Energy-based Autoregressive Generation for Neural Population Dynamics

Ningling Ge[1,2,3*], **Sicheng Dai**[1,2,3,4*], **Yu Zhu**[1,2,3,4†], **Shan Yu**[1, 3†]

[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artifcial Intelligence, University of Chinese Academy of Sciences
[3]State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology
[4]Beijing Academy of Artificial Intelligence
geningling2023@ia.ac.cn, daisicheng2023@ia.ac.cn, zhuyu2022@ia.ac.cn, shan.yu@nlpr.ia.ac.cn

Abstract

Understanding brain function represents a fundamental goal in neuroscience, with critical implications for therapeutic interventions and neural engineering applications. Computational modeling provides a quantitative framework for accelerating this understanding, but faces a fundamental trade-off between computational efficiency and high-fidelity modeling. To address this limitation, we introduce a novel Energy-based Autoregressive Generation (EAG) framework that employs an energy-based transformer learning temporal dynamics in latent space through strictly proper scoring rules, enabling efficient generation with realistic population and single-neuron spiking statistics. Evaluation on synthetic Lorenz datasets and two Neural Latents Benchmark datasets (MC_Maze and Area2_bump) demonstrates that EAG achieves state-of-the-art generation quality with substantial computational efficiency improvements, particularly over diffusion-based methods. Beyond optimal performance, conditional generation applications show two capabilities: generalizing to unseen behavioral contexts and improving motor brain-computer interface decoding accuracy using synthetic neural data. These results demonstrate the effectiveness of energy-based modeling for neural population dynamics with applications in neuroscience research and neural engineering.

Code — https://github.com/NinglingGe/Energy-based-Autoregressive-Generation-for-Neural-Population-Dynamics

Introduction

Neural population dynamics form a fundamental computational basis of brain function, where coordinated spike patterns across neuron ensembles encode sensory information (Romo and Salinas 2003; Panzeri et al. 2022), motor commands (Churchland et al. 2012; Gallego et al. 2017), and cognitive states (Mante et al. 2013; Rigotti et al. 2013). Elucidating these computational mechanisms not only provides mechanistic insight into cortical neural coding (Gallego et al. 2017; Safaie et al. 2023) but also advances therapeutic interventions for neurological disorders including Parkinson's disease (Little et al. 2013), and facilitates brain-computer interfaces (BCIs) for motor restoration in paralysis (Hochberg et al. 2012; Willett et al. 2023).

Computational modeling provides a quantitative framework to analyze neural mechanisms and population dynamics (Sussillo et al. 2015; Vyas et al. 2020). These approaches typically fall into **encoding and decoding models** (Mathis et al. 2024). Encoding models characterize how external variables are transformed into neural activity patterns by establishing statistical relationships between stimuli and neural responses (Walker et al. 2019; Wang et al. 2025), while decoding models take the inverse approach, reconstructing behavioral or stimulus variables from recorded neural activity (Gallego et al. 2017; Yoshida and Ohki 2020; Zhu et al. 2025). Decades of work have refined decoders that extract behavior and stimulus features from high-dimensional neural recordings (Sussillo et al. 2016; Ye et al. 2023; Azabou et al. 2024). In contrast, encoding models **remain underexplored**, despite their importance in revealing low-dimensional manifold dynamics (Gallego et al. 2017; Pandarinath et al. 2018) in motor cortex (Gallego et al. 2020; Safaie et al. 2023; Zhu et al. 2025) and enabling more interpretable, robust decoding for motor BCIs (Hochberg et al. 2012; Willett et al. 2023).

In this framework, neural encoding models can be further divided into predictive and generative models. Predictive models directly map stimuli to neural responses (Bashivan, Kar, and DiCarlo 2019; Walker et al. 2019), offering efficiency but limited ability to capture **maintain trial-to-trial variability** (Churchland et al. 2010; Ecker et al. 2014). Generative models fall into two main classes: VAE-based methods (Zhou and Wei 2020; Hurwitz et al. 2021; Keshtkaran et al. 2022) which sample from latent spaces conditioned on priors but fail to capture complex **population and single-neuron statistics**; and diffusion-based methods, such as LDNS (Kapoor et al. 2024) and GNOCCHI (McCart et al. 2024), which model response variability through latent-space distributions but require costly iterative sampling, leading to **inefficient estimation** of neural statistics.

To efficiently and effectively modeling, we develop a novel Energy-based Autoregressive Generation (EAG) framework. EAG employs an energy-based transformer that learns temporal dynamics in latent space through strictly proper scoring rules (Székely 2003), enabling efficient gen-

*These authors contributed equally.
†Corresponding Author

eration while achieving high fidelity and preserving trial-to-trial variability. The framework supports both unconditional generation for studying neural dynamics and conditional generation for modeling behavior-neural relationships. We evaluate EAG on synthetic Lorenz datasets and two real neural datasets from the Neural Latents Benchmark (Pei et al. 2021): MC_Maze and Area2_bump. The method achieves state-of-the-art (SOTA) generation quality with substantial computational gains, especially delivering a 96.9% speed-up over diffusion-based approaches. Beyond improvements in generation quality and computational efficiency, conditional generation applications demonstrate two capabilities: (1) generalization to unseen behavioral contexts, revealing generalizable computational mechanisms; and (2) up to a 12.1% improvement in motor BCI decoding accuracy when trained with EAG-generated data. These results demonstrate the effectiveness of energy-based modeling for neural population dynamics.

In conclusion, our main contributions are as follows:

- We develop the **novel EAG framework** that resolves the trade-off between computational efficiency and high-quality modeling through energy-based learning.
- We demonstrate that EAG achieves **SOTA generation quality** with substantial **efficiency improvements** over existing methods.
- We show that conditional generation enables **generalization** to unseen behavioral contexts and **improvement** of motor BCI decoding accuracy, demonstrating practical applications beyond basic neural modeling.

