A DISCRETE GENERATIVE MODEL OF NEURONAL SPIKING ACTIVITY ON MICROELECTRODE ARRAYS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models of neural activity could help characterize tissue dynamics, compare experimental conditions, and simulate population activity for applications ranging from disease and drug-response studies to closed-loop experimentation. Existing approaches, however, typically assume a fixed set of sorted neurons, whereas high-density microelectrode arrays produce extremely sparse, array-wide binary spike volumes in which the observed subset of electrodes varies across assays. We introduce a discrete generative model that represents this activity using a shared vocabulary of spatiotemporal motifs. A residual vector-quantized autoencoder learns the motif vocabulary, while a factorized masked transformer predicts where activity occurs and which motif appears at each active location. We evaluate the model on 31 assays spanning human brain organoids and acute *ex vivo* human hippocampal tissue. The learned motifs are broadly reused: assay identity explains only 9% of the entropy in motif use, and motif overlap across tissue types is comparable to overlap within them. When representation quality is evaluated independently of the generative prior, our approach achieves $5.2\times$ the voxel-level reconstruction average precision of a matched flat tokenizer. For masked completion and free generation, the full model achieves 1.4–$2.6\times$ the site-level average precision of the matched generative baseline and outperforms it across all four families of generation metrics. These results establish a compact, reusable representation for array-wide spiking activity without learned assay-specific parameters, providing a scalable foundation for generative modeling across diverse neural preparations.

1 INTRODUCTION

Cultured neural tissue on high-density microelectrode arrays is being used in closed-loop tasks (Kagan et al., 2022; Smirnova et al., 2023; Patel et al., 2025; Robbins et al., 2026). A *generative forward model* of spontaneous activity could support two uses: simulate the preparation, and give the unperturbed baseline against which stimulus-evoked change is measured. Controller design needs stimulation-conditioned dynamics, absent from these recordings (Section 6). Organoids develop rich spontaneous population dynamics (Trujillo et al., 2019; Sharf et al., 2022), and HD-CMOS arrays record them at single-electrode resolution across tens of thousands of sites (Ballini et al., 2014). Existing generative models of spiking activity work on sorted units (Kapoor et al., 2024; Minnick et al., 2026a); we sort too, but write each unit to its peak electrode and model the array-wide volume with no per-unit parameter. The activity is not uniform noise; recurring population events have structured spatial and temporal organization (Beggs & Plenz, 2003; van der Molen et al., 2026). This motivates asking whether such events can be represented by a shared motif vocabulary.

Both preparations show structured population activity (Sharf et al., 2022; van der Molen et al., 2026; Andrews et al., 2024a). Per-assay learned tables can conflate reusable structure with assay-specific memorization: removing assay identity strongly degrades our statistical references (Section 5.3). We aim instead for a shared model with no learned per-assay parameters.

Representing it is hard because of sparsity, not dimensionality. A clip is a $48 \times 120 \times 224$ binary volume with voxel occupancy 1.62×10^{-4}, and only 841–1,020 of the array's $26{,}400$ sites are routed in any assay, with a different subset each time. At this density the structure is *which electrodes*

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.

participate together, not image-like intensity. We therefore evaluate both site-level accuracy (which electrodes are active) and voxel-level accuracy (which electrode is active in which frame).

Contributions.

1. **A motif alphabet for sparse spike volumes.** A three-level residual VQ-VAE learns spatiotemporal patches; its residual paths form one deduplicated categorical alphabet, with a dedicated blank token for empty patches. When both tokenizers are given the ground-truth codes for a held-out region, ours achieves $5.2\times$ the voxel-level reconstruction AP of a flat-tokenizer baseline with the same grid and patch size (Section 5.1).
2. **Evidence that motifs are reused across observed assays.** Assay identity explains only 9% of code entropy, and vocabulary overlap across preparation types is similar in magnitude to overlap within them (Section 5.2). This analysis is possible because the corpus spans cultured organoid tissue and acute *ex vivo* human hippocampus.
3. **A factorized prior with controlled comparisons.** The prior predicts first where activity occurs and then which motif occupies each active location. It is conditioned on a fixed per-assay code but adds no learned assay-specific parameter table, so assay identity remains available while the number of learned parameters remains independent of the number of assays. We compare it with a same-grid flat-tokenizer baseline, directly supervised convolutional models, and assay-specific statistical references. Oracle decoding identifies temporal prediction as the principal bottleneck (80–98% of site-level versus 4–6% of voxel-level representational capacity) (Section 5).

2 Related work

The dichotomized Gaussian (Macke et al., 2009) and the coupled point-process GLM (Pillow et al., 2008; Truccolo et al., 2005) are the standard forward models for binary population activity, and both are fitted *per assay*: their parameters are a rate vector and a coupling or covariance matrix over that assay's electrodes. Those spatial parameters scale with the number of assays, so we include both as assay-specific statistical references and not as parameter-matched peers (Section 5.3). Sparsity limits what can be fitted, so each uses the stationary parameterization its literature prescribes (Appendix P).

Our own construction borrows from discrete generative modeling of images and video. Vector-quantized autoencoders (van den Oord et al., 2017; Razavi et al., 2019; Esser et al., 2021) turn continuous signals into token grids that an autoregressive or masked prior can model, and MaskGIT's parallel iterative unmasking (Chang et al., 2022) extended this to video in MAGVIT (Yu et al., 2023; 2024). Residual quantization is prior art from image and audio coding (Lee et al., 2022; Zeghidour et al., 2022); we use it without claiming it as a contribution. Two adaptations make the residual ladder usable by a masked prior: a sparse encoder for a canvas that is 89% empty, and flattening the ladder into one deduplicated categorical alphabet so the prior predicts one symbol per site. Our peer method, **MaskGIT-flat**, is the ordinary version of that construction at a matched token budget, and the oracle-code comparison in Section 5.1 takes the prior out of it.

The tissue and the instrument bring their own literature. Human brain organoids develop oscillatory and avalanche-structured spontaneous activity (Trujillo et al., 2019; Beggs & Plenz, 2003), and HD-CMOS arrays resolve it at single-electrode scale (Ballini et al., 2014). The organoid recordings were produced by Sharf et al. (2022). Later work combined them with further recordings and reported preconfigured, repeating population sequences (van der Molen et al., 2026). That literature characterizes such sequences; we learn an alphabet of them and build a generative model on the two deposits (Sharf, 2023; Andrews et al., 2024b), with spike detection and sorting following standard practice (Buccino et al., 2020).

Closest in aim are the generative models for neural recordings. LFADS (Pandarinath et al., 2018) and transformer successors (Ye & Pandarinath, 2021) infer latent dynamics underlying population activity; Spike-GAN (Molano-Mazon et al., 2018) generates spike trains adversarially, and LDNS (Kapoor et al., 2024) does so with a latent diffusion model conditioned on behavior. All operate on tens to a few hundred *sorted units*. Models built to span sessions exist: POYO (Azabou et al., 2023) tokenizes individual units and NDT2 (Ye et al., 2023) pretrains across sessions and subjects, but both learn a per-unit or per-session embedding, so their parameters still grow with the record-

ings covered. We model the array-level binary volume directly on the 120×224 canvas, with no unit-level parameterization and no learned per-assay table, which is what the reuse measurement in Section 5.2 justifies. A second recent line forecasts population spiking autoregressively at Neuropixels scale: a state-space forecaster trained on next-step spike counts also supports a linear behavioral readout (Minnick et al., 2026b), and a companion benchmark separates the usual aggregate correlation into temporal fidelity, spatial pattern accuracy and magnitude-invariant alignment (Minnick et al., 2026a). Both work on the next bin of sorted-unit counts, while we complete masked regions of a binary array volume, so neither their readout nor their metric transfers directly. A third line generates spikes for a downstream target: Wu et al. (2026) map sorted upstream neurons to firing probabilities for a small number of downstream neurons and sample spikes from those probabilities. Their generator is optimized by behavioral reward rather than direct downstream-spike supervision, although the behavioral decoder supplying that reward is itself trained on recorded downstream activity. Their objective is task-directed communication between two regions. All of this is adjacent work, addressing sorted-unit latent dynamics, population generation, or next-bin and transregional prediction, not the array-wide binary volume.

3 Data and task

We use 31 recordings from two open-access DANDI dandisets: 18 of human brain organoid slices (Sharf, 2023) and 13 of human *ex vivo* hippocampal slices resected during neurosurgery (Andrews et al., 2024b). Together, they contain 166 minutes of recording: 98 minutes from organoid slices and 68 minutes from hippocampal slices, in files that run from two minutes to half an hour (Appendix A). Both datasets were recorded using the same MaxWell Biosystems high-density CMOS array. The array contains $26{,}400$ sites at a pitch of $17.5\,\mu\mathrm{m}$, of which $1{,}024$ can be read simultaneously at 20 kHz. Per assay, 841–1,020 channels are routed, i.e. under 4% of the array, and a different subset each time. The model's canvas is therefore the array footprint itself, 120×220 sites padded to 120×224: routing varies between assays, physical coordinates do not, so one shared model spans every routing configuration. Both preparation types are included because an alphabet learned from cultured organoid tissue cannot be assumed to represent acute resected hippocampal tissue.

We treat each archive file as one *assay* with an assay ID, the unit of conditioning in the model. An assay ID is a dataset identifier and not a biological one: several assays can come from one organoid or one tissue preparation, recorded in different sessions or under different experimental conditions. The 18 organoid assays come from a source dataset of array recordings of six sectioned human brain organoids, L1–L6, grown from one donor-derived induced pluripotent stem cell line. The source dataset includes repeated recordings across developmental ages and drug conditions (Sharf et al., 2022). The conversion to NWB preserves neither the organoid nor the condition behind any assay, and three pairs of organoid assay files hold the same acquisition trace (Appendix A). The 13 slice assays record two slice preparations, from a woman of 52 and a man of 35, whose tissue was removed during temporal lobectomy for drug-refractory epilepsy (Andrews et al., 2024a). All splits are temporal within an assay and none is cross-preparation (Section 6).

Spikes are detected and sorted per assay with SpyKING Circus 2 (Buccino et al., 2020) and each curated unit is written to its peak electrode as a point event. Activity is segmented into 2,133 burst windows of 600 ms and max-pooled into 6 ms frames, so a frame records *whether* an electrode fired and not how many times. We distinguish three durations. The recordings run for 166 minutes, the burst windows cut out of them hold 21 minutes of activity, and the temporal split of those windows inside each recording gives 1069/426/638 windows, which is 10.7/4.3/6.4 minutes of training, validation and test material (Appendix D). A random contiguous span of each window is trimmed to 48 frames.

A clip is a binary volume $X \in \{0,1\}^{T\times H\times W}$ with $T = 48$, $H = 120$, $W = 224$: 1,290,240 voxels at occupancy 1.62×10^{-4}, about 209 spikes in 288 ms. It is cut into $(6, 15, 14)$ patches, producing a token grid of $8 \times 8 \times 16 = 1024$ sites. We write Z for the token field. It factors into an activity field A, which marks which sites are non-empty, and a motif field M_Z, which names what occupies each active one. Conditioning is a per-assay code g_r and a per-clip descriptor $\ell(X)$ (Section 4.3).

A task is a pair (q, M): an index $q \in \{0, 1, 2, 3\}$ and the binary mask M over the grid marking what is hidden. Each task is sampled with probability $\frac{1}{4}$ and evaluated separately. *Free generation* ($q = 0$) hides everything, so it has no visible-volume context and is not a completion task. *Causal*

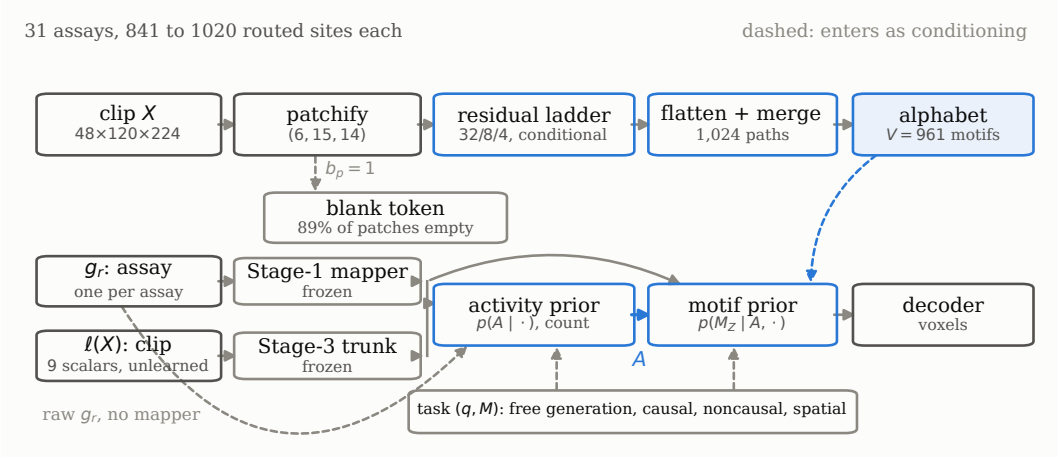

Figure 1: The pipeline. Empty patches bypass the quantizer; the three-level residual ladder is flattened to $V = 961$ motifs. Assay and clip context condition the activity and motif priors by the routes defined in Section 4.3, and the motif prior additionally reads the activity field emitted by the first prior.

completion is forward prediction from a kept prefix of 25–75% of frames. *Noncausal* completion interpolates from both sides, hiding a contiguous 30% of frames. *Spatial* completion imputes an unobserved region of the array, a box covering 25–60% of $H \times W$ at an aspect ratio between 0.5 and 2, hidden across every frame. Holes are snapped to the patch lattice before scoring. The four settings hide different fractions of a clip, so base rates differ and comparisons hold *within* a setting only (Appendix I).

One term is easily confused with free generation. *Reconstruction* refers only to the tokenizer path (encode a true clip, quantize and decode), and it measures what the alphabet can represent, independently of the prior.

4 METHOD

A tokenizer learns a shared alphabet of spatiotemporal motifs, after which a factorized prior models their arrangement (Figure 1). The main text defines each objective and constraint family; coefficients and curricula are in Appendix H.

4.1 A TOKENIZER WITH A ROUTE AROUND THE QUANTIZER

A convolutional stem embeds patches, a 2-layer transformer of width 64 with 4 heads attends over the $8 \times 8 \times 16$ grid, and a transposed-convolution renderer maps each token back to its 1260 voxels. Convolution supplies local feature extraction; global attention can relate co-participating electrodes that are not spatial neighbors. We did not train a matched all-convolutional core, so this choice is not isolated experimentally (Section 6).

At this occupancy, 88.9% of patches contain no spike. The encoder sets $b_p = 1$ for an empty patch and routes it to a dedicated blank embedding outside the content quantizers. In our trained tokenizer 91.7% of tokens take this route; under an otherwise matched dense design, blank patches consume level-1 content codes (Section 5.5). Patch size is likewise chosen for the tokenizer–prior pair: smaller patches reconstruct better but produce a larger, more blank-dominated token grid.

