CBraMod	0.809 ± 0.206
	LUNA	1.002 ± 0.193
BCI2000	ZUNA1.1	**0.247 ± 0.380**
	Spherical-spline	0.377 ± 0.963
	ST-EEGFormer	0.712 ± 0.694
	CBraMod	0.929 ± 0.089
	LUNA	1.003 ± 0.194
BerlinBCI	ZUNA1.1	**0.651 ± 0.420**
	Spherical-spline	0.868 ± 1.045
	ST-EEGFormer	0.681 ± 0.781
	CBraMod	0.944 ± 0.088
	LUNA	1.017 ± 0.176

Table 7. Channel reconstruction NMSE on dropped channels. Lower is better.

Dataset	Dropout (%)	ZUNA1.1	ZUNA1	ST-EEGFormer	Spherical-spline
ANPHY-Sleep	20	0.342 ± 0.015	0.422 ± 0.128	0.634 ± 1.359	**0.328 ± 0.024**
	50	0.419 ± 0.071	**0.395 ± 0.050**	0.716 ± 1.229	0.496 ± 0.076
	75	**0.475 ± 0.043**	0.633 ± 0.090	0.852 ± 1.336	0.637 ± 0.035
	90	**0.745 ± 0.088**	0.845 ± 0.067	0.981 ± 1.654	1.443 ± 0.156
BerlinBCI	20	0.602 ± 0.039	**0.573 ± 0.018**	0.645 ± 0.186	0.762 ± 0.027
	50	0.737 ± 0.022	0.674 ± 0.017	**0.669 ± 0.120**	1.040 ± 0.033
	75	1.039 ± 0.052	0.840 ± 0.018	**0.753 ± 0.129**	1.616 ± 0.092
	90	1.486 ± 0.069	1.135 ± 0.025	**0.938 ± 0.165**	2.718 ± 0.583
BCI2000	20	**0.367 ± 0.014**	0.391 ± 0.021	0.742 ± 0.228	0.385 ± 0.021
	50	**0.419 ± 0.020**	0.429 ± 0.020	0.689 ± 0.157	0.457 ± 0.023
	75	**0.507 ± 0.027**	0.573 ± 0.021	0.721 ± 0.226	0.777 ± 0.040
	90	**0.711 ± 0.020**	0.931 ± 0.024	0.871 ± 0.191	1.715 ± 0.159
AAD	20	0.890 ± 0.001	**0.764 ± 0.001**	0.982 ± 0.239	0.970 ± 0.001
	50	0.845 ± 0.001	**0.796 ± 0.001**	1.001 ± 0.401	0.898 ± 0.001
	75	0.831 ± 0.001	**0.804 ± 0.001**	1.830 ± 0.520	1.061 ± 0.001
	90	**0.836 ± 0.001**	0.846 ± 0.001	2.111 ± 0.576	1.608 ± 0.001

G Downstream Tasks

The training objective contains a term that rewards reconstruction (decoder RF-loss) and a term that encourages latent representations to be more Gaussian distributed (encoder MMD-loss); however, there is no explicit term that shapes latents to be informative or useful for downstream tasks. At first, in the style of REVE (El Ouahidi et al., 2025), we hypothesized that our loss function was not encouraging useful latent structure, and we ablated our Maximum Mean Discrepancy (MMD) loss that encourages latent Gaussian structure. Deleting MMD did seem to help classification accuracy by about 2% per task (absolute increase) on average, and motivated our switch to turning off MMD loss in our main training run at training step 400K (i.e. our "Hybrid" variation trained with MMD loss up to 400K training steps, and turned off thereafter).

However, this did not seem to help the accuracy degradation, which continued as training continued.

Table 8. Classification accuracy (0-1 scale) for the MMD loss ablation across downstream tasks. Bold indicates the highest classification accuracies. We find a peak early on in our accuracies, and then a sharp decay as the latent representations are increasingly specialized towards reconstruction.