Related Work

Neural-Behavioral Modeling. Understanding neural computation relies on encoding models that predict neural responses from behavioral variables and decoding models that infer behavioral states from neural signals (Mathis et al. 2024). Deep neural networks have advanced encoding models for stimulus-response mappings in visual cortex (Yamins and DiCarlo 2016; Kell et al. 2018; Walker et al. 2019; Bashivan, Kar, and DiCarlo 2019; Marks and Goard 2021; Vargas et al. 2024; Wang et al. 2025) and decoding models for extracting behavioral information from motor (Gallego et al. 2017) and visual cortices (Yoshida and Ohki 2020; Stringer et al. 2021). However, these approaches primarily focus on input-output mappings cannot capture the stochastic variability inherent in neural population dynamics.

Neural Spike Generation. Neural spike generation addresses data scarcity in brain-computer interface applications through synthetic data augmentation for decoder training and stability (Wen et al. 2023; Ma et al. 2023). Variational autoencoders and generative adversarial networks have been applied to neural population modeling and spike train synthesis (Pandarinath et al. 2018; Wen et al. 2023; Ma et al. 2023). Latent diffusion for neural spike generation (Kapoor et al. 2024; McCart et al. 2024) encodes spike trains into continuous latent spaces and applies diffusion models to generate behaviorally-conditioned neural activity, but requires computationally expensive iterative denoising steps (Ho, Jain, and Abbeel 2020). In contrast, this work applies energy-based modeling to autoregressive neural spike generation, enabling efficient sampling while maintaining high-quality spike pattern generation.

Energy-based Models. Energy-based models define probability distributions through energy functions, where probability is inversely related to energy (Hinton 2002; LeCun et al. 2006). These models enable direct sampling from learned distributions, providing a principled approach to generative modeling (Song and Ermon 2019). EBMs have been successfully applied across diverse domains including image generation (Song and Ermon 2019), natural language processing (Bakhtin et al. 2021), molecular design (Satorras, Hoogeboom, and Welling 2021), and protein structure design (Watson et al. 2023). In contrast, this work introduces EBMs to neural computational modeling for the first time.

Preliminaries

Strictly Proper Scoring Rules

Scoring rules evaluate probabilistic forecasts by comparing predicted distributions against observed outcomes. Given sample space $\mathcal{X}$ and probability measures $\mathcal{P}$ on $\mathcal{X}$, a scoring rule S maps predicted distributions p and observed samples x to extended real values:

$$S(p, x) : \mathcal{P} \times \mathcal{X} \mapsto \overline{\mathbb{R}}. \tag{1}$$

The expected score under the true distribution q quantifies prediction quality:

$$S(p, q) = \mathbb{E}_{x\sim q}[S(p, x)]. \tag{2}$$

A scoring rule is proper if truthful reporting maximizes expected scores:

$$S(p, q) \leq S(q, q), \quad \forall p, q \in \mathcal{P}. \tag{3}$$

Strict propriety occurs when equality holds exclusively for $p = q$, ensuring unique optimal predictions.

Classical scoring rules include the Brier score (Brier 1950), logarithmic score (Good 1952) and spherical score (Roby 1965). For continuous distributions, the energy score (Székely 2003) provides a strictly proper scoring rule that will be utilized in this work.

Scoring Rules for Generative Modeling

Strictly proper scoring rules provide objectives for training generative models through negative score loss functions:

$$\mathcal{L}_S(p, x) = -S(p, x). \tag{4}$$

The logarithmic score yields cross-entropy loss when maximized. Strict propriety ensures unique optimization targets.

For sequential data, autoregressive generation decomposes the loss across time steps:

$$\mathcal{L}_S(p, x) = -\sum_{t=1}^{T} S(p(\cdot|x_{<t}), x_t). \tag{5}$$

Expected loss minimization requires each conditional distribution $p(\cdot|x_{<t})$ to match the true conditional $q(\cdot|x_{<t})$.

Neural spike data lacks explicit likelihood forms, resulting direct score calculation intractable. This necessitates tractable score estimators that preserve strict propriety for neural spike generation.

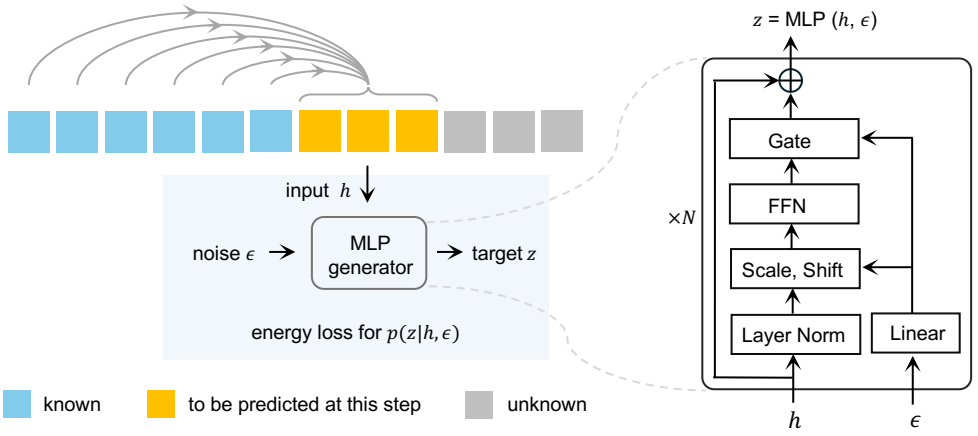


Figure 1: Energy-based Autoregressive Generation (EAG) framework. Known latent positions (blue) provide context for predicting masked positions (gray). The MLP generator incorporates noise $\boldsymbol{\epsilon}$ via adaptive layer normalization to enable stochastic generation, trained with energy loss for distributional prediction.

Methods

Latent Generation Framework

EAG adopts a two-stage paradigm that first learns compact neural representations and then performs efficient generation in the latent space. While Stage 1 follows established autoencoder approaches for fair comparison, EAG's core contribution lies in Stage 2, where we introduce a novel energy-based generation mechanism that fundamentally differs from existing diffusion-based methods.

Stage 1: Neural Representation Learning. Following LDNS (Kapoor et al. 2024), we employ their autoencoder architecture to obtain latent representations $\mathbf{z} \in \mathbb{R}^{d \times T}$ from neural spiking data $\mathbf{s} \in \mathbb{N}_0^{n \times T}$ and optional behavioral covariates $\mathbf{y}$, where $d \ll n$. This stage maps high-dimensional spike trains to a low-dimensional latent space under a Poisson observation model with temporal smoothness constraints. We use identical network architecture and training configuration as (Kapoor et al. 2024).