4.2 RESIDUAL HIERARCHY AND FLATTENED ALPHABET

Content patches use a three-level residual ladder (Lee et al., 2022; Zeghidour et al., 2022) over a VQ-VAE (van den Oord et al., 2017). For encoder output h,

$$z_1 = Q_1(h), \quad z_2 = Q_2(h - z_1 \mid z_1), \quad z_3 = Q_3(h - z_1 - z_2 \mid z_1, z_2), \quad \tilde{z} = z_1 + z_2 + z_3, \quad (1)$$

with 32, 8, and 4 entries per level. Each level searches only children of the previously selected path, giving $32 \times 8 \times 4 = 1024$ leaves. Level 1 captures the coarse motif and deeper levels refine its residual; cumulative decoding shows that both refinements improve the rendered spike volume (Section 5.5). Codebooks are EMA buffers rather than parameters updated by reconstruction gradients. Section 5.1 compares this ladder with a flat codebook at the same token budget. When both tokenizers are given their true codes, ours achieves 5.2× the voxel-level reconstruction AP of the flat baseline.

The masked prior requires one categorical variable per token. Each residual path is therefore materialized as the sum of its three code vectors; near-identical sums are merged, leaving $V = 961$ motifs (Appendix E). The tokenizer objective is

$$\mathcal{L}_{\text{tok}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{VQ}} + \lambda_c \mathcal{L}_{\text{ctx}} + \lambda_f \mathcal{L}_{\text{field}} + \lambda_s \mathcal{L}_{\text{spatial}} + \lambda_g \mathcal{L}_{\text{gap}} + \mathcal{R}_{\text{enc/blank}}. \tag{2}$$

Here $\mathcal{L}_{\text{rec}}$ is tolerance-aware sparse-spike reconstruction and $\mathcal{L}_{\text{VQ}}$ is commitment/usage regularization. The remaining terms match clip/patch descriptors, assay support and same-site short-gap statistics; $\mathcal{R}_{\text{enc/blank}}$ collects code-norm, blank-separation and encoder anti-collapse constraints. Appendix H defines every constituent term.

4.3 Global and local conditioning

Conditioning enters at two scales. Each assay ID has a fixed, seeded, L_2-normalized 64-D ± 1 code g_r, a dataset-level identifier and not a measured biological descriptor or learned embedding table. Stage 1 learns a 32-D representation $c_g = f_g(g_r)$ by predicting the assay's running union of active sites, minimizing $\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{supp}}^{\text{token}} + \mathcal{L}_{\text{supp}}^{\text{voxel}} + \lambda_{\text{sep}} \mathcal{L}_{\text{sep}} + \lambda_{\text{adj}} \mathcal{L}_{\text{adj}}$. The support terms cover the measured support while penalizing mass outside it; separation distinguishes assays whose support maps differ, and adjacency matches assay-level short-gap rates. The mapper is then frozen. The motif prior receives c_g, whereas the activity prior uses a learned projection of the raw g_r.

The local descriptor $\ell(X)$ contains nine unlearned scalars: log mean firing density; mass-weighted second moments in x, y, t; their three cross-terms; active site ratio; and a temporal trend score. On completion tasks it is computed from the whole clip, hidden region included, and handed over as a prompt: those tasks measure context-controlled completion, not prediction from visible activity alone. Stage 3 learns $c_\ell = f_\ell(\ell(X))$ with $\mathcal{L}_{\text{local}} = \text{NLL}_{\text{texton}} + \text{NLL}_{\text{flat}} + \mathcal{L}_{\text{summary}}$, matching a 128-texton distribution, the $V = 961$ flat-code histogram and per-code temporal summaries. The texton basis is built on level 1 because the ablation shows that these low-dimensional descriptors predict coarse motif identity better than the residual refinements (Section 5.5). After Stage 3, c_ℓ is frozen and conditions both priors. Neither context changes the VQ lookup itself: during tokenizer training it acts through decoded-output constraints, and during generation it conditions selection over the frozen alphabet.

4.4 A prior factorized into where and what

A single categorical head would have to model the overwhelming blank mass and the diversity of active motifs simultaneously. We instead write

$$p(Z \mid g_r, \ell, q, M) = \underbrace{p(A \mid P_A g_r, c_\ell, q, M)}_{\text{where/how much}} \underbrace{p(M_Z \mid A, c_g, c_\ell, q, M)}_{\text{which motif}}, \tag{3}$$

where P_A is the activity prior's learned projection of the raw assay code, $c_g = f_g(g_r)$ and $c_\ell = f_\ell(\ell(X))$. Both factors are MaskGIT transformers (Chang et al., 2022; Yu et al., 2023) with width 128, four layers, and four heads.

The **activity prior** emits a Bernoulli logit for each of the 1024 token cells plus a categorical head for total active count and reads the ROI occupancy of 16 coarse regions. Its objective is

$$\mathcal{L}_A = \mathcal{L}_{\text{cell}} + \lambda_K \mathcal{L}_{\text{count}} + \lambda_t \mathcal{L}_{t\text{-coact}} + \lambda_x \mathcal{L}_{x\text{-coact}} + \lambda_{\text{sup}} \mathcal{L}_{\text{support}}. \tag{4}$$

This separates the two sparse-data failure modes. Aggressive class weighting over-produces activity, while insufficient positive pressure collapses toward blank. Our per-cell BCE is unweighted, and the count, co-activation and support terms supply calibration and structure.

stage	what it trains	supervision target	frozen output used later
1	global embedder + spatial-map head	the assay's running union of active sites	frozen code-to-support mapper
2A	encoder, decoder, EMA codebooks	the clip itself, reconstructed through the quantizer	frozen motif alphabet, $V = 961$
3	lct trunk + heads (tokenizer frozen)	the clip's texton usage histogram	frozen nine-scalar-to-code mapper
4A	motif prior (tokenizer and mappers frozen)	true motifs, given the *true* activity field	motif generator
4B	activity prior	the true activity field and its total count	activity generator
4C	motif prior (activity prior frozen)	true motifs, given the activity field 4B *emits*	the prior used at inference

Checkpoint selection: 1 on spatial-map loss; 2A on val exact AUPRC; 3 on val multinomial NLL; 4A on val MRR; 4B on val NLL; 4C on val MRR.

Table 1: Training chronology. Each stage is frozen before the next begins. The supervision column identifies what each stage learns; full schedules and checkpoint criteria are in Appendix F.

The **motif prior** iteratively places one of the $V = 961$ motifs at each active cell. Its objective is

$$\mathcal{L}_M = \mathcal{L}_{\mathrm{CE}} + \lambda_n \mathcal{L}_{\mathrm{nbr}} + \lambda_d \mathcal{L}_{\mathrm{dist}} + \lambda_c \mathcal{L}_{\mathrm{ctx}} + \lambda_f \mathcal{L}_{\mathrm{field}} + \lambda_a \mathcal{L}_{\mathrm{adj}} + \lambda_s \mathcal{L}_{\mathrm{spatial}}. \tag{5}$$

Exact CE selects the target motif; neighborhood CE and expected code distance give graded credit to nearby entries in the frozen alphabet. The remaining terms soft-decode logits through the frozen tokenizer so clip/patch context, short-gap rates and assay support constrain the generated voxel field. Coefficients are per objective, not shared: λ_c, λ_f and λ_s name analogous constraints in Eq. 2 and Eq. 5 but take different values, all of which are in Appendix H.

4.5 Staged training and inference-time adaptation

The final adaptation corrects the mismatch between teacher-forced motif training and inference. Stage 4A learns motifs given the true activity map, whereas at generation the motif prior receives the activity field emitted by Stage 4B, whose cell-level recall is 0.72. A missed active cell is an input configuration absent under teacher forcing, not a noisier version of a seen one. Stage 4C freezes the activity prior and fine-tunes the motif prior on emitted maps. Table 1 gives the order and what each stage is allowed to learn. The emitted field remains soft inside the hole so uncertainty can be marginalized; replacing it with a hard map is worse than no adaptation (Section 5.5).

5 Experiments

Task-axis comparisons use paired Wilcoxon signed-rank tests with Benjamini–Hochberg correction across the family (Benjamini & Hochberg, 1995); we report q values and identify other analyses separately. Each conditioning result is compared with that model's random-context control; equal scores indicate unused conditioning, except on adherence, which scores obedience to the descriptor supplied. Average precision is stepwise, since trapezoidal interpolation is invalid in PR space (Davis & Goadrich, 2006). All results use a fixed assay-balanced sample of the test split: 279 clips, nine per assay across all 31 assays (Appendix I).

Baselines. We compare our method with three learned models and two statistical references. **MaskGIT-flat** is the matched peer, the ordinary construction this literature describes: a fully convolutional VQ tokenizer with one flat codebook of 1024 entries, compared with our deduplicated alphabet of 961, on the same grid and patch size at a matched token budget. Core, codebook and prior all differ, so it evaluates the combined design rather than any single component (Appendix P). **3D U-Net** is a direct-supervision reference, marked †: an *inpainter* trained on this exact hole distribution, so it cannot reconstruct. **3D CVAE** adds a conditional latent to that backbone, and that latent collapses (Appendix P). The **dichotomized Gaussian** (Macke et al., 2009) and **coupled GLM**

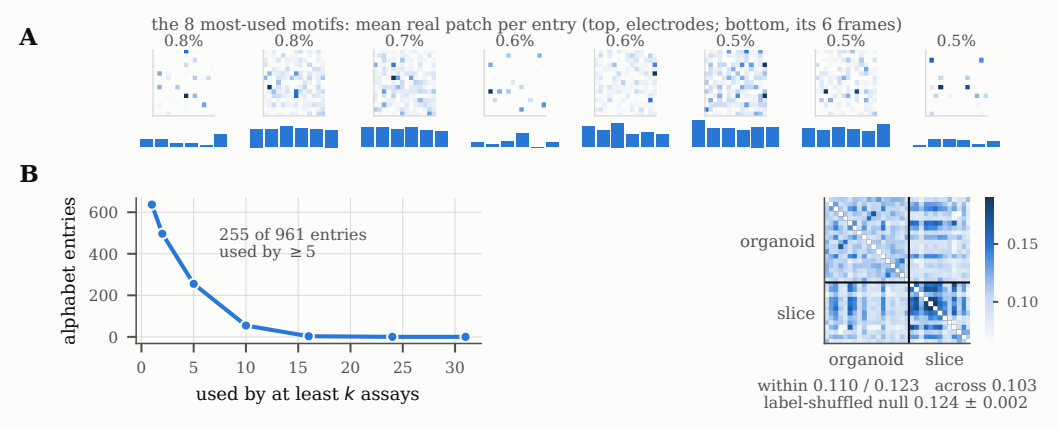

Figure 2: **What a motif represents, and whether it is reused. (A)** The most-used entries, each the mean *real* voxel patch assigned to it, not a decoder rendering; time collapsed above, six frames below. **(B)** Entries used by at least k assays, and pairwise vocabulary overlap by preparation type, both rarefied.

(Pillow et al., 2008; Truccolo et al., 2005) are fitted per assay as memorization references, marked *ref* (Appendix P).

5.1 The alphabet

Figure 2 shows what the entries represent and how broadly they are reused. Reconstruction measures representational capacity and is compared only with the other tokenizer. Ours reaches step-wise AP 0.2564 against MaskGIT-flat's 0.0269. It does so while using 619 of its 961 entries at perplexity 436.9, against 844 entries at 312.2: fewer entries carrying more effective diversity.

Removing the prior isolates the representation. When each tokenizer is given the *true* codes for a hole, AP is 0.2535 against 0.0486, a factor of 5.2. The advantage is not uniform across the two axes: at *site* level ours is slightly *lower*, 0.2296 against 0.2607, so the hierarchical alphabet improves resolution in *time* rather than in space. Section 5.3 shows that this is where the two arms separate.

5.2 Motifs are reused across assays

To test whether the alphabet is truly shared rather than partitioned by assay, we report three statistics in Table S2. 846 of 961 entries are in use across the test split, and 255 of them are used by at least five of the 31 assays at a matched token budget. Knowing which assay a clip came from removes only 9% of the code entropy.

The third statistic is why the corpus spans two preparation types. At matched sampling effort, mean pairwise overlap within organoid assays is 0.1095, within slice assays 0.1232, and *across* the two types 0.1033, so the same motifs appear in cultured organoid tissue and in acute human hippocampal slices. Overlap sits slightly below the label-shuffle null, so a small assay-specific component exists. These results do not establish transfer to unseen assays or preparations (Appendix B).

5.3 Task completion

Full results for each setting and the corresponding per-clip comparisons are in Appendix L and Figure S1. At site level we achieve 1.4–$2.6\times$ the peer's AP in all four settings, winning 64–88% of clips ($q \leq 6.05 \times 10^{-10}$). At voxel level the comparison splits: we win *spatial* ($q = 8.21 \times 10^{-13}$) and *free generation* ($q = 3.79 \times 10^{-4}$), while *causal* and *noncausal* are not significant.

The oracle-code decomposition localizes this. Our prior recovers 80–98% of its alphabet's site-level reference but only 4–6% of its voxel-level reference, MaskGIT-flat 44–57% and 16–30%. Our binding constraint is the prior's timing, theirs the flat alphabet.

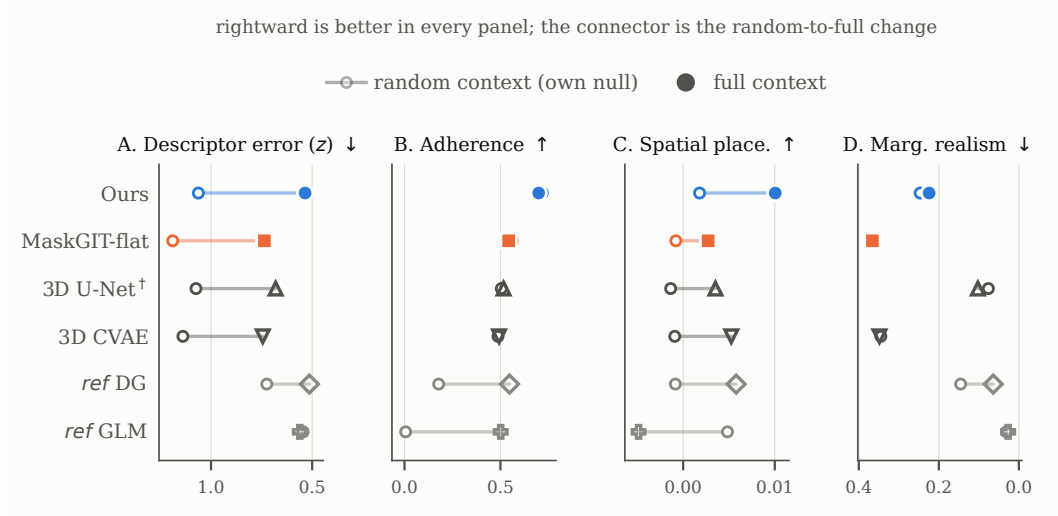

Figure 3: **Effect of conditioning for each model.** Hollow marker, that model's own random-context null; filled marker, full context; the connector is the random-to-full change. Coinciding markers mean unused conditioning in A, C and D; B scores obedience to the descriptor supplied, so a flat B is expected (Table S7). Panels are oriented so rightward is better and have independent x axes. All five context rungs appear in Table S7.

family	Ours	MaskGIT-flat	3D U-Net†	3D CVAE	*ref* DG	*ref* GLM
A. Descriptor error (z) ↓	0.5350 *[1.062]*	0.7362 *[1.189]*	0.6803 *[1.074]*	0.7441 *[1.139]*	0.5137 *[0.724]*	0.5610 *[0.545]*
B. Adherence ↑	0.6992 *[0.718]*	0.5429 *[0.561]*	0.5165 *[0.505]*	0.4925 *[0.487]*	0.5472 *[0.178]*	0.5012 *[0.006]*
C. Spatial placement, lookup-proof ↑	0.0101 *[0.002]*	0.0027 *[-0.001]*	0.0035 *[-0.001]*	0.0052 *[-0.001]*	0.0058 *[-0.001]*	-0.0049 *[0.005]*
D. Marginal realism ↓	0.2242 *[0.247]*	0.3661 *[0.368]*	0.1025 *[0.076]*	0.3484 *[0.345]*	0.0640 *[0.145]*	0.0268 *[0.031]*

Table 2: The four families at full context, with the corresponding random-context value shown in brackets.