Dataset	No MMD-50K	MMD–50K	MMD–400K	Hybrid–575K
Audiovisual	**0.53 ± 0.05**	0.40 ± 0.03	0.36 ± 0.04	0.35 ± 0.03
CVEP	0.88 ± 0.02	**0.88 ± 0.09**	0.87 ± 0.01	0.71 ± 0.07
Depression	0.89 ± 0.01	**0.91 ± 0.01**	0.89 ± 0.01	0.90 ± 0.02
ERN	**0.81 ± 0.01**	0.81 ± 0.03	0.80 ± 0.02	0.80 ± 0.02
LRP	0.77 ± 0.02	**0.78 ± 0.01**	0.78 ± 0.01	0.50 ± 0.01
Mental Arithmetic	**0.73 ± 0.02**	0.68 ± 0.02	0.67 ± 0.02	0.62 ± 0.05
Mental Imagery	**0.26 ± 0.01**	0.25 ± 0.01	0.22 ± 0.01	0.20 ± 0.01
N170	0.72 ± 0.01	**0.72 ± 0.01**	0.70 ± 0.01	0.58 ± 0.06
N2PC	0.58 ± 0.01	**0.62 ± 0.02**	0.55 ± 0.02	0.50 ± 0.01
N400	**0.63 ± 0.01**	0.61 ± 0.02	0.61 ± 0.02	0.59 ± 0.04

Table 9. Reconstruction NMSE under 50% random-uniform dropout. Bold indicates the lowest errors, showing that our final checkpoint is almost uniformly superior for reconstruction.

Dataset	No MMD-50K	MMD–50K	MMD–400K	Hybrid–575K
Audiovisual	0.53 ± 0.02	0.53 ± 0.02	0.51 ± 0.02	**0.50 ± 0.02**
Mental Arithmetic	1.01 ± 0.04	1.05 ± 0.04	**1.00 ± 0.04**	1.06 ± 0.04
Mental Imagery	0.56 ± 0.04	0.66 ± 0.04	0.55 ± 0.04	**0.54 ± 0.04**
CVEP	1.22 ± 0.06	1.88 ± 0.05	**1.19 ± 0.07**	1.24 ± 0.00
Depression	0.83 ± 0.07	0.86 ± 0.07	0.83 ± 0.07	**0.82 ± 0.07**
ERN	1.16 ± 0.04	1.08 ± 0.01	1.03 ± 0.03	**1.02 ± 0.01**
LRP	1.09 ± 0.02	1.09 ± 0.01	1.05 ± 0.01	**1.02 ± 0.01**
N170	1.08 ± 0.02	1.06 ± 0.01	**1.01 ± 0.01**	1.02 ± 0.01
N400	1.14 ± 0.04	1.10 ± 0.01	**1.02 ± 0.01**	**1.02 ± 0.01**
N2PC	1.07 ± 0.02	1.08 ± 0.01	1.05 ± 0.02	**1.03 ± 0.01**

In Figure 11, we show the model rankings on EEG foundation models including ZUNA1.1 using the NeuralBench metrics. Here, the "MMD-50K" checkpoint yields respectable performance on the NeuralBench datasets. However, our results cast serious doubt on the usage of masked autoencoders trained to reconstruct missing data for downstream tasks.

Finally, investigating the claim that the final encoder layer was overfit to the reconstruction task, we systematically explored whether building a linear classifier head on earlier layers of the encoder improves classification performance. Figure 12 shows balanced accuracy for classifiers built on top of different layers in the encoder for a few datasets in the NeuralBench benchmark. We observed some improved classification performance when using representations from earlier encoder layers, but nothing systematic. We include these results for completeness.