Stage 2: Energy-based Latent Generation. Given the learned latent representations $\mathbf{z}$ from Stage 1, EAG employs an energy-based autoregressive framework for latent generation. The approach predicts missing latent representations through masked autoregressive modeling guided by the energy score, a strictly proper scoring rule that does not require explicit likelihood computation. This formulation enables single-pass generation while preserving stochastic properties necessary for modeling trial-to-trial variability. The detailed methodology is presented in the following section.

Energy-based Autoregressive Generation

Building upon the latent representations $\mathbf{z}$ learned in Stage 1, EAG employs energy-based modeling for neural population dynamics generation through autoregressive prediction, as shown in Figure 1. The framework addresses a fundamental challenge in neural generative modeling: predicting distributions over continuous latent spaces without explicit likelihood computation while preserving the stochastic nature of neural variability.

Energy Loss The energy score provides a strictly proper scoring rule for continuous latent variables in $\mathbb{R}^d$. For model distribution p_θ generating latent samples $\mathbf{z}$ and data distribution with ground truth latents $\mathbf{z}_{\text{data}}$, the energy score with parameter $\alpha \in (0, 2)$ is defined as:

$$S(p_\theta, \mathbf{z}_{\text{data}}) = \mathbb{E}[\|\mathbf{z}_1 - \mathbf{z}_2\|^\alpha] - 2\mathbb{E}[\|\mathbf{z} - \mathbf{z}_{\text{data}}\|^\alpha] \quad (6)$$

where $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}$ are independent samples from p_θ. The energy score achieves strict propriety for $\alpha \in (0, 2)$, ensuring that optimal predictions correspond to the true latent distribution.

The energy loss can be unbiasedly estimated using two independent latent samples $\mathbf{z}_1, \mathbf{z}_2$ from the model distribution:

$$\mathcal{L}_{\text{energy}}(p_\theta, \mathbf{z}_{\text{data}}) = \|\mathbf{z}_1 - \mathbf{z}_{\text{data}}\|^\alpha + \|\mathbf{z}_2 - \mathbf{z}_{\text{data}}\|^\alpha - \|\mathbf{z}_1 - \mathbf{z}_2\|^\alpha \quad (7)$$

This formulation performs distributional prediction by balancing two objectives: the first two terms minimize prediction error while the third term maintains sample diversity. Essentially, the energy loss trains the model to generate samples that are both accurate and appropriately variable, capturing the distributional properties necessary for modeling neural trial-to-trial variability. Unlike diffusion models that require iterative sampling steps, this approach enables direct distributional prediction in a single forward pass while maintaining generative capabilities.

Energy Transformer Architecture The energy transformer architecture enables stochastic latent generation through noise-conditioned output layers while maintaining standard transformer processing for temporal dynamics modeling.

Input Processing. Latent representations $\mathbf{z} \in \mathbb{R}^d$ are mapped to model dimension d_{model} through linear projection layers. Positional encodings are applied to maintain temporal ordering information across the latent sequence.

Stochastic Output Generation. As illustrated in Figure 1, the output layer incorporates random noise $\boldsymbol{\epsilon}$ through a

multi-layer perceptron generator to enable stochastic latent generation. The noise vector of dimension d_{noise} is sampled from uniform distribution $[-0.5, 0.5]$ and embedded to dimension d_{mlp} through learned linear transformations.

The MLP generator employs residual blocks with adaptive layer normalization to inject noise into latent predictions, as shown in the architectural detail of Figure 1. For the i-th residual block with input $\mathbf{h}^i$:

$$\begin{aligned} \mathbf{h}^i_{\boldsymbol{\epsilon}} &= (1 + \text{scale}(\boldsymbol{\epsilon})) \cdot \text{LN}(\mathbf{h}^i) + \text{shift}(\boldsymbol{\epsilon}) \\ \mathbf{h}^{i+1} &= \mathbf{h}^i + \text{gate}(\boldsymbol{\epsilon}) \cdot \text{FFN}(\mathbf{h}^i_{\boldsymbol{\epsilon}}) \end{aligned} \tag{8}$$

where $\text{shift}(\cdot)$, $\text{scale}(\cdot)$, and $\text{gate}(\cdot)$ are learned linear transformations that interpret noise input as adaptive parameters for controlling the scale and bias of feature transformations. The gating mechanism allows the model to selectively incorporate stochastic variations based on the current context.

Masked Autoregressive Generation. As depicted in Figure 1, the framework employs masked autoregressive modeling where known latent positions (blue) provide temporal context for predicting masked positions (gray). During training, random masking ratios sampled uniformly from $[0.7, 1.0]$ are applied to latent sequences, ensuring the model learns to handle various degrees of missing information. The masking strategy randomly selects time points to predict while preserving causal relationships in the temporal sequence.

During inference, latent time points are generated progressively with masking ratio decreasing from 1.0 to 0 following a cosine schedule. This approach enables bidirectional attention during training for improved context utilization while maintaining autoregressive properties during generation. The progressive unmasking allows the model to iteratively refine predictions based on previously generated latent states.

Conditional Generation For behavior-conditioned neural generation, EAG incorporates behavioral covariates $\mathbf{y}$ into the input sequence as conditioning tokens. Behavioral variables are embedded through a linear projection layer to match the latent dimension, then concatenated with the latent representations.

During training, behavioral conditions are randomly replaced with learnable null tokens for 10% of trials to enable classifier-free guidance. This dropout strategy allows the model to learn both conditional and unconditional generation within a unified framework.

At inference, the model generates latent representations for both the given behavioral condition $\mathbf{h}_c$ and null condition $\mathbf{h}_u$, with the final representation computed using classifier-free guidance:

$$\mathbf{h} = \gamma \cdot \mathbf{h}_c + (1 - \gamma) \cdot \mathbf{h}_u \tag{9}$$

where γ controls the strength of behavioral conditioning. This enables flexible control over the degree of behavioral constraint during generation.

Results

Our experiments proceed in four stages. First, we conduct unconditional generation on three diverse distinct datasets, and benchmark its performance against VAE-based and diffusion-based generative models on two real neural datasets. Second, we demonstrate that EAG achieves 30× higher sampling efficiency than diffusion models, enabling faster generation without sacrificing quality. Third, EAG generalizes well to unseen labels on conditional generation, capturing neural-behavioral relationships while preserving trial variability. Finally, we show that EAG enhances downstream BCI decoding performance.