Two reference classes outperform us. A static per-assay site map beats all four learned models at both levels in every setting ($q \leq 3.1 \times 10^{-10}$), despite carrying no clip-specific information. Its site-level AP falls from 0.6796 to 0.0284 when the clip's assay is withheld, as do the two lookup references. Both directly supervised convolutional models also beat us on voxel AP in all four settings, the U-Net by the largest margin, and neither holds a per-assay table. Direct supervision remains stronger for voxel-level imputation; our contribution is a reusable discrete representation, which an inpainter with an uncompressed skip path does not produce. Every arm above ours across the 32 paired comparisons is in one of these two classes, and none is the matched peer (Appendix P).

5.4 Generation

Generation is scored on four families, each per clip and paired (Figure 3, Table 2). **A, descriptor error**: z-scored MAE between the local descriptor recomputed from the sample and the true clip's, so 1.0 is the error of an unrelated clip. **B, adherence**: agreement with the descriptor the model was *given*, not the true one. **C, spatial placement**: map correlation against the clip's own electrodes minus the same map scored against a different clip of the same assay: the component a fixed site map cannot reproduce. **D, marginal realism**: relative error on the canonical short-gap rates. Average precision has no place in this table: it ranks a probability map, so the conditional mean maximizes it, and the site map of Section 5.3 duly outranks every learned arm on it (Figure S1).

We outperform the matched generative peer in all four families and every other learned model in three; the U-Net leads on marginal realism. Family C is the only one a site map cannot reproduce, and there the U-Net reaches 0.0035 against our 0.0101. Context makes the U-Net *worse* on the family it wins, so that win is a pooled statistic and not a conditioning result (Appendix K).

MaskGIT-flat gains more from conditioning: its median improvement is 0.2561 on $848/1116$ clips, versus our 0.2108 on $772/1116$ ($q = 3.6 \times 10^{-37}$). **We therefore make no claim that our conditioning is stronger.**

Our model instead shows stronger per-clip discrimination: its within-assay correlation with the true ROI spike count is an order of magnitude above MaskGIT-flat's, and it alone exceeds the arithmetic control in all three settings. No arm combines discrimination with calibration; our decoded field over-counts (Appendix N).

5.5 Design choices and their controls

Every choice in Section 4 has a control, tabulated in Appendix E and O. The blank route leaves all 32 level-1 parents available for content, where a dense arm at the same seed and schedule uses 25 and shares 13 with blanks (AP 0.0501 against 0.0358). Every ladder level contributes: re-decoding the same codes at cumulative depth adds $+0.0832$ then $+0.1202$ AP on 100% of clips. Patch size is a trade-off and not an optimum (Table S3).

The two conditioning codes are complementary: the clip code better predicts the temporal marginal ($R^2 = 0.9027$ versus 0.6063), whereas the assay code reaches ΔNLL 0.3036 against the clip code's 0.2072; neither predicts the temporal centroid. The texton basis sits on level 1 because $z_1 > z_1{+}z_2 > \text{flat}$ is monotone across all three arms, and the motif prior's median rank is 9 of 961 against the strongest null's 53. Final adaptation raises MRR under emitted activity maps from 0.1898 to 0.2035, against an oracle ceiling of 0.2149, an effect 11.4 times the seed standard deviation. Adapting to hard maps instead gives 0.1427, worse than not adapting at all.

6 Limitations

The corpus is biologically small: 31 assays from at most six organoids and two patients, including three pairs with the same acquisition trace, and no split here is cross-preparation. We therefore claim *scalability*, a parameter cost that stays flat in the number of assays, and not generalization to unseen preparations. Removing the per-assay table is not the absence of memorization, and Section 5.2 shows the alphabet is shared among training assays, not that it transfers (Appendix R).

Two comparison classes favor other methods. The static per-assay site map is the strongest within-assay spatial reference because it stores assay-specific structure (Section 5.3). The matched peer wins several pooled summaries while we win the shape-sensitive terms, and the U-Net wins marginal realism (Table S8). Because pooled statistics are less sensitive to conditioning, our claims rely on paired per-clip tests. The CVAE's latent collapses despite standard countermeasures, leaving the continuous-latent question open (Appendix P).

Token granularity is the main weakness. A $(6, 15, 14)$ patch spans 210 electrodes and 36 ms; the prior reaches 98% of the site-level oracle-code reference but only 5% of the voxel reference, making within-token *timing* the principal failure mode. Two fixes failed (Appendix S), and smaller patches collapse the prior (Appendix R). The decoded field over-produces spikes; our per-clip *discrimination* is the strongest of any arm (Section 5.4), but the two are separate properties and no arm here has both. Finally, our tokenizer pairs a transformer core with a convolutional stem and renderer while the matched peer's is fully convolutional, so this is not a controlled comparison: that arm also changes the codebook structure and the blank route (Appendix R).

7 Conclusion

Discrete motifs are a workable representation for ultra-sparse HD-MEA spike volumes. Learned once over a corpus spanning cultured organoid tissue and acute *ex vivo* human hippocampus, the alphabet achieves $5.2\times$ the voxel-level reconstruction AP of a matched flat tokenizer given the true codes, and is measurably reused across assays. A factorized prior over it achieves 1.4–$2.6\times$ the site-level AP of the matched peer, with no learned per-assay parameter table, while within-token timing remains the dominant failure mode and a static site map the stronger spatial predictor (Section 6). Neither the blank route nor the where/what factorization is specific to spike trains.

ETHICS STATEMENT

No new human or animal data were collected for this work. Both corpora are open-access, de-identified secondary data obtained from the DANDI Archive (Sharf, 2023; Andrews et al., 2024b), and we report the approvals of the studies that generated them.

The *ex vivo* hippocampal recordings were obtained from tissue removed during temporal lobectomy with hippocampectomy for drug-refractory epilepsy. The tissue was resected for a clinically indicated reason and would otherwise have been discarded. The source study reports that informed consent was obtained from patients prior to surgical resection, that there was no participant compensation, and that the work was approved by the University of California, San Francisco institutional review board (Andrews et al., 2024a).

The organoid recordings are from human iPSC-derived brain organoids, which do not constitute human subjects research. The originating study reports that the iPSC line was reprogrammed from skin fibroblasts obtained by biopsy following written informed consent from the donor, under Washington University School of Medicine institutional review board approvals 201104178 and 201306108 (Sharf et al., 2022). That study is the source of the organoid recordings used here and provides the relevant consent information; the later compilation that re-deposits them reports no human ethics approval of its own.

The archive records neither participant identifiers nor, for the organoid recordings, the mapping from assay to preparation; we note in Section 3 that this limits what the corpus can support scientifically, and it also means no re-identification risk is introduced by releasing our per-assay provenance table.

On dual use: a generative forward model of spontaneous activity in cultured and *ex vivo* neural tissue is a research instrument for simulation, null-model construction and closed-loop experiment design. We make no clinical claim, the model is not diagnostic, and nothing here should be read as characterizing the donors whose tissue produced the recordings.

REPRODUCIBILITY STATEMENT

Every number and figure panel is generated from a committed data artifact rather than entered manually. Each reported value is therefore traceable to the run that produced it. The anonymized code release contains the model, the training stages, all six evaluation arms including ours, the evaluation harness, the table and figure generators, and the commands that drive them. Its checkpoint manifest records which checkpoint is shipped for each stage and how it was selected. The anonymized supplement carries the code, every checkpoint the evaluation and generation paths load, and the report artifacts the table and figure generators read, so both generators run from it directly. Absolute paths are replaced by a placeholder, and the exploratory and superseded checkpoints are omitted; the checkpoint manifest lists them all and marks which ship. Evaluation is performed under a fixed protocol over an assay-balanced sample of the test split (279 clips, nine per assay, across all 31 assays), with the task axis averaging 8 Monte Carlo samples per clip under seed 20260822 and the distributional battery drawing one sample per rung. Pinning fixes the clip set, the protocol and the committed artifacts; the battery's draws are unseeded, so its statistics carry the finite-sample variation quantified in Appendix I rather than reproducing bit for bit. The clips are sampled from each assay's held-out windows with replacement, but the draw is seeded, so every arm is scored on identical clips in identical order (Appendix I). Data are the two public dandisets cited above; Appendix D gives the preprocessing chain in full, including which of the two stored spike representations is read.

USE OF AI STATEMENT

Large language model assistance (Claude, Anthropic; ChatGPT, OpenAI) was used throughout this project, largely in an agentic coding setting. Specifically, it assisted with implementation of the model, training stages, baseline methods, evaluation harness, analysis scripts, and figure and table generators. It was also used for initial manuscript drafts and subsequent editing, code review and debugging, methodological and experimental-design critique, literature search and summarization, and preparation of this submission.

The research direction, the hypotheses, the experimental design, and every claim made here are the authors'. No number in this paper was produced by a language model: all results come from the committed artifacts of executed runs, and the generator scripts exist precisely so that no value can be introduced by transcription. The authors verified the results, read the code, and are responsible for the paper's content, including any errors.

REFERENCES

John P. Andrews, Jinghui Geng, Kateryna Voitiuk, Matthew A. T. Elliott, David Shin, Ash Robbins, Alex Spaeth, Albert Wang, Lin Li, Daniel Solis, Matthew G. Keefe, Jessica L. Sevetson, Julio A. Rivera de Jesús, Kevin Donohue, H. Hanh Larson, Drew Ehrlich, Kurtis I. Auguste, Sofie R. Salama, Vikaas Sohal, Tal Sharf, David Haussler, Cathryn Cadwell, David Schaffer, Edward F. Chang, Mircea Teodorescu, and Tomasz J. Nowakowski. Multimodal evaluation of network activity and optogenetic interventions in human hippocampal slices. *Nature Neuroscience*, 27(12): 2487–2499, 2024a. doi: 10.1038/s41593-024-01782-5.

John P. Andrews, Jinghui Geng, Kateryna Voitiuk, Matthew A. T. Elliott, David Shin, Ash Robbins, Alex Spaeth, Albert Wang, Lin Li, Daniel Solis, Matthew G. Keefe, Jessica L. Sevetson, Julio A. Rivera de Jesús, Kevin Donohue, H. Hanh Larson, Drew Ehrlich, Kurtis I. Auguste, Sofie R. Salama, Vikaas Sohal, Tal Sharf, David Haussler, Cathryn Cadwell, David Schaffer, Edward F. Chang, Mircea Teodorescu, and Tomasz J. Nowakowski. Multimodal evaluation of network activity and optogenetic interventions in human hippocampal slices, 2024b. URL https://dandiarchive.org/dandiset/001132. DANDI Archive, dandiset 001132, version 0.241130.1903.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J. Mendelson, Blake Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1948.

Marco Ballini, Jan Müller, Paolo Livi, Yihui Chen, Urs Frey, Alexander Stettler, Amir Shadmani, Vijay Viswam, Ian Lloyd Jones, David Jäckel, Milos Radivojevic, Marta K. Lewandowska, Wei Gong, Michele Fiscella, Douglas J. Bakkum, Flavio Heer, and Andreas Hierlemann. A 1024-channel CMOS microelectrode array with 26,400 electrodes for recording and stimulation of electrogenic cells in vitro. *IEEE Journal of Solid-State Circuits*, 49(11):2705–2719, 2014. doi: 10.1109/JSSC.2014.2359219.

John M. Beggs and Dietmar Plenz. Neuronal avalanches in neocortical circuits. *Journal of Neuroscience*, 23(35):11167–11177, 2003. doi: 10.1523/JNEUROSCI.23-35-11167.2003.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Alessio Paolo Buccino, Cole Lincoln Hurwitz, Samuel Garcia, Jeremy Magland, Joshua H. Siegle, Roger Hurwitz, and Matthias H. Hennig. SpikeInterface, a unified framework for spike sorting. *eLife*, 9:e61834, 2020. doi: 10.7554/eLife.61834.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *ICML*, 2006.

Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Brett J. Kagan, Andy C. Kitchen, Nhi T. Tran, Forough Habibollahi, Moein Khajehnejad, Bradyn J. Parker, Anjali Bhat, Ben Rollo, Adeel Razi, and Karl J. Friston. In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron*, 110(23):3952–3969.e8, 2022. doi: 10.1016/j.neuron.2022.09.001.

Jaivardhan Kapoor, Auguste Schulz, Julius Vetter, Felix Pei, Richard Gao, and Jakob H. Macke. Latent diffusion for neural spiking data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 118119–118154, 2024.

Diederik P. Kingma, Tim Salimans, Rafal Józefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4743–4751, 2016.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022.

Jakob H. Macke, Philipp Berens, Alexander S. Ecker, Andreas S. Tolias, and Matthias Bethge. Generating spike trains with specified correlation coefficients. *Neural Computation*, 21(2):397–423, 2009.

John R. Minnick, Jinghui Geng, Kamran Hussain, Jesus Gonzalez-Ferrer, Ash Robbins, Mohammed A. Mostajo-Radji, David Haussler, Jason K. Eshraghian, and Mircea Teodorescu. SpikeProphecy: A large-scale benchmark for autoregressive neural population forecasting. *arXiv preprint arXiv:2605.12992*, 2026a.

John R. Minnick, Jesus Gonzalez-Ferrer, Kamran Hussain, Jinghui Geng, Ash Robbins, Mohammed A. Mostajo-Radji, David Haussler, Jason Eshraghian, and Mircea Teodorescu. Implicit behavioral decoding from next-step spike forecasts at population scale. *arXiv preprint arXiv:2605.12999*, 2026b.

Manuel Molano-Mazon, Arno Onken, Eugenio Piasini, and Stefano Panzeri. Synthesizing realistic neural population activity patterns using generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, L. F. Abbott, and David Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15:805–815, 2018. doi: 10.1038/s41592-018-0109-9.

Dhruvik Patel, Md Sayed Tanveer, Jesus Gonzalez-Ferrer, Alon Loeffler, Brett J. Kagan, Mohammed A. Mostajo-Radji, and Ge Wang. A computational perspective on NeuroAI and synthetic biological intelligence. *arXiv preprint arXiv:2509.23896*, 2025. doi: 10.48550/arXiv.2509.23896.

Jonathan W. Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M. Litke, E. J. Chichilnisky, and Eero P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454:995–999, 2008.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, 2019.

Ash Robbins, Hunter E. Schweiger, Sebastian Hernandez, Alex Spaeth, Kateryna Voitiuk, David F. Parks, Tjitse van der Molen, Jinghui Geng, Isabel Cline, Kenneth S. Kosik, Sofie R. Salama, Tal Sharf, Mohammed A. Mostajo-Radji, David Haussler, and Mircea Teodorescu. Goal-directed learning in cortical organoids. *Cell Reports*, 45(2):116984, 2026. doi: 10.1016/j.celrep.2026.116984.

Tal Sharf. Intrinsic activity in human brain organoid slice, 2023. URL https://dandiarchive.org/dandiset/000732. DANDI Archive, dandiset 000732, the converted organoid deposit used here; the deposited files record Tal Sharf as experimenter and the Kosik Lab, UCSB, as the originating laboratory. This deposit no longer resolves in the DANDI archive; the same MaxOne material remains available from the originating study through Dryad, doi:10.25349/D9031Z.