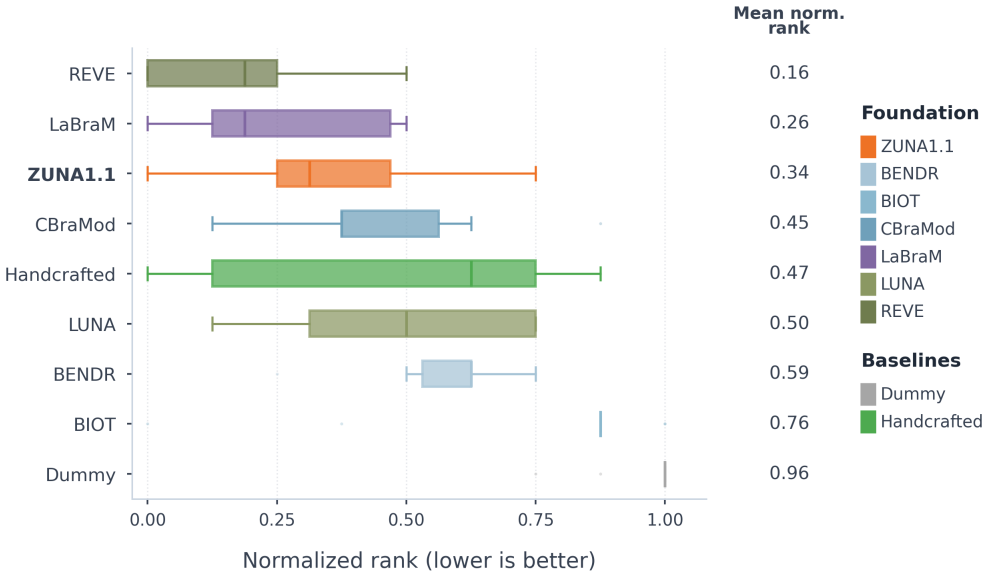
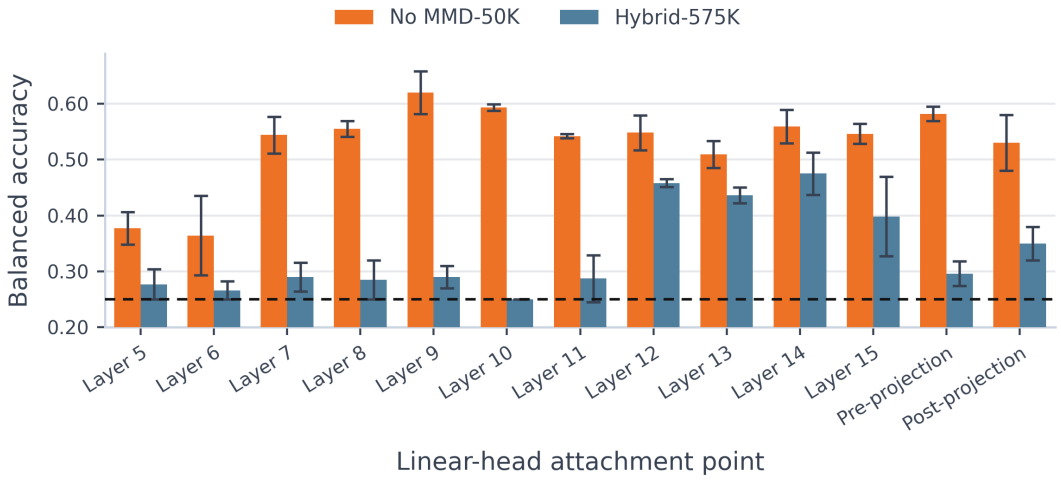
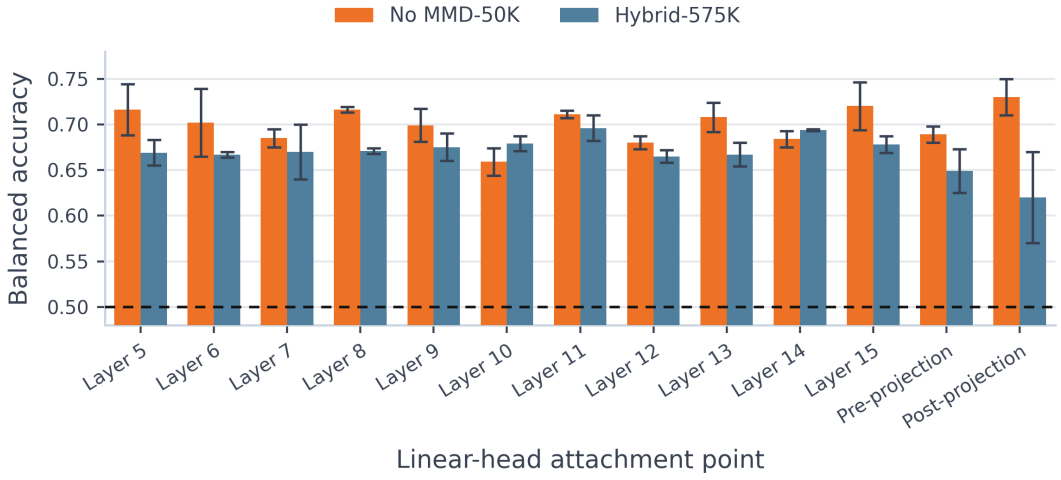


Figure 11. **Model ranking on downstream tasks.** Models are ordered by their mean normalized rank, displayed on the right hand side of the boxes. Lower is better.