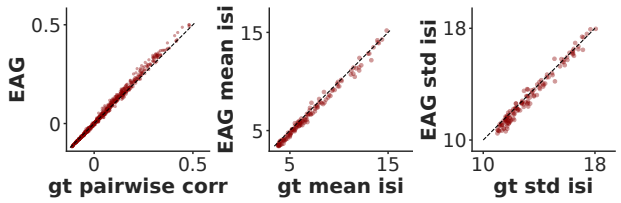


Figure 2: Unconditional generation on Lorenz Dataset. The generated data closely matches the ground truth across metrics: pairwise correlation, mean-isi, and std-isi.

Datasets. We evaluate our method on one synthetic dataset and two real neural datasets. The synthetic dataset is generated by simulating 128-dimensional neural spiking activity using a Lorenz attractor. The two real datasets, MC_Maze and Area2_Bump, are from the Neural Latents Benchmark (Pei et al. 2021), which record neural activity from different cortical areas of macaque monkeys performing various tasks. Detailed descriptions are provided in the supplementary.

Baselines. To rigorously benchmark the generative performance of EAG, we compared it against a suite of established baseline models drawn from both VAE-based and diffusion-based models. The VAE-based baselines include Targeted Neural Dynamical Modeling (TNDM) (Hurwitz et al. 2021), Poisson-identifiable VAE (pi-VAE) (Zhou and Wei 2020), and AutoLFADS (Sussillo et al. 2016; Pandarinath et al. 2018; Sedler and Pandarinath 2023; Keshtkaran et al. 2022). Additionally, we include Latent Diffusion for Neural Spiking (LDNS) (Kapoor et al. 2024) as a diffusion-based baseline, given its recent advances over AutoLFADS. Detailed settings are provided in the supplementary.

Metrics. For evaluation, we adopted the comprehensive set of metrics previously proposed in the LDNS study (Kapoor et al. 2024). The four metrics include population spike count distribution, pairwise spike-count correlations, mean inter-spike interval (ISI), and ISI standard deviation, capturing both population-level and single-neuron spiking statistics.

EAG Generates High-quality Neural Spike Data

To validate effectiveness of EAG framework, we first evaluate it under an unconditional generation setting on three datasets characterized by diverse distributions and high variability. We demonstrate that EAG achieves state-of-the-art performance against both VAE-based and diffusion-based methods.

Lorenz Dataset. We first apply EAG to a synthetic dataset

Dataset	Method	D_{KL} psch	RMSE pairwise corr	RMSE mean isi	RMSE std isi
MC_Maze	TNDM	0.0028 ± 6.0e-5	0.0027 ± 1.2e-5	0.057 ± 0.004	0.029 ± 0.001
	pi-VAE	0.0063 ± 2.0e-4	0.0031 ± 1.1e-5	0.064 ± 0.002	0.034 ± 0.001
	AutoLFADS	0.0040 ± 2.2e-4	0.0026 ± 1.3e-5	0.039 ± 0.003	0.029 ± 0.001
	LDNS	0.0039 ± 3.0e-4	0.0025 ± 1.1e-4	0.037 ± 0.001	0.023 ± 0.0001
	EAG	**0.0014 ± 2.0e-4**	**0.0024 ± 1.0e-5**	**0.024 ± 0.001**	**0.018 ± 0.0024**
Area2_Bump	TNDM	0.0027 ± 2.9e-4	0.0077 ± 1.0e-4	0.049 ± 0.009	0.039 ± 0.003
	pi-VAE	0.0067 ± 4.2e-4	0.0088 ± 7.9e-5	0.050 ± 0.007	0.029 ± 0.004
	AutoLFADS	0.0032 ± 3.2e-4	0.0081 ± 1.2e-5	0.048 ± 0.003	0.031 ± 0.006
	LDNS	0.0020 ± 1.2e-4	0.0076 ± 1.4e-5	0.050 ± 0.002	0.034 ± 0.002
	EAG	**0.0018 ± 1.6e-4**	**0.0075 ± 9.1e-5**	**0.035 ± 0.004**	**0.025 ± 0.003**

Table 1: Model metrics comparison. D_{KL} for the population spike count histogram and RMSE comparisons. Results are reported as mean ± standard deviation over 5 folds. EAG outperforms all baselines (Wilcoxon, $p < 0.001$), except for pairwise correlation vs. LDNS (n.s., $p = 0.09$). Bold indicates the best-performing model; underlined values indicate the second best.

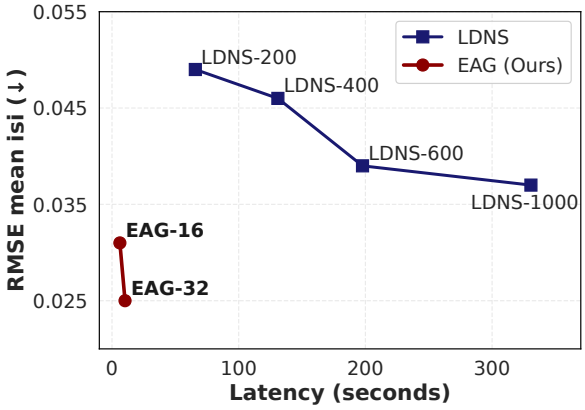


Figure 3: The latency/quality trade-off for EAG and LDNS. We vary number of diffusion steps (200, 400, 600, 1000) of LDNS and number of autoregressive steps (16, 32) of EAG. EAG-32 achieves a 96.9% reduction in latency, and a 32.4% improvement on RMSE mean ISI compared to LDNS-1000.

generated from a 3D Lorenz system, where 128-dimensional neural observations over 256 time steps are derived from projections of the underlying attractor dynamics. In the first stage, we train an autoencoder to effectively reconstruct the firing rates. Then we train EAG within the latent space in the second stage. Supp. Figure S1 and Figure S2 visualize Lorenz rates generated by EAG and the corresponding spike trains obtained via Poisson sampling, which are visually indistinguishable from the real data. We quantitatively evaluate EAG's performance, as visualized in Figure 2. EAG captures both population-level features (e.g. pairwise correlation) and single-neuron statistics (e.g., ISI mean and variance). Additionally, Supp. Figure S3 demonstrates that both AE-reconstructed and EAG-generated rates accurately capture the ground-truth pairwise correlation structure of the synthetic Lorenz rates.