Tal Sharf, Tjitse van der Molen, Stella M. K. Glasauer, Elmer Guzman, Alessio P. Buccino, Gabriel Luna, Zhuowei Cheng, Morgane Audouard, Kamalini G. Ranasinghe, Kiwamu Kudo, Srikantan S. Nagarajan, Kenneth R. Tovar, Linda R. Petzold, Andreas Hierlemann, Paul K. Hansma, and Kenneth S. Kosik. Functional neuronal circuitry and oscillatory dynamics in human brain organoids. *Nature Communications*, 13:4403, 2022. doi: 10.1038/s41467-022-32115-4.

Lena Smirnova, Brian S. Caffo, David H. Gracias, Qi Huang, Itzy E. Morales Pantoja, Bohao Tang, Donald J. Zack, Cynthia A. Berlinicke, J. Lomax Boyd, Timothy D. Harris, Erik C. Johnson, Brett J. Kagan, Jeffrey Kahn, Alysson R. Muotri, Barton L. Paulhamus, Jens C. Schwamborn, Jesse Plotkin, Alexander S. Szalay, Joshua T. Vogelstein, Paul F. Worley, and Thomas Hartung. Organoid intelligence (OI): the new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science*, 1:1017235, 2023. doi: 10.3389/fsci.2023.1017235.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28, pp. 3483–3491, 2015.

Wilson Truccolo, Uri T. Eden, Matthew R. Fellows, John P. Donoghue, and Emery N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93(2):1074–1089, 2005. doi: 10.1152/jn.00697.2004.

Cleber A. Trujillo, Richard Gao, Priscilla D. Negraes, Jing Gu, Justin Buchanan, Sebastian Preissl, Allen Wang, Wei Wu, Gabriel G. Haddad, Isaac A. Chaim, Alain Domissy, Matthieu Vandenberghe, Anna Devor, Gene W. Yeo, Bradley Voytek, and Alysson R. Muotri. Complex oscillatory waves emerging from cortical organoids model early human brain network development. *Cell Stem Cell*, 25(4):558–569.e7, 2019. doi: 10.1016/j.stem.2019.08.002.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Tjitse van der Molen, Alex Spaeth, Mattia Chini, Sebastian Hernandez, Gregory A. Kaurala, Hunter E. Schweiger, Cole Duncan, Sawyer McKenna, Jinghui Geng, Max Lim, Julian Bartram, Tobias Gänswein, Aditya Dendukuri, Zongren Zhang, Jesus Gonzalez-Ferrer, Kiran Bhaskaran-Nair, Aidan L. Morson, Cole R. K. Harder, Linda R. Petzold, Dowlette-Mary Alam El Din, Jason Laird, Maren Schenke, Lena Smirnova, Bradley M. Colquitt, Mohammed A. Mostajo-Radji, Paul K. Hansma, Mircea Teodorescu, Andreas Hierlemann, Keith B. Hengen, Ileana L. Hanganu-Opatz, Kenneth S. Kosik, and Tal Sharf. Preconfigured neuronal firing sequences in human brain organoids. *Nature Neuroscience*, 29(1):123–135, 2026. doi: 10.1038/s41593-025-02111-0. Published version of the study behind dandiset 000732; its Methods report the organoid cohort size this paper cites.

Shenghui Wu, Zhiwei Song, Xiang Zhang, Yifan Huang, Shuhang Chen, Xiang Shen, Jieyuan Tan, Mingdong Li, Ziyi Wang, Yujun Chen, Kai Liu, Dario Farina, Jose C. Principe, and Yiwen Wang. A generative spike prediction model using behavioral reinforcement for re-establishing neural functional connectivity. *Nature Computational Science*, 6(2):179–192, 2026. doi: 10.1038/s43588-025-00915-5.

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with neural data transformers. *Neurons, Behavior, Data analysis, and Theory*, 5(3), 2021. doi: 10.51628/001c.27358.

Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3523.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: Masked generative video transformer. In *CVPR*, 2023.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion: Tokenizer is key to visual generation. In *ICLR*, 2024.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994.

A DATA PROVENANCE

Table S1 lists every assay used, including what the archive does and does not report. It is generated from the provenance artifact in the code release. The recordings contain 166 minutes in total, or 2.8 hours: 98 minutes from organoid slices and 68 minutes from *ex vivo* hippocampal slices. Their durations vary substantially, from 127 s to 1,940 s. Spike extraction retains only the burst windows, which contain 21 minutes of activity across 2,133 windows, split into 10.7/4.3/6.4 minutes for training, validation and testing.

The biological cohort behind those recordings is smaller than the assay count, and the archive documents it only in part. Every file in the organoid dandiset names the same laboratory and the same experimenter, describes itself as intrinsic activity in a human brain organoid slice, and carries a MaxWell device string, so all 18 of them are array recordings of sectioned organoids and none is a whole-organoid probe recording. The study behind that material reports six organoids recorded on such arrays, named L1–L6 and distinguished by array number in the accompanying data deposit, all grown from the single induced pluripotent stem cell line F12442.4 reprogrammed from skin fibroblasts of one consenting tissue donor (Sharf et al., 2022). We therefore describe the organoid material by organoids, cell line and donor, and the hippocampal material by slices and patients. The same source includes repeated recordings of one organoid across developmental time points, control and diazepam conditions, and four-hour interval series, so several assays can come from one organoid. A later study compiles this deposit together with recordings of its own into a larger cohort of eight human organoids (van der Molen et al., 2026); that cohort is not the corpus used here, and we do not describe our assays by it.

The 13 hippocampal assays come from two patients, a woman of 52 and a man of 35, and form two slice preparations: nine assays under the subject prefix `10F` and four under `11G`. The archive files each slice as its own subject, so its subject count is a slice count, and the source study reports a cohort of seven patients (Andrews et al., 2024a).

Identity the conversion did not preserve. The organoid conversion wrote `neuroconv` defaults into every identity field: one subject `sub-U`, one age range covering a human lifetime, one session timestamp shared by all 18 files, a UUID identifier, and a session description saying only that the file was generated automatically. We searched each file for the MaxWell array number, the original raw HDF5 file name, and the experiment group. We searched all metadata fields and the raw bytes of each complete file but found none of them. No assay in Table S1 can therefore be assigned to one of L1–L6, and none can be assigned an experiment group or a drug condition. The routing hash is the only resolution the files retain, and it gives 13 distinct configurations for 18 assays. A shared routing configuration is evidence of a shared array, and therefore of a shared organoid, but not of a shared recording: an organoid measured twice keeps its routing.

Three pairs of assay files hold the same acquisition trace. Comparing the stored signal rather than the metadata shows what the routing hash cannot. In three pairs the acquisition dataset is identical: `obj-1e5gdym` with `obj-7c50zw`, `obj-1cs0t8m` with `obj-at6hmv`, and `obj-1q4jnyp` with `obj-en3uzn`. Decompressed and compared sample by sample over the whole array and the whole recording, each pair agrees on every one of its 3.65–3.67×10^9 values. The two members of a pair are still different files: they differ in size, in whole-file checksum, in the NWB version they declare and in whether the electrode table carries redundant position columns, so each pair is one recording converted twice, not one file deposited twice. Nothing in the archive says why, and we do not infer a cause.

What the model consumes is not identical. The two conversions were spike-sorted separately and sorting is not deterministic, so no burst window in one copy carries the same spike event set as any window in the other. For a typical window the best exact-event agreement with the other copy is a Jaccard index of 0.002–0.004, and the two copies differ by 3–18% in total spike count and by 6–26% in active sites, with correspondingly different short-gap rates and local descriptors. The windows do cover the same recording times, to a median of 4–54 ms and never more than 0.34 s.

Because each copy is split in acquisition order, and the two orderings agree, the same segment falls in the same split in both copies. Over the 95 segments the three pairs share, the split assignment differs for only one segment, which is assigned to validation in one copy and test in the other. No

segment is in the training split of one copy and the test split of the other. We therefore keep all 31 assays. The duplication is a property of the archive rather than of the experiment, the model was trained on 31 assays, and no result has been recomputed on a reduced corpus.

idx	assay	dandiset	prep.	subject	age	sex	chan.	dur. (s)	routing
0	sub-U_ses-20231119T113000_obj-13gb78s	000732	organoid	U	P1D/P100Y	U	1020	180	13207200a2
1	sub-U_ses-20231119T113000_obj-15bpqpa	000732	organoid	U	P1D/P100Y	U	1018	862	7415ef1a7d
2	sub-U_ses-20231119T113000_obj-195c1dk	000732	organoid	U	P1D/P100Y	U	1020	180	cead3bf0a8
3	sub-U_ses-20231119T113000_obj-1cs0t8m†	000732	organoid	U	P1D/P100Y	U	1020	180	173ca3dd33
4	sub-U_ses-20231119T113000_obj-1e5gdym†	000732	organoid	U	P1D/P100Y	U	1020	180	c005c7a3f4
5	sub-U_ses-20231119T113000_obj-1kcinv8	000732	organoid	U	P1D/P100Y	U	841	1940	262898a949
6	sub-U_ses-20231119T113000_obj-1q4jnyp†	000732	organoid	U	P1D/P100Y	U	1014	180	2eb3b59799
8	sub-U_ses-20231119T113000_obj-1yigkds	000732	organoid	U	P1D/P100Y	U	1020	180	dabafefa61
9	sub-U_ses-20231119T113000_obj-1ywxnj8	000732	organoid	U	P1D/P100Y	U	1020	378	d59ef839d8
10	sub-U_ses-20231119T113000_obj-7c50zw†	000732	organoid	U	P1D/P100Y	U	1020	180	c005c7a3f4
11	sub-U_ses-20231119T113000_obj-at6hmv†	000732	organoid	U	P1D/P100Y	U	1020	180	173ca3dd33
12	sub-U_ses-20231119T113000_obj-cjtijx	000732	organoid	U	P1D/P100Y	U	1020	180	13207200a2
13	sub-U_ses-20231119T113000_obj-en3uzn†	000732	organoid	U	P1D/P100Y	U	1014	180	2eb3b59799
14	sub-U_ses-20231119T113000_obj-fvpkkt	000732	organoid	U	P1D/P100Y	U	1020	180	34dc02c543
15	sub-U_ses-20231119T113000_obj-hom2cn	000732	organoid	U	P1D/P100Y	U	1020	180	a2d71f17ba
17	sub-U_ses-20231119T113000_obj-oqeor8	000732	organoid	U	P1D/P100Y	U	1020	180	a86ef3df2c
19	sub-U_ses-20231119T113000_obj-skejyf	000732	organoid	U	P1D/P100Y	U	1020	180	34dc02c543
20	sub-U_ses-20231119T113000_obj-zvgjcc	000732	organoid	U	P1D/P100Y	U	1020	180	bb7f365328
21	sub-10F-0_ses-20240928T165427	001132	slice	10F-0	P52Y	F	887	313	51f17a2f10
22	sub-10F-1_ses-20240928T165018	001132	slice	10F-1	P52Y	F	1009	127	c52a5b44c7
23	sub-10F-2_ses-20240928T165008	001132	slice	10F-2	P52Y	F	1009	213	c52a5b44c7
24	sub-10F-3_ses-20240928T165423	001132	slice	10F-3	P52Y	F	1009	200	c52a5b44c7
25	sub-10F-4_ses-20240928T165229	001132	slice	10F-4	P52Y	F	1009	310	c52a5b44c7
26	sub-10F-5_ses-20240928T165424	001132	slice	10F-5	P52Y	F	1009	320	c52a5b44c7
27	sub-10F-6_ses-20240928T165427	001132	slice	10F-6	P52Y	F	1009	360	c52a5b44c7
28	sub-10F-7_ses-20240928T165426	001132	slice	10F-7	P52Y	F	1009	390	c52a5b44c7
29	sub-10F-8_ses-20240928T165425	001132	slice	10F-8	P52Y	F	1009	390	c52a5b44c7
30	sub-11G-0_ses-20240928T192015	001132	slice	11G-0	P35Y	M	894	200	cc989abd68
31	sub-11G-1_ses-20240928T165430	001132	slice	11G-1	P35Y	M	894	390	cc989abd68
32	sub-11G-2_ses-20240928T165427	001132	slice	11G-2	P35Y	M	894	510	cc989abd68
33	sub-11G-3_ses-20240928T165427	001132	slice	11G-3	P35Y	M	894	360	cc989abd68

Table S1: The 31 assays. *prep.* is the preparation type reported by the source dandiset. *routing* is a hash of the routed electrode set: assays sharing a hash were recorded through the identical channel selection. Note that all `sub-U` rows carry the same subject field, age and session timestamp. These are conversion defaults written by `neuroconv`, not measurements, and they collapse a cohort of at most six organoids into one apparent subject. † marks the six assays that form the three pairs holding the same acquisition trace.

Assays present in the archive but not used. Three further `sub-U` recordings are present in dandiset 000732 and are not in the table: `obj-1wcxx1y` (2187 s), `obj-m6lpfz` (1167 s) and `obj-paunmv` (2044 s). All three are targeted sub-region recordings for which the spike-extraction stage produced no burst windows, so there is nothing to tokenize; they are excluded because the extraction stage produced no files, not because of any selection we made. No recording was dropped on the basis of its activity, its statistics, or how any model performed on it.

B VOCABULARY SHARING

The entropy and overlap statistics in full. Conditioning on assay identity takes the code entropy from 6.2770 nats to 5.7057, a retained fraction of 0.9090 under a Miller–Madow correction. The correction biases against the shared-alphabet conclusion, since plug-in entropy is downward-biased at these sample sizes. Raw, unrarefied overlap tracks assay length at $r = +0.96$ and inverts the ordering between preparation types, which is why every overlap reported in Section 5.2 is at matched sampling effort. Rarefied overlap sits slightly below the label-shuffle null (0.1087 against 0.1243 ± 0.0023, $z = -6.69$), so a small assay-specific component does exist.

Ladder depth, in figures. Re-decoding the same codes at cumulative depth, the third level adds $+0.1202$ AP on 100% of clips ($p = 3.5 \times 10^{-12}$), whereas the summary-statistic analysis attributes only $+0.019$ to it.

Rarefaction removes the strong dependence of overlap on sampling effort. Recordings contribute between 117 and 1756 tokens and raw pairwise overlap tracks that sampling effort at $r = +0.96$, so the unrarefied matrix reports a within-organoid mean of 0.394 against a within-slice mean of 0.221,

quantity	value
alphabet entries in use	846 of 961
used by ≥ 5 assays, rarefied	255
$H(\text{code})$	6.2770 nats
$H(\text{code} \mid \text{assay})$	5.7057 nats
retained fraction	**0.9090**
rarefied overlap, within organoid	0.1095
rarefied overlap, within slice	0.1232
rarefied overlap, across preparation types	0.1033
null label-shuffled	0.1243 ± 0.0023

Table S2: Vocabulary sharing across the 31 assays, measured on the frozen tokenizer over the test split. The blank token is excluded: it is 91.7% of all tokens and every assay emits it, so including it would drive every overlap statistic to ≈ 1.0 while measuring nothing. Overlap is rarefied to a common token budget and compared with a label-shuffle null. Discussed in Section 5.2.

a preparation difference that is entirely a difference in assay length, and one that reverses under matched effort. The reported budget is the largest common budget available for all 31 assays; the ordering is stable across every budget from 40 to 117, and above 117 the confound visibly returns as longer assays begin to dominate.