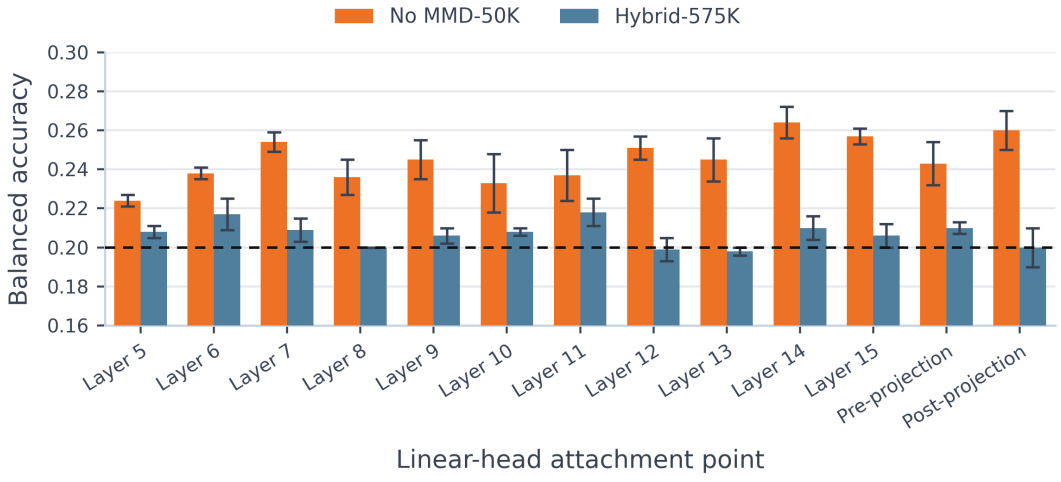


(a) Audiovisual

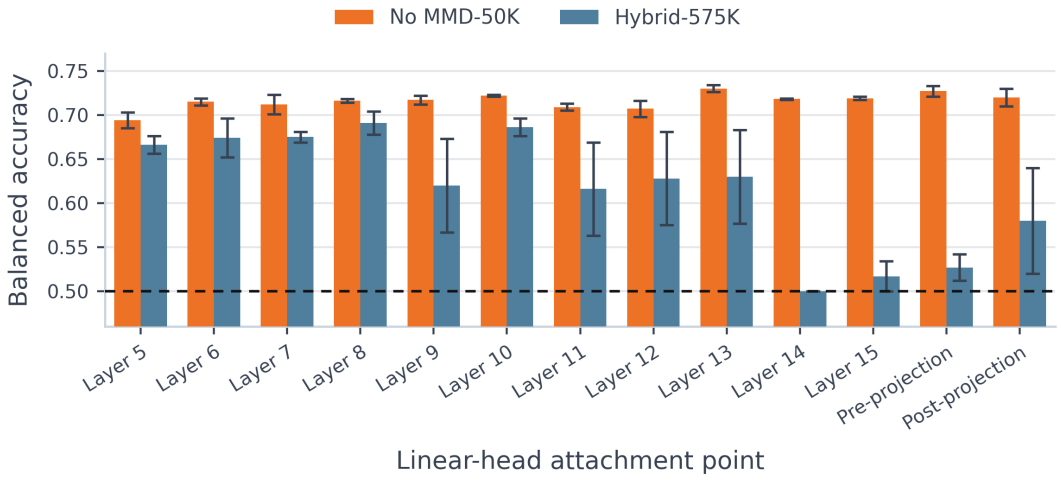


(b) Mental Arithmetic

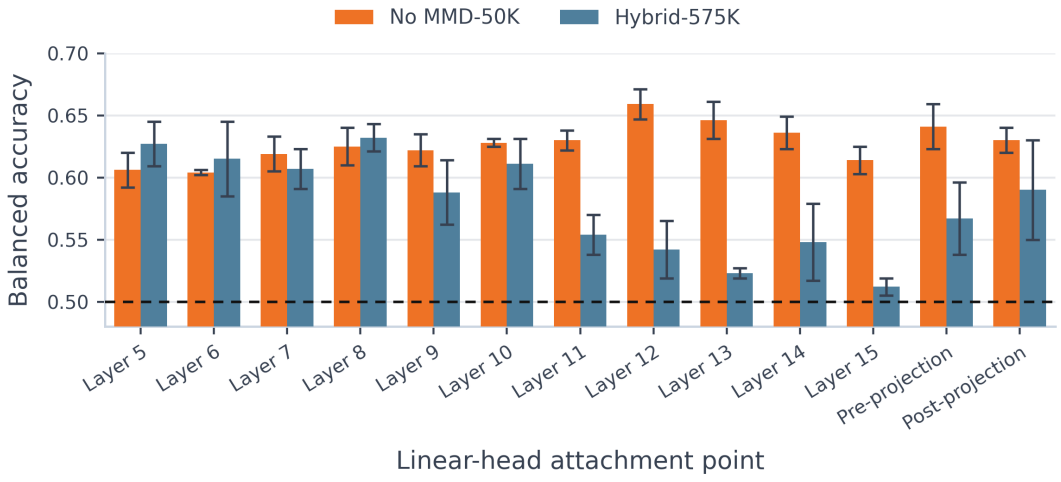
Figure 12. Balanced accuracy for linear classifiers attached at different ZUNA representation layers. The black dashed line marks chance-level performance, matching the Dummy baseline in Table 2. We ran this experiment to probe the representational quality due to previous hypotheses in EEG literature that the final layer of masked autoencoders is prone to overfitting to the reconstruction task. We do not see uniform evidence of this across datasets.