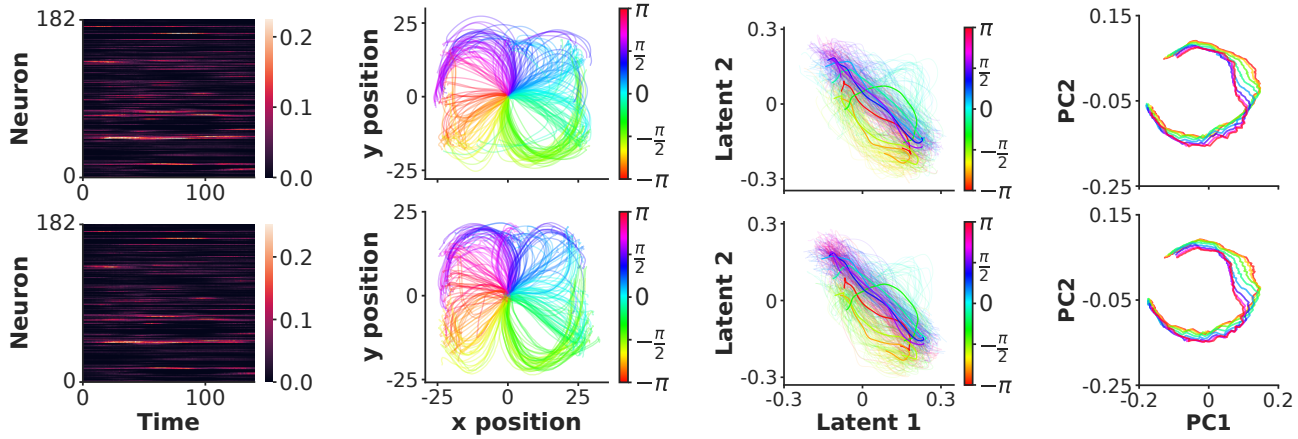
MC_Maze Dataset. We then benchmark EAG on MC_Maze (Churchland and Kaufman 2022; Pei et al. 2021), a widely used real neural dataset often referred to as the "neural MNIST," recorded from premotor and primary motor cortex as a monkey performs a delayed center-out reaching task with spatial barriers. EAG is shown to generate highly realistic sparse spike trains (Supp. Figure S4). Quantitative comparison against VAE-based and diffusion-based baselines demonstrates that EAG consistently achieves the best performance across all four evaluation metrics as Table 1 (visualization in Supp. Figure S5). Notably, even after augmenting baselines with spike-history inputs (a trick known to benefit LDNS), EAG still achieves the best performance (Supp. Table S1). In addition, a plain autoregressive Transformer trained with MSE loss performs significantly worse than EAG (Supp. Table S1), highlighting the essential contribution of the energy loss.

Area2_Bump Dataset. To further test EAG's robustness on limited data and non-autonomous neural activity, we evaluate it on the Area2_Bump dataset, a small-scale neural dataset containing about 300 trials recorded from the somatosensory cortex during a bump-perturbed reaching task. Despite the data scarcity, EAG continues to generate spike trains that are both visually and statistically aligned with the real data (Supp. Figure S6, Supp. Figure S7), outperforming all baselines as Table 1 shows. Additionally, Supp. Table S2 shows that EAG maintains the best performance on all evaluation metrics compared to spike-history-augmented versions.

Collectively, these results highlight EAG's strong capacity to model both population-level and single-neuron-level patterns and generate biologically realistic spike data, with stable performance across varying data scales and cortical regions.

EAG Samples with High Efficiency

To assess the efficiency of EAG in addition to its generation quality, we conducted a systematic comparison with diffusion-based models. Unlike LDNS, which relies on iterative denoising steps during inference, EAG generates in a single forward pass, resulting in orders-of-magnitude improvements in sampling speed. We trained four LDNS models with diffusion steps ranging from 200 to 1000, and two



(a) Firing rates (b) Decoded trajectory (c) Single-trial neural latents (d) Averaged neural PCs

Figure 4: Generalization to unseen angle labels. (a) real firing rates and sampled firing rates. (b) Decoded trajectory from real rates and sampled rates conditioned on unseen angle labels. (c) Single-trial neural trajectories in latent space extracted from real and sampled activity. Colorbars in (b) and (c) indicate the condition angle $\alpha_{\text{condition}}$. (d) The first 2 principal components averaged over eight reach directions of real and sampled firing rates. For all panels, top: real data; bottom: sampled data.

EAG models with autoregressive steps 16 and 32, and evaluated both generation quality and inference latency. Here, latency refers to the time taken to generate the same number of trials as in the training dataset. (∼2000 trials).

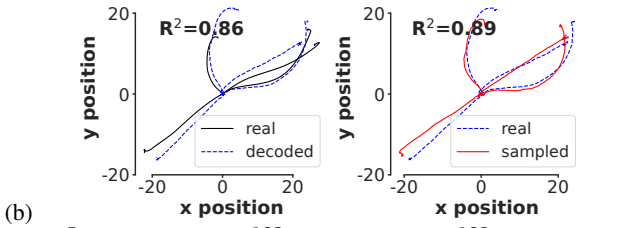
Taking single-neuron generation quality as a representative case, Figure 3 illustrates the relationship between RMSE mean-isi and generation latency across varying steps of LDNS and EAG. In terms of sample latency, EAG-32 generates 2008 trials in just 10.29s while LDNS-1000 requires 330.64s, achieving a 96.9% speed-up. Against the minimal-step LDNS-200, EAG-32 still achieves an 84.4% reduction in latency. In terms of generation quality, even minimal-step EAG-16 outperforms LDNS-1000. Specifically, EAG-32 delivers a 49.0% gain over LDNS-200 and a 32.4% gain over LDNS-1000. Similar trends are also observed across other quality metrics, as detailed in Supp. Figure S8 and Supp. Table S3. Moreover, scaling analysis further shows that EAG's runtime and memory usage remain largely unaffected by increases in neuron count or time length (details in Supp.Figure S4). These results demonstrate that energy-based autoregressive generation achieves higher efficiency than diffusion-based methods while maintaining better quality and stable to scaled datasets.