C Ethics and data use

The ethics statement before the references gives the consent and approval language reported by each source study. Additional detail: dandiset `000732` is the deposit we downloaded and is the version used here. It no longer resolves in the archive, and the same organoid material is distributed by the originating study through Dryad, with a README that names the six organoids and their array numbers (Sharf et al., 2022). Dandiset `001603` is the later compilation that incorporates that material alongside recordings of its own (van der Molen et al., 2026). Dandiset `001132` is deposited at version `0.241130.1903`. Both dandisets we use are marked `dandi:OpenAccess` in their archive metadata. Neither carries participant identifiers. We collected no new human or animal data, ran no experiment on tissue, and had no contact with participants.

D Preprocessing

Spike extraction. Each assay is sorted with SpyKING Circus 2 through SpikeInterface (Buccino et al., 2020): a fourth-order Butterworth bandpass at 300–6000 Hz, a detection neighborhood of $100\,\mu$m, and a waveform window of 2.0 ms before and after the peak. Each curated unit is written to its peak electrode.

Burst-window detection. Continuous recordings are reduced to burst windows before anything else runs. The per-channel binary raster is summed across channels into a population trace, smoothed with a 2000-sample boxcar (100 ms at 20 kHz, edge padded), and its Hilbert envelope is normalized. The dominant frequency of the mean-removed envelope below 5 Hz sets a characteristic burst period, and peaks are taken from the smoothed trace with a minimum separation of 0.75 of that period and a prominence of 20 spikes. Each accepted peak yields the window $[t-0.2\,\text{s},\ t+0.4\,\text{s}]$, i.e. 600 ms with the peak one third of the way in; windows extending past either end of the recording are dropped. This produces the 2,133 windows used here. The extraction runs upstream of this code release, which begins at the stored window arrays; the parameters above are the ones the extraction used.

Choice of stored spike representation. The extraction step writes two arrays per burst window, and reading the wrong one changes every number in this paper. `binary_unit_burst` stores *point events*: one sample per spike at the unit's peak electrode, with a mean run length of exactly 1.0 samples. `binary_ch_burst` stores the sorter's *waveform extent*: the 4 ms window around each

detection painted onto every electrode in the neighborhood, with a mean run length of 86.8 samples. The loader reads the point-event family. Reading the waveform family instead inflates the voxel rate by roughly $22\times$, and the regression test asserts both run lengths so the two cannot be silently swapped.

Binning, cropping, padding. A stored window is $(120, 220, 12000)$ at one $20\,\text{kHz}$ sample per bin, i.e. $50\,\mu\text{s}$ per bin and $600\,\text{ms}$ per window; there are $2{,}133$ such windows across the 31 assays. Max-pooling by 120 raw bins gives $6\,\text{ms}$ frames, so a frame records *whether* an electrode fired in that interval and not how many times. At the observed firing rate, this temporal pooling discards little information, but it is the reason a voxel is binary and not a count. A window yields 100 pooled frames; a contiguous span of 50 pooled frames (a 6000-sample crop) is drawn at random, then end-cropped to 48 frames, the nearest multiple of the patch's temporal extent of 6. The array footprint is 120×220 and is padded on the width axis to 120×224, a multiple of the patch's spatial extent of 14. The result is a $48 \times 120 \times 224$ binary volume of 1,290,240 voxels.

The two occupancy rates. A whole $600\,\text{ms}$ burst window has occupancy 7.59×10^{-5}. A clip is a $288\,\text{ms}$ span drawn inside such a window, and the draw lands preferentially on the active part, so clip occupancy is about twice that: 1.62×10^{-4}, or 209 spikes per clip. Every number in this paper is computed on clips and quotes the clip rate. Both rates are measured on 48 validation clips at the shipped clip geometry, so they describe the data the model is fitted to rather than a held-out estimate; they are descriptive statistics of the corpus and no result depends on them.

E ARCHITECTURE AND DESIGN CURVES

Level warm-up by loss weight. Levels are activated one at a time by ramping their loss weights. Scaling the level outputs instead would rescale the vectors the EMA accumulates, biasing the very target the codebook is converging to; staged activation by weight gave a large improvement over activating all three simultaneously.

Deduplication criterion. The merge tests whether two entries are the same vector. Usage frequency measures a different quantity, how often an entry is chosen, so a frequency filter deletes rare but distinct motifs and leaves genuine duplicates in place. An earlier version of this step used a frequency filter.

Reconstructions from four assays, with their spatial-map and local-code adherence measured on the assay-balanced evaluation sample, are in Figure S2.

Stage controls, in figures. Against the strongest rung of the null ladder, the motif prior's cross-entropy is lower by 2.04 nats. The two conditioning codes are complementary: the local code reaches R^2 0.9027 on the temporal marginal against the global code's 0.6063, while the global code reaches ΔNLL 0.3036 against the local code's 0.2072. No conditioning predicts the temporal centroid (0.0224–0.1204). A sparse encoder keeps all 32 parent codes available for content against 25 for a dense arm at identical seed and schedule.

Architecture. A $3 \times 3 \times 3$ convolutional stem takes the single input channel to 8 and feeds a patch embedding; the encoder is then a 2-layer transformer of width 64 with 4 heads over the $8 \times 8 \times 16$ token grid, with global attention. The decoder is its mirror with causal-in-time attention, followed by a transposed-convolution patch renderer that maps a token embedding to its 1260 output voxels. The pipeline is therefore a convolution–transformer hybrid. As noted in Section 6, the missing comparison is between the transformer *core* and a convolutional core, not between the entire pipeline and a convolutional pipeline. The code dimension is 64. The residual ladder has 32, 8 and 4 entries at levels 1, 2 and 3, each level a dense tree conditioned on the path above it, so the tree has $32 \times 8 \times 4 = 1024$ leaves. Codebooks are EMA buffers and carry no gradient. Both MaskGIT priors are width 128, 4 layers, 4 heads, dropout 0.1; the activity prior additionally carries a count head over $K_{\max} = 312$ bins and reads the ROI occupancy of a $2 \times 2 \times 4$ partition of the grid into 16 regions of 64 cells, so the shape of the hole enters as context at region granularity. There is no regional count target; the count head predicts one total for the clip.

Deduplication. Merging is by pairwise relative distance at 0.05, giving 961 distinct entries from 1024 leaves. The result is insensitive to moderate changes in the threshold: raising it from 0.02 to 0.2 moves the number of colliding pairs only from 107 to 121, and the median nearest-neighbor relative distance across the alphabet is 0.543. The merged group is a well-separated tail.

patch (T, H, W)	voxels	tokens	blank %	active tok.	spikes/active	capture$_{K=32}$
$(24, 15, 14)$	5040	256	79.6	52.2	4.01	0.0600
$(12, 30, 14)$	5040	256	76.0	61.4	3.41	0.0617
$(6, 30, 28)$	5040	256	73.1	68.9	3.04	0.0699
$(12, 15, 14)$	2520	512	84.5	79.1	2.65	0.0607
$(6, 30, 14)$	2520	512	81.8	92.9	2.25	0.0675
$(24, 15, 7)$	2520	512	86.4	69.4	3.01	0.0559
$(6, 15, 14)$ (shipped)	1260	1024	88.9	113.6	1.84	0.0866
$(3, 15, 14)$	630	2048	92.7	148.9	1.41	0.1385
$(6, 15, 7)$	630	2048	93.4	134.4	1.56	0.1060
$(6, 10, 7)$	420	3072	95.3	143.3	1.46	0.1259
$(3, 15, 7)$	315	4096	95.9	166.8	1.26	0.1809
$(3, 10, 7)$	210	6144	97.2	173.9	1.20	0.2281
$(2, 8, 7)$	112	11520	98.4	189.4	1.11	0.3799

Table S3: The patch-size sweep, at fixed data. *capture*$_{K=32}$ is the fraction of patch variance a 32-entry k-means alphabet recovers, i.e. a model-free proxy for how well a small alphabet can describe patches of that size. It improves monotonically as the patch shrinks, and so does the blank fraction: at $(3, 15, 14)$ the grid is 2048 tokens of which 92.7% are empty. The shipped choice is the largest patch at which capture is still rising steeply.

Patch size. The sweep is why we describe patch size as a design choice and not an optimum. Reconstruction alone would favor the smallest patch. However, smaller patches enlarge the token grid and increase the blank fraction; a prior trained on a grid that is 97% blank collapses toward blank predictions. The selected patch size therefore reflects a tokenizer–prior trade-off whose limitations are described in Section 6.

Ladder depth. Table S4 quantifies the contribution of each level.

decode depth	AP ↑	gain over previous
z_1	0.0772	–
$z_1{+}z_2$	0.1604	+**0.0832** (100% of clips)
$z_1{+}z_2{+}z_3$	0.2806	+**0.1202** (100% of clips)

Table S4: Decoding the *same* codes at cumulative depth. Nothing is retrained between rows, so the only variable is how much of the residual sum reaches the decoder. The third level is not inert: it adds $+0.1202$ AP on 100% of 64 clips, even though a summary-statistic η^2 analysis of the same codes attributes only $+0.019$ to it. The effect has to be evaluated at the decoder output, not through a summary statistic.

Decoder-side conditioning. An earlier tokenizer carried a gated cross-attention branch giving the decoder direct access to the conditioning codes, and it never became active: over 300 epochs, 21 paired conditional and unconditional validation series (AP, tolerant AP, best-F_1 and its threshold, the BCE and VQ losses, and the three reference reconstruction losses) are bit-identical, maximum absolute difference exactly 0.0. Because the reconstruction objective provides no incentive to use this conditioning path, increasing the gate magnitude did not activate it. In the shipped model, conditioning therefore reaches generation through the priors and affects the tokenizer only through output constraints.

The residual after three levels. An adapter predicting a continuous mixture weight over the codebook hull, in place of a discrete third level, logged 24 diagnostic series identically zero for 300 epochs. What remains after the ladder is close to quantization noise and about 4% predictable in any basis we tried, so a fourth continuous stage has nothing to fit.

Sparse encoder. Against a dense arm at identical seed, data and schedule, the sparse encoder keeps all 32 level-1 parent codes on content patches, while the dense arm uses 25 and shares 13 of those 25 with blank patches: more than half its alphabet is doing double duty. Validation AP is 0.0501 against 0.0358. The dense arm ablates the sparse design as a *package* and does not isolate quantizer routing: two auxiliary losses key off the blank mask and go inactive when nothing is blank. It answers "why this encoder design", not "why this quantizer".

F TRAINING PROTOCOL

Stages are trained in sequence and earlier stages are frozen thereafter; no stage is trained jointly with another. The tokenizer runs 300 epochs with the three ladder levels activated in turn by ramping their loss weights, never by scaling the level outputs. A level scale $\neq 1$ rescales the vectors the EMA accumulates and biases the target the codebook converges to. Staged activation measured $3.4\times$ better than activating all three at once.

The local-context mapper trains for 300 epochs against a multinomial likelihood, with its marginal baseline accumulated online as a Dirichlet(1) posterior mean across every batch of every epoch, not from a single collection pass. The prior stages use warmup followed by cosine decay. The selection metric differs across stages by design: the motif prior selects on MRR, the activity prior on validation NLL, and the adaptation stage on validation MRR. The activity prior in particular must *not* select on a rank metric: rank metrics are blind to calibration, and the activity prior's output is a probability that is later thresholded. Neither the activity prior's count teacher nor its early stopping is allowed to act before the count handoff completes, since a teacher-forced model is solving an easier problem and its score must not set the selection bar.

G HYPERPARAMETERS

Table S5 is the full training configuration, one row per stage.

stage	what	lr	wd	schedule	warm	ep.	clip	pat.	selected on
1	Global-code mapper pretrain	1×10^{-3}	1×10^{-4}	none	–	200	–	20	spatial-map loss
2A	Tokenizer (3-level residual VQ-VAE)	1×10^{-3}	1×10^{-4}	cosine to 1×10^{-5}	–	300	–	40	val exact AUPRC
3	Local-code mapper (texton multinomial)	1×10^{-3}	1×10^{-2}	cosine to 1×10^{-5}	–	300	–	40	val multinomial NLL
4A	Motif prior (MaskGIT over V=961)	3×10^{-4}	1×10^{-2}	warmup + cosine	10	600	1.0	150	val MRR
4B	Activity prior (where, plus count head)	2×10^{-4}	1×10^{-2}	warmup + cosine	5	120	1.0	30	val NLL
4C	Adaptation of the motif prior to emitted maps	3×10^{-5}	1×10^{-2}	warmup + cosine	5	120	1.0	40	val MRR

Shared across every stage: batch size 4 with 8 accumulation steps, i.e. an effective batch of 32; CUDA autocast, FP16; dropout 0.1 in both priors; seed 0. Stages run in sequence and every earlier stage is frozen thereafter.

Per-stage settings. **1** λ_{sep} 0.0001; **2A** λ_{ctx} 2.8, $\lambda_{\text{ctx field}}$ 0.05, $\lambda_{\text{code norm}}$ 0.1, pos weight 100 to 1 over 100 epochs; **3** n textons 128, texton basis z1, batches per epoch 120; **4A** $\lambda_{\text{z1 neighbor ce}}$ 0.05, λ_{ctx} 0.25, λ_{adj} 0.25, λ_{spatial} 0.25, full mask prob 0.15; **4B** λ_{bce} 1.0, pos weight 1.0, λ_{count} 0.065, λ_{spatial} 0.02, random mask prob 0.5; **4C** emitted-map ramp 0 to 0.8 over 20 epochs, readout gumbel, soft field on.

Table S5: Every setting the shipped pipeline trains under, one row per stage. The values are read out of the training source automatically and not transcribed, so a renamed argument raises instead of leaving a stale value in the table. Only the shipped stages appear.

Two settings differ from the others by design. The activity prior is the only stage selected on a likelihood: rank metrics are blind to calibration, and its output is a probability that is later thresholded. The adaptation stage runs an order of magnitude below the stage it adapts, since it corrects a mismatch between two trained modules instead of learning the task again.