(c) Mental Imagery



(d) N170



(e) N400

Figure 12. Balanced accuracy for linear classifiers attached at different ZUNA representation layers, continued.

H TRAINING CORPUS

Both ZUNA variants were trained on the Temple University EEG corpus (Obeid & Picone, 2016) and assorted datasets from OpenNeuro.

Crucially, we verified absolutely zero leakage in both downstream classification and reconstruction results with our four main datasets here.

Table 10 lists the OpenNeuro dataset identifiers in the training corpus.

As discussed in the main paper and various points in the appendix, the main differences in the training datasets of the two model variants is the preprocessing, particularly with our new quality metric and segment length generalization.

Crucially, AAD, ANPHY-Sleep, BCI2000 and BerlinBCI evalation datasets are hosted on Zenodo or PhysioNet, and we verified are not included in OpenNeuro.

Table 10. OpenNeuro dataset identifiers used in the training corpus.

OpenNeuro IDs
ds001785 ds001787 ds001849 ds001971 ds002034 ds002218 ds002336 ds002338
ds002578 ds002680 ds002691 ds002720 ds002721 ds002778 ds002893 ds003004
ds003039 ds003061 ds003194 ds003195 ds003420 ds003421 ds003458 ds003474
ds003478 ds003490 ds003506 ds003509 ds003516 ds003517 ds003518 ds003519
ds003522 ds003523 ds003555 ds003570 ds003574 ds003602 ds003620 ds003626
ds003638 ds003655 ds003670 ds003690 ds003710 ds003739 ds003751 ds003753
ds003768 ds003774 ds003800 ds003801 ds003810 ds003822 ds003838 ds003846
ds003885 ds003887 ds003969 ds003987 ds004000 ds004010 ds004017 ds004018
ds004019 ds004022 ds004024 ds004033 ds004040 ds004043 ds004067 ds004105
ds004106 ds004117 ds004118 ds004119 ds004120 ds004121 ds004122 ds004123
ds004147 ds004151 ds004152 ds004196 ds004200 ds004252 ds004256 ds004262
ds004264 ds004279 ds004284 ds004295 ds004306 ds004315 ds004317 ds004324
ds004347 ds004348 ds004350 ds004356 ds004357 ds004367 ds004369 ds004388
ds004389 ds004395 ds004444 ds004446 ds004448 ds004475 ds004477 ds004502
ds004505 ds004511 ds004515 ds004532 ds004561 ds004563 ds004572 ds004574
ds004577 ds004579 ds004580 ds004582 ds004584 ds004587 ds004588 ds004595
ds004602 ds004603 ds004621 ds004626 ds004635 ds004657 ds004660 ds004661
ds004706 ds004718 ds004745 ds004752 ds004784 ds004796 ds004802 ds004816
ds004817 ds004840 ds004841 ds004842 ds004843 ds004844 ds004849 ds004850
ds004851 ds004852 ds004853 ds004854 ds004855 ds004860 ds004883 ds004902
ds004917 ds004942 ds004951 ds004952 ds004995 ds005021 ds005034 ds005048
ds005087 ds005089 ds005095 ds005106 ds005114 ds005121 ds005131 ds005170
ds005189 ds005207 ds005262 ds005273 ds005296 ds005305 ds005307 ds005340
ds005342 ds005363 ds005383 ds005385 ds005397 ds005403 ds005406 ds005410
ds005416 ds005420 ds005429 ds005486 ds005505 ds005506 ds005507 ds005510
ds005511 ds005512 ds005520 ds005530 ds005565 ds005571 ds005586 ds005594
ds005620 ds005672 ds005688 ds005692 ds005697 ds005779