EAG Generalizes to Unseen Contexts

EAG's modeling capability generalizes effectively to unseen behavioral contexts. To systematically evaluate this property, we condition EAG on behavior variables not observed during training and assess whether the model can generate realistic neural activity consistent with the given context.

We first condition EAG on the **monkey's initial reach direction** α_{init}. Figure 4a shows real rates (top panel) and generated rates (bottom panel) conditioned at a novel angles ($\sim$ 60°), which align closely (additional samples in Supp. Figure S10). The generated activity is not a direct

(a)



(b)

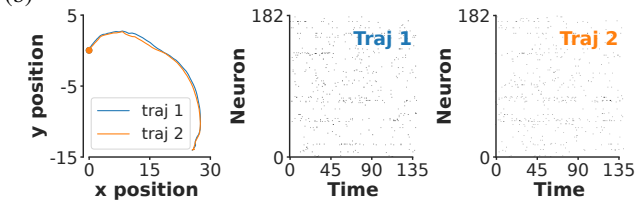


Figure 5: Generalization to unseen velocity labels. (a) Real hand trajectory and decoded trajectory (left panel), decoded trajectory from real rates and sampled rates (right panel). (b) Sampled spike trains conditioned on two nearly identical trajectories show trial-to-trial variability.

copy but shows natural trial-to-trial variability. EAG's ability to model variability is further examined in velocity-conditioned generation. Next, using the pipeline described in Supp. Figure S9, we decode trajectories via a pretrained ridge regression model from EAG-sampled rates conditioned on unseen labels. As shown in Figure 4b, the decoded (bottom) and true trajectories (top) closely match. To verify that EAG effectively captures the underlying mechanisms of neural-activity relations, we visualize real and sampled single-trial latents from the EAG-extracted latent space (Figure 4c). Despite unseen labels, latent distributions align with angles, matching real data patterns. PCA on firing rates, averaged over eight directions, further reveals highly

Method	Co-smoothing bps(↑)	Behavior decoding(↑)	PSTH R^2(↑)
GRU	0.267/**0.270**	0.868/**0.880**	0.507/**0.533**
SLDS	0.218/**0.233**	0.794/**0.801**	0.478/**0.492**
LFADS	0.324/**0.347**	0.901/**0.906**	0.579/**0.594**
NDT	0.272/**0.305**	0.776/**0.838**	0.546/**0.581**
AutoLFADS	0.347/**0.350**	0.906/**0.911**	0.594 /**0.607**
NDT(ray)	0.335/**0.363**	0.875/**0.904**	0.589/**0.596**

Table 2: Metrics before and after EAG augmentation on MC_Maze. Metrics follow the NLB evaluation and are formatted as before/after(improvement ratio). Bold indicates improvement after EAG augmentation.

similar real and generated trajectories (Figure 4d). These results demonstrate EAG's capacity to model latent neural–behavioral structure, enabling generalization to unseen conditions.

Then we consider a more fine-grained behavioral label by conditioning EAG on the hand velocity (v_x, v_y) at each time point. EAG produces firing rates that closely resemble the real ones when conditioned on an entirely unseen velocity trajectory, as shown in Supp. Figure S11. Figure 5a shows the ridge model bias (left panel) when decoding from real rates, while the right panel shows trajectories decoded from EAG-sampled rates, achieving a higher similarity with the ground truth ($R^2 = 0.89$) compared to LDNS ($R^2 = 0.65$). Notably, even for two trajectories that are virtually identical, EAG-sampled spike trains retain strong trial-to-trial variability (Figure 5b). This ability to generate realistic neural activity from hypothetical movements while preserving variability are vital for downstream applications such as BCI decoding (Wen et al. 2023; Degenhart et al. 2020).

EAG Enhances BCI Performance

To assess the practical impact of EAG-generated data, we investigate its effectiveness in improving BCI decoding performance through data augmentation. Specifically, using two Neural Latent Benchmark datasets, MC_Maze and Area2_Bump, we evaluate changes in decoding accuracy for multiple baseline models when trained with and without onefold EAG-based augmentation. As shown in Table 2, EAG consistently improves decoding on MC_Maze, with larger gains in more complex models. In particular, the transformer-based Neural Data Transformer (NDT) exhibits the most pronounced gain, reaching up to 12.1%. To ensure that these improvements are not artifacts of suboptimal training, we further apply ray-based hyperparameter optimization for both LFADS and NDT (autoLFADS and NDT-ray), and compare performance with and without augmentation. The improvements are reduced but persist, validating effectiveness and the robustness of EAG's augmentation benefits.

Similar results are observed on the smaller Area2_Bump dataset (Supp. Table S5). Given its limited size relative to MC_Maze, we hypothesize that larger scale augmentation brings more decoding gains. We test the effect of scaling EAG augmentation to 1×, 2×, and 4×, using NDT as the decoder. As shown in Table 3, accuracy rises as data scale increased, with the largest jump yielding up to a 54.7% improvement at 2× augmentation in NDT. This confirms that, for small datasets, data scarcity represents a more pressing bottleneck in decoding tasks, and increased augmentation yields greater gains.

Method	Co-smoothing bps(↑)	Behavior decoding(↑)	PSTH R^2(↑)
NDT-2size	0.106/**0.164**	0.512/**0.621**	0.321/**0.422**
NDT-4size	0.106/**0.161**	0.512/**0.631**	0.321/**0.425**
NDT(ray)-2size	0.208/**0.227**	0.794/**0.854**	0.414/**0.499**
NDT(ray)-4size	0.208/**0.228**	0.794/**0.834**	0.414/**0.509**

Table 3: Metrics of multi-scale EAG augmentation on Area2_Bump. Evaluation of 2× and 4× multi-scale augmentation using NDT and NDT(ray) decoders.