H LOSS TERMS

Section 4 names each objective by symbol. Table S6 is where each one is defined.

term	coefficient	what it constrains
Stage 1 $\mathcal{L}_{\mathrm{global}}$		
support, token grid	outside 2.0, mass 1.0	cover the recording's active sites, penalize mass outside them, match the total
support, full array	outside 3.0, mass 2.0	the same at electrode resolution; ramped in as the token arm decays
separation	0.0001	repel the predicted maps of recordings whose true maps already differ, at margin 0.25
adjacency	0.5	match the recording's same-site short-gap rates
Stage 2A $\mathcal{L}_{\mathrm{tok}}$		
tolerant spike reconstruction	100 to 1 over 100 epochs	exact BCE at a decaying positive weight, plus max-pooled hit, peak-margin and multi-count terms
vector quantization	1.0	commitment at $\beta = 0.25$ plus a usage-entropy term at 0.001 against a detached codebook; codebooks are EMA at decay 0.95
context	2.8	the nine local moments recomputed from the decoder's own logits must match the clip's true descriptor
context field	0.05	the same nine moments per patch rather than per clip
spatial, token grid	0.001	no mass where the recording's support map forbids it
spatial, full array	0.0001	the same at electrode resolution, ramped in late
short-gap rate	0.01	same-site inter-spike gaps must not exceed or fall short of the recording's measured rates
code norm	0.1	a soft ceiling on encoder output norms, so codes do not drift out of the EMA's reach
blank hinge	0.05	empty patches must not produce spike logits
blank/active separation	0.01	the decoder's blank and content embeddings stay apart
encoder isotropy	1.0	a variance floor and an off-diagonal correlation penalty, against encoder collapse
Stage 3 $\mathcal{L}_{\mathrm{local}}$		
texton multinomial NLL	1.0	the predicted distribution over the 128-texton basis must match the clip's soft texton histogram
flat-code multinomial NLL	1.0	the same against the $V{=}961$ alphabet
regional summary	1.0	squared error on per-code mean time and per-frame active fraction
Stage 4A / 4C $\mathcal{L}_M$		
categorical cross-entropy	1.0	the correct motif at each masked active cell
neighborhood CE	0.05	partial credit over the five codebook-nearest alternatives at temperature 0.25, so a near miss in alphabet geometry is not a total miss
expected code distance	0.05	the whole predicted distribution is pulled toward the target's neighborhood, not just its mode
context	0.25	soft-decode the motif logits through the frozen tokenizer; the result's nine moments must match the clip
context field	0.05	the same moment match evaluated per patch, so a clip cannot be right on average and wrong everywhere
adjacency	0.25	the decoded volume's short-gap rates must match the recording's
spatial	0.25	no decoded mass outside the recording's support map
Stage 4B $\mathcal{L}_A$		
per-cell BCE	1.0	active against blank at each token cell, at positive weight 1.0 – deliberately not the class-balancing ratio, which over-produces
count	0.065	one categorical over the clip's total active count, with an ordinal neighbor term at 0.25 and an expected-distance term at 0.05 so a near-miss count is scored as one
temporal co-activation	0.1	the predicted field's co-activation rate along time must match the batch's
spatial co-activation	0.1	the same co-activation match taken across the array instead of along time
spatial support	0.02	no predicted activity in columns the true map leaves empty

Coefficients are constant unless noted; the ramps and curricula are in Appendix F. Present in the code and inactive in the shipped configuration: Stage 2A, token-profile entropy (the within-token blur experiment; reported as a failure and not shipped); Stage 4A/4C, the second entry of loss_weights (0.1) (the alpha head it weighted was removed; the motif prior reads loss_weights[0] only); Stage 2A, false-positive weighting (tolerant_spike_loss is called with fp_weight 0, so the hallucination penalty is inactive); Stage 2A, latent masking, CFG dropout (both off in the shipped run).

Table S6: Every term of every stage's objective, with its coefficient and what it constrains. Section 4 names these terms; this is where they are defined. Coefficients are read out of the source automatically and not transcribed. Two coefficients require clarification because their effective runtime values differ from the first values visible in the source: Stage 2A's context weight is rebound at run time from 0.1 to 2.8, and Stage 4B's coefficients come from the second of two configuration dictionaries, the first of which the loss never reads.

I METRIC DEFINITIONS AND MULTIPLE COMPARISONS

Hidden fraction per evaluation setting. Free generation hides 1.000 of the token grid, causal 0.587, noncausal 0.396 and spatial 0.339. These are fractions of the *token* grid, not of the frame axis: a causal prefix keeps 25–75% of frames and a noncausal hole covers a contiguous 30% of them, but holes are snapped to the patch lattice before scoring, which rounds a partly covered token into the hole. Because the base rate differs by column, a number is comparable across arms within a column and never across columns; no aggregate over the four settings is reported anywhere in this paper.

Average precision. Every AP value is computed stepwise, never trapezoidally: linear interpolation in PR space is invalid (Davis & Goadrich, 2006) and here inflates the saturating flat tokenizer by $+0.2239$ against our $+0.0015$. *Site-level* AP scores whether the right electrodes are active anywhere in the clip; *voxel-level* AP scores whether the right electrode is active in the right frame. At this sparsity the two answer different questions and are never averaged together. The canvas is padded from 220 to 224 columns so the width divides the patch grid, and those four columns are inside the scored region: they hold no electrode, so they are negatives for every arm alike and neither adds nor removes a positive. The descriptor and spatial-map metrics un-pad first, since the stored per-assay map is unpadded.

Distributional metrics. Relative error terms are $|\hat{s} - s|/s$ on a named statistic computed per clip. The statistics are the spike rate, the k-frame persistence rates, the mean avalanche size, the burst rate and the mean inter-spike interval. KS terms are two-sample Kolmogorov–Smirnov distances on the inter-spike-interval and avalanche-size distributions. Bin edges and feature lists come from one shared definition that every arm imports, so no two arms can be scored on different bins. Two statistics are excluded as degenerate before any test is run: spatial co-activation is degenerate on 99.2% of clips and single-frame persistence on 28.9%. The exclusion is applied identically to every arm.

Against a single generation composite. Selecting the activity prior on a single scalar combining them fails a basic check: placing the *oracle* activity map into that composite *loses* by 0.0127. A selection metric an oracle cannot win is not measuring what it claims to, so the activity prior selects on validation NLL and the four families are reported separately.

Multiple comparisons. All paired tests are Wilcoxon signed-rank, on the same clips in the same order, with pairing asserted, not assumed. Benjamini–Hochberg (Benjamini & Hochberg, 1995) is applied jointly over the 32 task-axis comparisons (two metric families by four settings by every learned arm against ours), and q is reported for that family. Analyses outside it are identified where they appear and report their own statistic: the ladder-depth increment is a paired test within one model and quotes p, and the vocabulary-overlap null is a z score against a label shuffle.

Scored clips. The temporal test split holds 638 burst windows, unevenly distributed over the 31 assays. Evaluation does not walk that split directly. It scores an assay-balanced set of 279 clips, nine per assay, each drawn from that assay's held-out windows with replacement and cropped under a fixed seed, so a long recording does not dominate and a short one is still represented. The draw is deterministic: the same seed yields the same 279 clips in the same order for every arm, which is what makes the comparisons paired. Nine per assay is the test quota; training and validation use thirty and six.

The four settings are sampled per clip, so a setting scores a subset of those clips and not all of them: 279 on free generation, 278 on causal and on noncausal, and 243 on spatial. A clip is dropped from a setting when its hole contains no spike, since average precision is undefined with no positive, and a spatial box at this occupancy encloses no spike more often than a temporal hole does. The task axis additionally averages 8 Monte Carlo samples per clip under seed 20260822; the distributional battery draws one sample per clip per rung and is not seeded, so repeating it redraws those samples.

The clip set is drawn before evaluation and cannot have been chosen after seeing a result: the draw is seeded, and the loader then reads its 279 indices unshuffled, so the order is fixed and the harness seed governs only the model's own randomness, namely Monte Carlo draws, the Gumbel readout and

mask draws. Every arm therefore sees identical clips in an identical order, which is the precondition for the paired tests. The comparison harness verifies it instead of assuming it, by requiring two runs to agree bitwise on a model-free arm computed from the same ground truth before it will pair them.

The cost sets the budget. One clip in the task axis requires an iterative MaskGIT decode per Monte Carlo sample, in four settings, for each of six arms and their nulls; the full protocol reported here is several GPU-hours. An earlier version of this work ran 12 and 8 batches, which, because the prefix is ordered by assay, reached only 6 and 4 assays. Those numbers are not reported here. The distribution terms in particular are finite-sample statistics whose *absolute* value moves with the budget: one composite we measured shifted by $+0.036$ between 4 and 8 validation batches on a fixed checkpoint, which is why no number is carried between protocols.

J Full conditioning ladders

Section 5.4 reports each family at two rungs, the arm's own random-context control and full context. Table S7 gives all five, which is what shows whether an arm improves *monotonically* as context is added or only differs at the endpoints. Among the families where improvement is the right expectation, A, C and D, two arms fail and the endpoint view hides it: the U-Net degrades on D as context accumulates, and the GLM is flat across every rung. Family B is excluded from that reading because it scores obedience to the descriptor supplied, not proximity to the true clip (Table S7). *local only* is a control and not a rung, since it withholds the assay code entirely.

K Pooled marginal statistics

The U-Net's win on marginal realism has a mechanical explanation. A deterministic conditional-mean field, re-thresholded, sits closest to the pooled marginal exactly when it is least committed to a particular clip, which is why its score degrades from its own random-context null to full context while ours improves.

Pooled statistics are relatively insensitive to conditioning: after aggregation over clips, a model can match the correct marginal while assigning it to the wrong clip. We therefore treat these statistics as complementary diagnostics and base the conditioning claims on per-clip paired comparisons.

L Per-setting battery

Figure S1 and Table S9 are the completion accuracy behind Section 5.3, per setting and per arm.

One readout for every arm. Every arm is scored through the same readout code path. An arm that emits a score is thresholded; an arm that emits a probability is sampled. This matters because an early version of the coupled GLM under-produced spikes by $4\times$ purely from a rate clamp in its own sampler, which is a property of the baseline's implementation and not of the method, and could have been mistaken for a property of the baseline. Own-count columns (accuracy given the true spike count) are reported for every task in the battery, because at this sparsity the count and the placement are separable and an arm can win one while losing the other.

Free generation against reconstruction. Task 0 has an ROI of 1.000: nothing is visible. It is therefore free generation, and its output is bitwise identical for every arm regardless of which clip was nominally being completed. It is included as the no-visible-volume reference, and it is not a fourth completion task.

Figure S3 shows one clip completed under all four settings, with the observed remainder marked so that copied spikes are not read as generated ones.

M Qualitative panels

Both figures here are selected examples, so the selection rule is stated and the distribution it selected from is drawn alongside. Panels are maximum projections over a 16-frame window, not single

arm	random	local only	global only	global partial local	global full local
A. Descriptor error (z) ↓					
Ours (4C+soft)	1.0623	0.8473	0.6738	0.6172	0.5350
MaskGIT-flat	1.1892	0.9595	0.8347	0.7528	0.7362
3D U-Net (det.)†	1.0741	0.8415	0.7711	0.6740	0.6803
3D CVAE	1.1388	0.9099	0.8373	0.7581	0.7441
Dich. Gaussian	0.7243	0.5183	0.6009	0.5298	0.5137
Coupled GLM	0.5447	0.5584	0.5560	0.5648	0.5610
B. Adherence ↑					
Ours (4C+soft)	0.7183	0.3081	0.6945	0.7212	0.6992
MaskGIT-flat	0.5608	0.2535	0.5457	0.4971	0.5429
3D U-Net (det.)†	0.5047	0.3280	0.5042	0.5120	0.5165
3D CVAE	0.4872	0.3364	0.4885	0.4751	0.4925
Dich. Gaussian	0.1779	0.5150	0.5210	0.5416	0.5472
Coupled GLM	0.0057	0.4973	0.5094	0.5039	0.5012
C. Spatial placement, lookup-proof ↑					
Ours (4C+soft)	0.0018	-0.0012	-0.0058	0.0104	0.0101
MaskGIT-flat	-0.0008	0.0003	-0.0040	0.0039	0.0027
3D U-Net (det.)†	-0.0014	-0.0004	0.0010	0.0047	0.0035
3D CVAE	-0.0009	0.0010	-0.0014	-0.0003	0.0052
Dich. Gaussian	-0.0009	0.0049	0.0018	0.0014	0.0058
Coupled GLM	0.0048	-0.0052	0.0066	-0.0028	-0.0049
D. Marginal realism ↓					
Ours (4C+soft)	0.2467	0.2285	0.2523	0.2145	0.2242
MaskGIT-flat	0.3684	0.3808	0.3704	0.3842	0.3661
3D U-Net (det.)†	0.0762	0.1088	0.0981	0.0881	0.1025
3D CVAE	0.3450	0.4169	0.3454	0.3311	0.3484
Dich. Gaussian	0.1455	0.0661	0.0681	0.0619	0.0640
Coupled GLM	0.0311	0.0321	0.0382	0.0334	0.0268

Table S7: All five conditioning rungs, all four families, all six arms. *random* supplies a mismatched context and is the control, not a rung. For families A, C and D the target is the true clip, so monotone improvement from *random* to *global full local* is the evidence that an arm reads its conditioning; the 3D U-Net moves the wrong way on family D, getting worse as context is added. Family B is scored differently and is exempt: it measures agreement with the descriptor the model was *handed*, so an obedient arm matches whichever descriptor it receives and is flat across the ladder, and a high score under *random* is the expected behavior rather than a failure to condition. The informative movement on B is the collapse at *local only*, where our arm falls to 0.3081: without the assay code it cannot act on the descriptor at all.

statistic	4C+soft (ship)	MaskGIT-flat		DG (Macke'09)		GLM (Pillow'08)	
	median	median	q	median	q	median	q
firing rate	0.2619	0.0523	1.3×10^{-143}	0.0667	6.2×10^{-121}	0.2365	7.0×10^{-3}
4-frame persistence	0.7147	0.5278	9.1×10^{-20}	0.4021	5.8×10^{-53}	0.4512	7.2×10^{-49}
mean avalanche size	0.4948	0.4602	2.8×10^{-1}	0.3485	4.5×10^{-1}	0.6832	5.0×10^{-34}
ISI distribution (KS)	0.3465	0.5000	1.2×10^{-66}	0.2055	2.0×10^{-91}	0.2083	9.8×10^{-102}
avalanche size (KS)	0.5000	0.5000	4.0×10^{-1}	0.5000	7.0×10^{-6}	0.6667	8.8×10^{-34}
aggregate	0.4769	0.4404	3.7×10^{-11}	0.3733	7.6×10^{-30}	0.5138	1.5×10^{-1}

Table S8: Per-clip relative error on the pooled summaries, median over clips, with the paired difference against ours corrected by BH within this family. Lower is better throughout. The matched peer wins the rate and persistence terms and we win the inter-spike-interval distribution; the two lookup arms win the terms a per-assay site map determines and lose the shape terms. These values are measured using the generation-regime sample sets, which contain four samples per clip. This protocol differs from that used in Table 2, so the results are not interchangeable.