α	D_{KL} psch	RMSE pairwise corr	RMSE mean isi	RMSE std isi
1.0	0.0014 ± 2.0e-4	0.0024 ± 1.0e-5	0.024 ± 0.001	0.018 ± 0.0024
1.25	0.0026 ± 2.3e-4	0.0024 ± 1.0e-5	0.033 ± 0.007	0.018 ± 0.0007
1.5	0.0013 ± 9.8e-5	0.0025 ± 9.0e-6	0.022 ± 0.001	0.018 ± 0.0018
1.75	0.0017 ± 1.1e-4	0.0024 ± 1.2e-5	0.026 ± 0.001	0.019 ± 0.0007
2.0	0.0541 ± 7.8e-4	0.0028 ± 1.1e-5	0.051 ± 0.004	0.027 ± 0.0007

Table 4: Ablations on strict propriety. Results of varying the exponential coefficient α in the energy loss, highlighting importance of strict propriety.

Ablation Studies

Previous experiments typically set the exponential coefficient α to 1 in energy loss (Equation 7). However, the energy score remains strictly proper for all $\alpha \in (0, 2)$. Since setting $\alpha < 1$ can lead to instability during early training due to unbounded gradients, we focus on the range $\alpha \in [1, 2)$. Notably, when $\alpha = 2$, the energy score is still proper while not strictly proper. The looser constraint ($\mathbb{E}_{p_\theta}[\mathbf{z}] = \mathbb{E}_q[\mathbf{z}_{\text{data}}]$) in this case is insufficient to effectively guide the model to accurately capture realistic neural dynamics. As demonstrated in Table 4, models trained with $\alpha = 2$ exhibit clear deficiencies in generating realistic spiking activity, whereas models with $\alpha \in [1, 2)$ perform consistently well with only minor differences. Based on these results, we adopt $\alpha = 1$ as our default setting throughout all experiments.

Conclusion

We developed a novel Energy-based Autoregressive Generation (EAG) framework that resolves the fundamental tradeoff between computational efficiency and high-fidelity modeling through latent energy-based learning. EAG achieves state-of-the-art generation quality with substantial computational efficiency improvements over existing methods, particularly diffusion-based approaches. Conditional generation applications demonstrate generalization to unseen behavioral contexts and improvement of motor BCI decoding accuracy using generated neural data. These results establish that neural population dynamics can be effectively modeled through direct energy-based generation in latent space, eliminating computationally expensive iterative sampling procedures. This framework provides a foundation for applications requiring both computational efficiency and biological realism in neural population modeling.

Acknowledgments

This work was supported by the Lingang Laboratory, Grant No.LGL-1987 and the Strategic Priority Research Program of Chinese Academy of Sciences (XDB1010302).

References

Azabou, M.; Pan, K. X.; Arora, V.; Knight, I. J.; Dyer, E. L.; and Richards, B. A. 2024. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *The Thirteenth International Conference on Learning Representations*.

Bakhtin, A.; Deng, Y.; Gross, S.; Ott, M.; Ranzato, M.; and Szlam, A. 2021. Residual energy-based models for text. *Journal of Machine Learning Research*, 22(40): 1–41.

Bashivan, P.; Kar, K.; and DiCarlo, J. J. 2019. Neural population control via deep image synthesis. *Science*, 364(6439): eaav9436.

Brier, G. W. 1950. The statistical theory of turbulence and the problem of diffusion in the atmosphere. *Journal of Atmospheric Sciences*, 7(4): 283–290.

Churchland, M.; and Kaufman, M. 2022. MC_Maze: macaque primary motor and dorsal premotor cortex spiking activity during delayed reaching. *Data set*.

Churchland, M. M.; Cunningham, J. P.; Kaufman, M. T.; Foster, J. D.; Nuyujukian, P.; Ryu, S. I.; and Shenoy, K. V. 2012. Neural population dynamics during reaching. *Nature*, 487(7405): 51–56.

Churchland, M. M.; Yu, B. M.; Cunningham, J. P.; Sugrue, L. P.; Cohen, M. R.; Corrado, G. S.; Newsome, W. T.; Clark, A. M.; Hosseini, P.; Scott, B. B.; et al. 2010. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3): 369–378.

Degenhart, A. D.; Bishop, W. E.; Oby, E. R.; Tyler-Kabara, E. C.; Chase, S. M.; Batista, A. P.; and Yu, B. M. 2020. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nature biomedical engineering*, 4(7): 672–685.

Ecker, A. S.; Berens, P.; Cotton, R. J.; Subramaniyan, M.; Denfield, G. H.; Cadwell, C. R.; Smirnakis, S. M.; Bethge, M.; and Tolias, A. S. 2014. State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1): 235–248.

Gallego, J. A.; Perich, M. G.; Chowdhury, R. H.; Solla, S. A.; and Miller, L. E. 2020. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature neuroscience*, 23(2): 260–270.

Gallego, J. A.; Perich, M. G.; Miller, L. E.; and Solla, S. A. 2017. Neural manifolds for the control of movement. *Neuron*, 94(5): 978–984.

Good, I. J. 1952. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1): 107–114.

Hinton, G. E. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8): 1771–1800.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hochberg, L. R.; Bacher, D.; Jarosiewicz, B.; Masse, N. Y.; Simeral, J. D.; Vogel, J.; Haddadin, S.; Liu, J.; Cash, S. S.; Van Der Smagt, P.; et al. 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398): 372–375.

Hurwitz, C.; Srivastava, A.; Xu, K.; Jude, J.; Perich, M.; Miller, L.; and Hennig, M. 2021. Targeted neural dynamical modeling. *Advances in Neural Information Processing Systems*, 34: 29379–29392.

Kapoor, J.; Schulz, A.; Vetter, J.; Pei, F.; Gao, R.; and Macke, J. H. 2024. Latent diffusion for neural spiking data. *Advances in Neural Information Processing Systems*, 37: 118119–118154.

Kell, A. J.; Yamins, D. L.; Shook, E. N.; Norman-Haignere, S. V.; and McDermott, J. H. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3): 630–644.

Keshtkaran, M. R.; Sedler, A. R.; Chowdhury, R. H.; Tandon, R.; Basrai, D.; Nguyen, S. L.; Sohn, H.; Jazayeri, M.; Miller, L. E.; and Pandarinath, C. 2022. A large-scale neural network training framework for generalized estimation of single-trial population dynamics. *Nature Methods*, 19(12): 1572–1577.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; Huang, F.; et al. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Little, S.; Pogosyan, A.; Neal, S.; Zavala, B.; Zrinzo, L.; Hariz, M.; Foltynie, T.; Limousin, P.; Ashkan, K.; FitzGerald, J.; et al. 2013. Adaptive deep brain stimulation in advanced Parkinson disease. *Annals of neurology*, 74(3): 449–457.

Ma, X.; Rizzoglio, F.; Bodkin, K. L.; Perreault, E.; Miller, L. E.; and Kennedy, A. 2023. Using adversarial networks to extend brain computer interface decoding accuracy over time. *elife*, 12: e84296.

Mante, V.; Sussillo, D.; Shenoy, K. V.; and Newsome, W. T. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474): 78–84.

Marks, T. D.; and Goard, M. J. 2021. Stimulus-dependent representational drift in primary visual cortex. *Nature communications*, 12(1): 5169.

Mathis, M. W.; Rotondo, A. P.; Chang, E. F.; Tolias, A. S.; and Mathis, A. 2024. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21): 5814–5832.

McCart, J. D.; Sedler, A. R.; Versteeg, C.; Mifsud, D.; Rigotti-Thompson, M.; and Pandarinath, C. 2024. Diffusion-Based Generation of Neural Activity from Disentangled Latent Codes. *ArXiv*, arXiv–2407.

Pandarinath, C.; O'Shea, D. J.; Collins, J.; Jozefowicz, R.; Stavisky, S. D.; Kao, J. C.; Trautmann, E. M.; Kaufman,

M. T.; Ryu, S. I.; Hochberg, L. R.; et al. 2018. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10): 805–815.

Panzeri, S.; Moroni, M.; Safaai, H.; and Harvey, C. D. 2022. The structures and functions of correlations in neural population codes. *Nature Reviews Neuroscience*, 23(9): 551–567.

Pei, F.; Ye, J.; Zoltowski, D.; Wu, A.; Chowdhury, R. H.; Sohn, H.; O'Doherty, J. E.; Shenoy, K. V.; Kaufman, M. T.; Churchland, M.; et al. 2021. Neural latents benchmark'21: evaluating latent variable models of neural population activity. *arXiv preprint arXiv:2109.04463*.

Rigotti, M.; Barak, O.; Warden, M. R.; Wang, X.-J.; Daw, N. D.; Miller, E. K.; and Fusi, S. 2013. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451): 585–590.

Roby, T. B. 1965. Belief states and the uses of evidence. *Behavioral science*, 10(3): 255–270.

Romo, R.; and Salinas, E. 2003. Flutter discrimination: neural codes, perception, memory and decision making. *Nature Reviews Neuroscience*, 4(3): 203–218.

Safaie, M.; Chang, J. C.; Park, J.; Miller, L. E.; Dudman, J. T.; Perich, M. G.; and Gallego, J. A. 2023. Preserved neural dynamics across animals performing similar behaviour. *Nature*, 623(7988): 765–771.

Satorras, V. G.; Hoogeboom, E.; and Welling, M. 2021. E (n) equivariant graph neural networks. In *International conference on machine learning*, 9323–9332. PMLR.

Sedler, A. R.; and Pandarinath, C. 2023. lfads-torch: A modular and extensible implementation of latent factor analysis via dynamical systems. *arXiv preprint arXiv:2309.01230*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Stringer, C.; Michaelos, M.; Tsyboulski, D.; Lindo, S. E.; and Pachitariu, M. 2021. High-precision coding in visual cortex. *Cell*, 184(10): 2767–2778.

Sussillo, D.; Churchland, M. M.; Kaufman, M. T.; and Shenoy, K. V. 2015. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, 18(7): 1025–1033.

Sussillo, D.; Jozefowicz, R.; Abbott, L.; and Pandarinath, C. 2016. Lfads-latent factor analysis via dynamical systems. *arXiv preprint arXiv:1608.06315*.

Székely, G. J. 2003. E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05): 1–18.

Vargas, A. M.; Bisi, A.; Chiappa, A. S.; Versteeg, C.; Miller, L. E.; and Mathis, A. 2024. Task-driven neural network models predict neural dynamics of proprioception. *Cell*, 187(7): 1745–1761.

Vyas, S.; O'Shea, D. J.; Ryu, S. I.; and Shenoy, K. V. 2020. Causal role of motor preparation during error-driven learning. *Neuron*, 106(2): 329–339.

Walker, E. Y.; Sinz, F. H.; Cobos, E.; Muhammad, T.; Froudarakis, E.; Fahey, P. G.; Ecker, A. S.; Reimer, J.; Pitkow, X.; and Tolias, A. S. 2019. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12): 2060–2065.

Wang, E. Y.; Fahey, P. G.; Ding, Z.; Papadopoulos, S.; Ponder, K.; Weis, M. A.; Chang, A.; Muhammad, T.; Patel, S.; Ding, Z.; et al. 2025. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640(8058): 470–477.

Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; et al. 2023. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976): 1089–1100.

Wen, S.; Yin, A.; Furlanello, T.; Perich, M. G.; Miller, L. E.; and Itti, L. 2023. Rapid adaptation of brain–computer interfaces to new neuronal ensembles or participants via generative modelling. *Nature biomedical engineering*, 7(4): 546–558.

Willett, F. R.; Kunz, E. M.; Fan, C.; Avansino, D. T.; Wilson, G. H.; Choi, E. Y.; Kamdar, F.; Glasser, M. F.; Hochberg, L. R.; Druckmann, S.; et al. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976): 1031–1036.

Yamins, D. L.; and DiCarlo, J. J. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3): 356–365.

Ye, J.; Collinger, J.; Wehbe, L.; and Gaunt, R. 2023. Neural data transformer 2: multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36: 80352–80374.

Yoshida, T.; and Ohki, K. 2020. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature communications*, 11(1): 872.

Zhou, D.; and Wei, X.-X. 2020. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. *Advances in neural information processing systems*, 33: 7234–7247.

Zhu, Y.; Song, C.; Ouyang, W.; Yu, S.; and Huang, T. 2025. Neural Representational Consistency Emerges from Probabilistic Neural-Behavioral Representation Alignment. *arXiv preprint arXiv:2505.04331*.