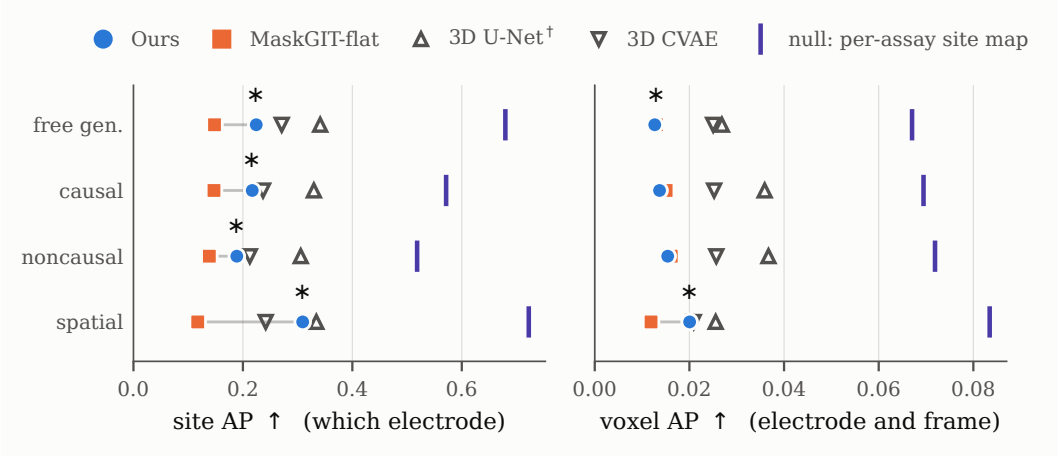

Figure S1: **Completion accuracy, paired per clip.** The left panels show site-level accuracy; the right panels additionally require the correct frame. $*$: the paired test favors us over the matched peer at $q < 0.05$; the gray connector joins that pair. †: no reusable representation (Appendix P). The violet rule is the static per-assay site map, a lookup that is constant within an assay and carries no clip-specific information. It still outranks every learned arm here. That is a property of average precision, which ranks a probability map and is therefore maximized by the conditional mean, not by a sample; generation is scored in Figure 3. Panels have independent x axes.

arm	free gen. (0)	causal (1)	noncausal (2)	spatial (3)
Site AP ↑: which electrode				
Ours	0.2244	0.2170	0.1889	0.3092
MaskGIT-flat	0.1481*	0.1472*	0.1387*	0.1174*
3D U-Net†	0.3412	0.3297	0.3057	0.3342
3D U-Net, seed 2†	0.4045	0.3414	0.3079	0.3475
3D CVAE	0.2709	0.2368	0.2127	0.2417*
null assay site map, seen	0.6796	0.5714	0.5184	0.7224
null assay site map, unseen	0.0284	0.0230	0.0215	0.0350
Voxel AP ↑: which electrode and *frame*				
Ours	0.0127	0.0137	0.0154	0.0200
MaskGIT-flat	0.0130*	0.0151°	0.0162°	0.0119*
3D U-Net†	0.0269	0.0359	0.0367	0.0255
3D U-Net, seed 2†	0.0322	0.0339	0.0355	0.0253
3D CVAE	0.0250	0.0252	0.0257	0.0209
null assay site map, seen	0.0670	0.0695	0.0719	0.0835
null assay site map, unseen	0.0022	0.0021	0.0030	0.0034

Table S9: Average precision inside the hole, both levels, all four tasks. $*$: the paired test favors ours; $\circ$: not significant; unmarked: the other arm wins. Read within a column only: ROI fractions differ ($1.00/0.59/0.40/0.34$) and so do the clip counts ($279/278/278/243$), so the columns have different base rates. Nothing is bolded because the arms serve different roles and are not a single ranked benchmark.

frames: at this occupancy a frame carries a handful of spikes over $26{,}880$ sites, so an instant is visually empty and shows nothing about motif structure. Rows are chosen by the fraction of *achievable* spatial-map correlation the model attains, $r_{\text{model}}/r_{\text{truth}}$, where the denominator is the clip's own ground truth against its assay's training site map. Selection, the ranking it came from and every number quoted below are written to a committed artifact by the panel generator.

Spatial-map correlation carries the tolerance the rest of the pipeline already uses. Placement is learned through the global code and a spatial-violation term, never supervised per electrode, so both the training loss and the generation metrics accept a prediction within one site of an allowed one, a dilation of radius 1; the same radius is applied here, to the model map and to the reference alike so that the ceiling moves with it. The reference is the clip's own ground truth against the same training site map, and it is far below 1 for a reason that bounds any single-clip generator: the map is an average over many clips, while one clip is a sparse draw from it, a few hundred spikes over $26{,}880$ sites. Two real clips of the same assay agree only to a median r of 0.7282 across the 31 assays. Panels therefore report the attained fraction, model over reference, not a raw correlation.

N Count discrimination and its arithmetic control

How many spikes a clip contains and where it puts them are separable at this sparsity, so the battery reports both. Within an assay, our correlation with the true ROI spike count is 0.9138–0.9638 across the three completion settings, against 0.0716–0.4892 for MaskGIT-flat.

That result must be compared with an arithmetic control and not with zero. Every model is given the clip's local code. It is computed on the whole clip, including the hidden region, and that is deliberate: the completion tasks here supply a description of the clip and ask the model to fill the gaps consistently with it, so the descriptor is a prompt and not something the model is expected to infer. A consequence is that its first feature is log mean firing density, so $\exp(\mathrm{lct}_0) \times |\mathrm{ROI}|$ predicts the count with no model at all and scores 0.4550–0.8328. It provides a model-free baseline that a useful count head should exceed, is computed alongside the other model-free references, and is exceeded by our model in all three completion settings.

Discrimination and calibration are separate capabilities and no arm here has both, and the two count quantities this battery reports must be kept apart. The own-count column sums the decoded voxel probabilities inside the hole and takes that sum as the arm's expected spike count; ours runs about $6\times$ the true count, and the only arm whose bias stays within $\pm 25\%$ everywhere is a per-assay lookup (Elkan, 2001). That figure belongs to the decoded probability field, not to the activity prior's categorical count head, which is a separate output. Measured on its own, the head is unbiased under the random masks it trains on and over-predicts by 13.2% on train clips and 15.2% on validation clips under the fixed protocol that validation and generation use, whose distribution of hole sizes it never saw. One fitted multiplicative scale on held-out train clips takes that bias to 2.4% and count MAE from 10.500 to 8.913; an affine fit and a hole-size covariate add nothing over it. The shipped default leaves the scale at 1, so no generation number in this paper depends on the correction. Ranking the right sites and emitting the right number of them are different problems. The count head sets the second at token level: its logits give the K active cells the Gumbel top-K readout draws. How many voxels that becomes is then set by the decoder and the voxel readout, which is why the token-level bias above and the decoded over-count are different numbers.

O Null construction

A margin over uniform says very little at $V = 961$. Table S10 therefore scores the motif prior against a ladder of four progressively harder model-free lookups.

The four rungs of Table S10 are cumulative lookups fitted on the training split and applied unchanged at test time. *Uniform* is $1/V$. *Global marginal* is the empirical code distribution pooled over all assays. *Per-assay marginal* conditions that distribution on the assay identity. *Per-assay $\times$ position* conditions it on the assay and the grid cell jointly, and is the strongest: it is the best a pure lookup table can do without seeing the clip.

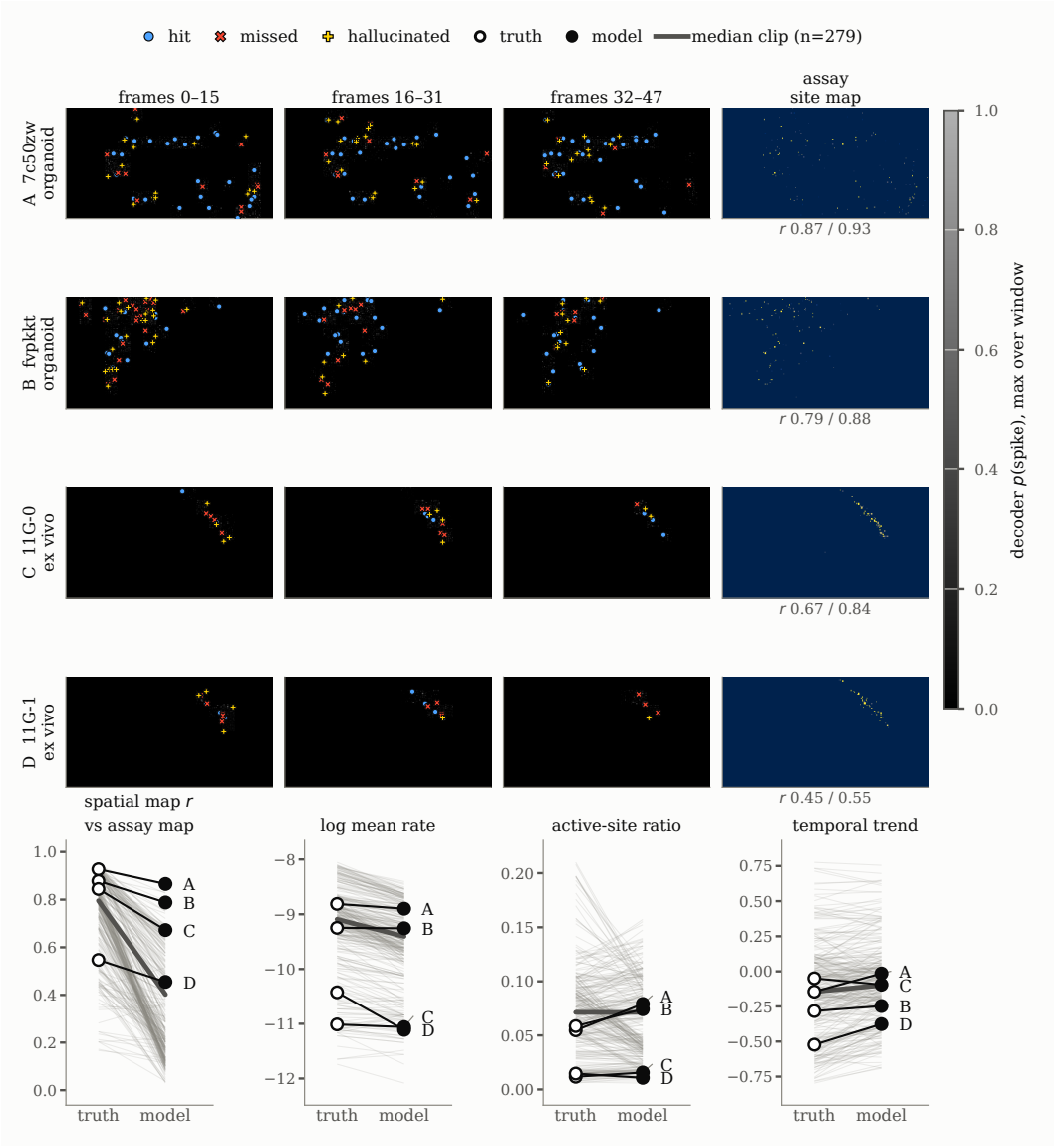

Figure S2: Tokenizer reconstruction, four assays, two per corpus. Marks are the thresholded reconstruction against the truth; the grayscale ground is the decoder's probability. The right column is the assay's training site map, the reference the correlation beneath each row is measured against. These rows attain 0.9335, 0.8973, 0.7962 and 0.8315 of their achievable map correlation and are therefore far from typical: over all 279 test clips the median falls from 0.7955 to 0.2335, and only a fraction 0.0179 of clips place better than their own truth does. The strip is paired (one segment per clip, heavy line the median paired change) because the pooled version disagrees: active-site ratio has marginal medians 0.0713 and 0.0479, a third apart, while the median *per-clip* change is 0.0 with a fraction 0.491 of clips moving up. The rate deficit is the real weakness on display, a median −0.3129 in log mean rate with only a fraction 0.0538 of clips over-producing. The tokenizer is given the true local code, so preserving it here is close to circular; Figure S3 is where conditioning is actually tested.

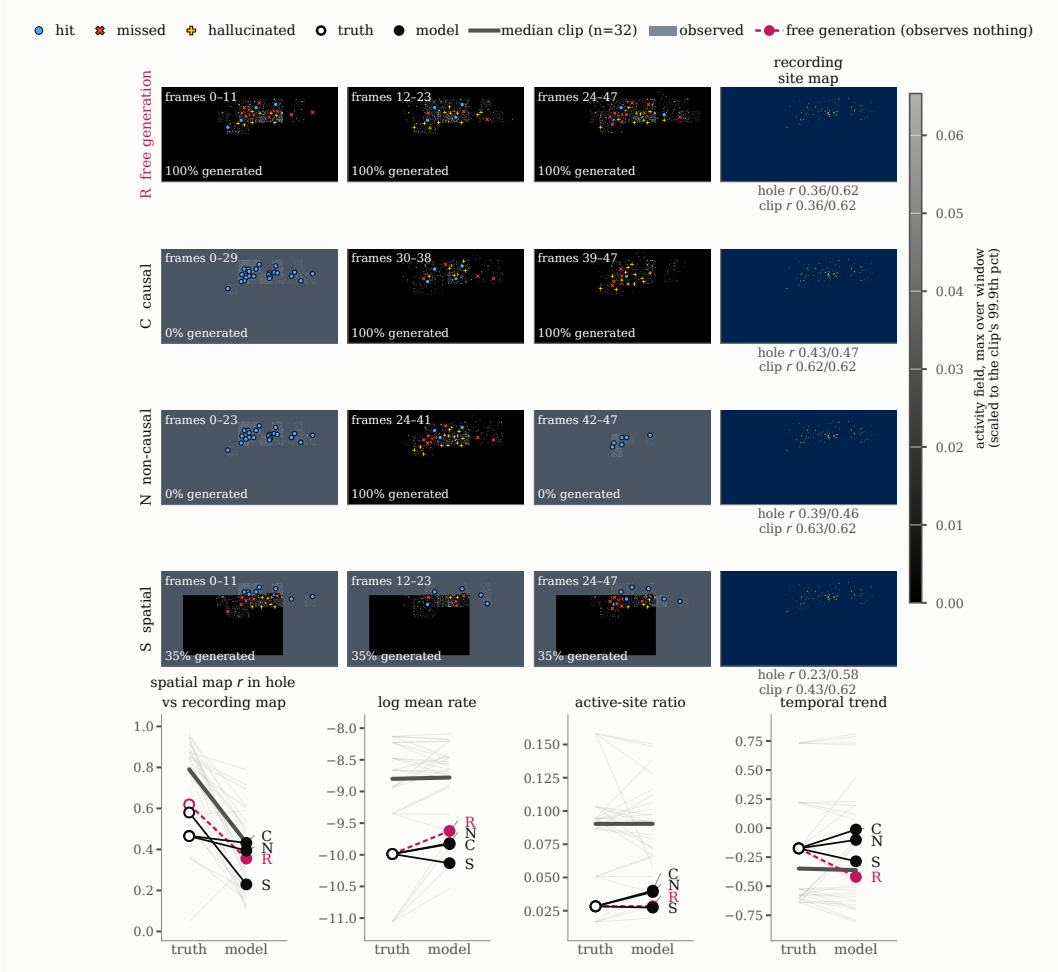

Figure S3: One ground-truth clip under all four settings. The dump indexes the same clip under every task, so these rows are one clip seen through the whole conditioning ladder and not four unrelated clips. The strip therefore shows a single hollow truth point connected to four corresponding model points. Washed regions are observed, not generated, and each panel states the fraction it produced; without that mark, spikes copied from the visible remainder could be mistaken for generated hits. The map column is measured inside the hole only. A whole-clip correlation would be symmetric (the copied remainder is identical in the model's clip and in the truth) but not comparable across these rows: for a temporal hole every site is observed in some frame, so projecting over time fills the canvas from copied frames, whereas under the spatial hole a fraction 0.352 of sites is never observed at all. Inside the hole the attained fraction is 0.9258 from a prefix and 0.8486 with two-sided context, against 0.5741 with nothing observed and 0.3954 when the hole is a region of the array. Both inpainting settings clear the zero-context reference, so the visible frames are being used to place activity and not merely to fill the rest of the volume. Free generation is drawn dashed in crimson because it observes nothing and is therefore not a completion setting at all (Appendix L); it is the zero-context reference with which the other three settings are compared, and the spatial setting is the only one that falls below it. The whole-clip value is printed under each site map for reference and runs 0.5741–1.0181; the gap between the two columns is the copied share, not a result. The spatial setting is the weak one, and it is where the static site map of Appendix L beats us.

predictor	top-1 ↑	top-5 ↑	median rank ↓	MRR ↑	CE (nats) ↓
uniform	0.0024	0.0044	481	0.0021	6.8680
global marginal	0.0078	0.0321	156	0.0305	6.2895
per-assay marginal	0.0149	0.0599	102	0.0497	6.3026
per-assay × position	0.0675	0.2512	53	0.1529	6.1713
motif prior (ours)	**0.0860**	**0.3668**	**9**	**0.2168**	**4.1296**

Table S10: The motif prior against its four-rung null ladder over $V = 961$. Median rank is quoted because the distribution has a long tail and the mean (44.65) misrepresents it. The ladder is shown in full so the margin is compared with its hardest rung, the per-assay code distribution at that grid position, and not with uniform. The MRR here is teacher-forced: the prior is scored against the true activity field. Table S14 reports the deployment figure instead, under the activity field the prior itself emits, which is lower and is the number the shipped model is selected on.

The random rung as a control. The conditioning ladder in Appendix J has a *random* rung in which the model is given another clip's context. It is tempting to interpret the correlation across all five rungs as a measure of conditioning strength, but the random rung supplies an assay index that bypasses the ladder's ordering entirely, so a pooled correlation over the five rungs measures the wrong thing. The quantity we report is the within-assay gap between the full-context and random-context rungs, which is paired and does not mix the two effects.

P BASELINE IMPLEMENTATIONS

Stationary fits for both statistical baselines. A full per-electrode covariance over the $26{,}880$ padded canvas locations is not estimable from the available windows at this firing rate, so neither a per-channel covariance nor per-electrode coupling can be fitted. Both models are therefore used at the stationary parameterization their own literature prescribes; this is a property of the data, not a weakening of the baselines.

Free generation: mean against median. We win free generation on the paired test while posting a marginally lower *mean* voxel AP: we are ahead on 62% of clips, and a few clips on which the flat arm does much better move the mean without moving the median. The paper therefore quotes the paired test.

Win-loss count across the paired comparisons. Across the two metric families and four settings, other methods outperform ours in 23 of the 32 paired comparisons; 7 go to us and 2 are ties. All 23 wins are achieved by the directly supervised convolutional inpainters (the U-Net, its seed replicate and the CVAE), and *none* by the matched generative peer, against which we take 6 of 8 with the other two tied.

The U-Net as an upper reference. It beats us on voxel AP in all four completion settings ($q \leq 3.53 \times 10^{-5}$), and it is deterministic: one context yields one field, so every "sample" is that field re-thresholded. Its marginal-realism score worsens as context is added (Section 5.4), and it has no entry in the alphabet comparison because it has no discrete codebook. The dichotomized Gaussian likewise emits a static per-site field that cannot read the visible remainder.

Cost of withholding the site map. Replacing the per-assay map with a single global one takes GLM adherence from 0.5012 to 0.0188 and DG's from 0.5472 to 0.2076, and DG's within-assay placement gap falls fivefold, from 0.0058 to 0.0011. Adherence at 0.0188 means the GLM does not degrade so much as stop tracking the requested context altogether; its own placement gap is negative both with the map and without it, so that capability was never there to lose. This is the measurement behind Section 5.3: what these arms score is the map.

Table S11 is the parameter count behind that claim.

MaskGIT-flat. MaskGIT-flat uses a single codebook of 1024 entries on the identical $8 \times 8 \times 16$ grid and $(6, 15, 14)$ patch, together with the same prior family and conditioning inputs. Token bud-

arm	fitted (shared)	shared maps	per assay	at 1000 assays
Ours	8,620,946	0	**0**	0
MaskGIT-flat	13,232,961	0	**0**	0
3D U-Net†	8,372,673	860,160	**0**	0
3D CVAE	9,360,601	860,160	**0**	0
ref Dich. Gaussian	26	26,880	**26,881**	26,881,000
ref Coupled GLM	5,194	53,792	**26,880**	26,880,000

Table S11: Parameters shared across assays versus tabulated per assay. The per-assay column counts one value per padded canvas location, not per physical electrode and not per routed channel. The two lookup arms grow linearly in the number of assays; every learned arm, including the peer, is flat. At 31 assays that is 833k tabulated values, and at a thousand it would be 26.9M.

get, alphabet size and training budget are matched. Its tokenizer is the ordinary one this literature describes: strided-convolution patch embedding, four residual convolutional blocks, one codebook, mirrored decoder, with no attention anywhere in it. It is built that way deliberately, since comparing against our own tokenizer with the ladder switched off would not provide an independent architectural baseline. The consequence is that the arm does *not* isolate the residual ladder: it varies the tokenizer core, the codebook structure, the blank route and the auxiliary losses together. What it bounds is the pair, and the oracle-code comparison in Section 5.1 removes only the prior from that pair, not the encoder architecture.

3D U-Net and 3D CVAE. The U-Net is an inpainter: it takes the masked volume and the mask, and is trained with the same loss on the same hole distribution. It has no autoencoding path, so it cannot reconstruct, and its free-generation output is the all-masked corner of the same task, driven by FiLM on the conditioning codes and thresholded at a fitted log-linear rate. The CVAE is the same backbone with a latent variable added and nothing else changed, a conditional VAE in the standard form (Sohn et al., 2015), with a recognition network $q(z \mid x, c)$ used only in training, a *conditional* prior $p(z \mid c)$ and a generator $p(x \mid z, c)$, where c is the visible remainder, the hole mask and the two conditioning codes. It therefore samples rather than regresses, and its free-generation output is a genuine draw from that conditional prior, not a posterior mean.

Both required an explicit positional embedding: FiLM conditioning is spatially uniform, so without one a convolutional stack cannot represent a per-assay site map at all, and the U-Net's spatial-map correlation falls from 0.2757 to 0.0552.

The CVAE's latent nevertheless collapses, and Table S12 records the measures used to prevent posterior collapse and the resulting KL. These measures include a conditional prior, a down-weighted KL, KL warm-up, per-channel free bits (Kingma et al., 2016), and charging KL only where the hole is; two independent fits collapse despite these measures.

We interpret this collapse as consistent with the conditional structure of the task rather than evidence, by itself, of an optimization failure. The conditional prior cannot see inside the hole: at generation time nothing about the held-out region is available to it, so the skip connections already supply much of the available conditioning; the ELBO therefore has little incentive to retain additional latent information. Forcing usage with a larger β floor or higher free bits would increase latent activity without establishing that it carries useful held-out information. The consequence for the tables is that this arm behaves like the deterministic one, which is why the two convolutional arms move together throughout.

Dichotomized Gaussian and coupled GLM. Both are fitted per assay at the stationary parameterization their own literature prescribes. The data force this choice: at a rate of 1.62×10^{-4} and roughly 200 spikes per clip, a full per-electrode covariance or a per-electrode GLM over the $26{,}880$ padded canvas locations is not estimable from the available windows. Table S13 is the test that establishes what these two arms are actually using.

fit	epochs	peak val KL/dim	final val KL/dim	collapsed
shipped	31	0.1300	0.0100	yes
no positional embedding	40	0.0270	0.0089	yes

Both fits carry the standard anti-collapse measures: a *conditional* prior $p(z \mid c)$ in place of $\mathcal{N}(0, I)$; the KL down-weighted to $\beta = 0.05$; KL warm-up over the first 30% of training; free bits at 0.05 applied *per channel*, which is where collapse happens; the KL charged only on the latent cells the hole touches, so the latent is not taxed for what the skip connections already carry; and the posterior log-variance initialized at -4.0. The latent is a 4-channel grid and not a global vector, so it can say *where* the extra spikes go.

Table S12: The CVAE's latent, and the measures taken against collapse. Two independent fits, differing only in whether the backbone carries a positional embedding, both end below the free-bits floor. Settings and outcomes are read from the fit reports.

ref arm	within-assay gap		adherence		map r	
	with	without	with	without	with	without
Dich. Gaussian	+0.0058	+0.0011	+0.5472	+0.2076	0.5807	0.0694
Coupled GLM	-0.0049	-0.0004	+0.5012	+0.0188	0.5883	0.0624

Table S13: The two lookup arms with their per-assay site map replaced by the global one, same models and same clips. The GLM's adherence does not merely degrade, it goes negative: without the map it stops tracking the requested context at all. In this evaluation the two arms therefore function primarily as assay-specific memorization references.

Q SEED VARIANCE

The adaptation stage produces a modest improvement, so we test whether it exceeds variation across seeds. Table S14 reports four runs.

run	best epoch	epochs	val MRR	deployment MRR
shipped	24	64	0.21359	0.20351
seed 101	56	96	0.21254	0.20581
seed 202	35	75	0.21225	0.20447
seed 303	80	120	0.21377	0.20649
mean $\pm$ sd			0.21304 ± 0.00076	0.20507 ± 0.00134
ref no adaptation, soft field				0.18981
ref no adaptation, hard map				0.14266
ref oracle activity map				0.21488

Table S14: The adaptation stage across four independent seeds, each re-initialized from the unadapted control and run through the real dispatch path. The effect over the soft-field reference is $+0.01526$ against an across-seed sd of 0.00134, i.e. $11.4\times$, and the *worst* seed still beats that reference by $+0.01370$. The selected epoch varies from 24 to 80 while the value moves 0.0015, so early stopping is picking a plateau and not a lucky checkpoint.

Table S14 also shows two important features of checkpoint variability. The shipped checkpoint is the *worst* of the four on deployment MRR. It was selected on validation MRR, and validation turns out to be a poor predictor of deployment across seeds: seed 101 has the worst validation score and the second-best deployment score. Swapping in seed 303 after seeing the test numbers would be selecting on test, so the val-selected checkpoint ships and the penalty is reported. And the same seed test settles the direction of adaptation.

Direction of the adaptation stage. The reverse direction, which freezes the motif prior and tunes the activity prior on the downstream objective, faces three limitations. The sampled map is discrete, so the gradient is a straight-through surrogate through a top-K whose true Jacobian is zero almost everywhere; the trainable surface cannot change the *ranking* that decides which cells light; and

"make the map closer to truth" is already the activity prior's own training objective, pursued with full gradient access. Measured over 56 epochs, every candidate sat inside one seed standard deviation, with motif MRR declining at $t = -10.45$. That entire range inside 0.92 seed sd distinguishes it from the shipped direction, whose effect is 11.4 times the same quantity.

R LIMITATIONS: SUPPORTING DETAIL

The nature of the input. The model consumes a binary volume over the array footprint, but that volume is built from sorted units: each curated unit is written to its peak electrode as a point event (Appendix D). So the input is a spatial raster of sorted unit locations, not raw threshold crossings and not unsorted multi-unit activity. What the phrase *array-wide binary volume* claims is about the model and not the recording: there is no per-unit parameter, no unit correspondence across assays, and no unit-indexed output head. Sorting acts as an upstream filter that decides which events enter the canvas.

Scope of the forward model. Every clip here is spontaneous activity. Nothing in the corpus carries electrical stimulation, and the conditioning interface has no input for one, so the model gives the distribution of activity a preparation produces when left alone. That supports simulation and supplies the unperturbed baseline against which a stimulus-evoked change would be measured. It does not by itself support offline controller design, which needs transition dynamics conditioned on applied stimulation; reaching that would mean extending the conditioning interface to carry stimulus events and training on recordings that contain them.

Resolution and its uses. The prior recovers 80–98% of its alphabet's site-level oracle-code reference and 4–6% of the voxel-level one, and the within-token temporal profile is close to uniform (Appendix S). The model should therefore be read as generating which electrodes are active and the envelope of a burst, not the millisecond placement of spikes within an active patch. Uses that turn on exact spike timing are outside what the present resolution supports.

Recordings, preparations and memorization. The organoid assays come from a cohort of at most six organoids and the archive does not say which assay came from which; the slice assays come from two slice preparations, one per neurosurgical patient. Generalization to an unseen preparation is not only untested but impossible under the current conditioning interface: the per-assay code is a seeded random vector, so a new preparation supplies the model with no information (Section 5.3). Assay-specific information certainly does live in the shared weights. The claim is not that the model has memorized nothing, only that what it memorizes does not grow a table as assays are added.

Smaller patches. They reconstruct better but leave 97% of tokens blank, and a prior trained on a grid that empty collapses onto predicting blank. The patch size is therefore a tokenizer–prior trade-off resolved in favor of the prior, not an optimum for either alone.

Exclusion of the blank token from the reuse statistics. It is 91.7% of all tokens and every assay emits it, so including it would push every pairwise overlap toward 1.0 while measuring nothing. Excluding it therefore makes the overlaps reported in Section 5.2 conservative and not artificially favorable.

Token granularity. Our alphabet's site-level oracle-code reference on free generation is 0.2296 and the prior reaches 0.2244, 98% of it. The voxel-level reference is 0.2535 and the prior reaches 0.0127, 5% of it. The gap between those two fractions quantifies the within-token timing problem: the model places activity on very nearly the right electrodes and then distributes it wrongly in time inside the 36 ms token.

Evidence on the tokenizer core. A convolutional VQ tokenizer is measured here: MaskGIT-flat's is fully convolutional, with no attention anywhere in it, on the same patch and grid at a matched token budget, and given the true codes it represents held-out activity a factor of 5.2 worse than ours (Section 5.1). That is the closest evidence available, and it is not a controlled comparison: the same

arm also replaces the three-level ladder with one flat codebook and drops the blank route, so core and alphabet vary together.

A second bound comes from the U-Net, which reaches spatial-map correlation 0.2757 with a positional embedding and 0.0552 without, against our 0.3050. The positional-embedding ablation is consistent with the expected mechanism: convolution is translation-equivariant whereas the electrode array has absolute site identity, so an explicit spatial basis is needed to represent a fixed site map. It bounds how much of our spatial result is attributable to the transformer and how much to the patch grid and the positional information, but the U-Net differs from our tokenizer in objective and supervision as well as in structure.

S WITHIN-TOKEN BLUR

A $(6, 15, 14)$ patch is 1260 voxels and the decoder places probability across it too evenly. Over content tokens, the model's within-token temporal profile entropy is 1.7226 nats against a real-data value of 0.3282 and a completely flat ceiling of 1.7918; spatially it is 3.8674 against 0.2478 with a ceiling of 5.3471. Two fixes were tried and both are reported as failures.

A *peak term* rewarding concentration within the token converges only as a late fine-tune; from scratch it prevents the tokenizer converging at all. As a fine-tune it sharpens all four profile axes, but at a substantial cost in AP. A *token-entropy constraint* on the within-token profile is dose-ordered and also too expensive: at the higher of two doses it moves the temporal profile entropy from 1.7801 in the matched unconstrained arm to 1.7622 and the spatial from 4.4521 to 4.2710 (1.2% of the temporal gap to real data and 4.3% of the spatial) for a 28% fall in validation AP as the ramp completes.

All three numbers are read from the *final* checkpoint and not the validation-selected one. The term is ramped in late, so validation selected a checkpoint three epochs after activation, at which point every diagnostic still showed the untreated model; a ramped-in term measured anywhere but the last epoch will appear to do nothing.

Neither fix is in the shipped model; the tolerant spike loss remains the only blur control. The blur is why Section 6 names within-token timing, not spatial placement, as the dominant residual error.

T STAGE 3: THE LOCAL-CONTEXT MAPPER

The nine local-context scalars, and the level-1 basis. They are log mean firing density; the second moments of the spike mass in x, y and t together with their three cross-terms; the active-site ratio; and a temporal trend score. All are computed directly from the volume, with nothing learned. The clustering target is a 128-texton basis built on the first ladder level rather than on the flat alphabet because level 1 carries motif identity, while levels 2 and 3 carry fine-grained details that the nine clip-level statistics cannot predict.

conditioning	ΔNLL flat $\uparrow$	ΔNLL texton $\uparrow$	R^2 t-centroid $\uparrow$	R^2 t-marginal $\uparrow$
`lct` only	0.2072	0.0943	0.1156	0.9027
`gct` only	0.3036	0.0861	0.0224	0.6063
both	0.2999	0.1017	0.1204	0.8885

Table S15: The three conditioning arms at epoch 300, sharing one frozen encoder pass so the unique contribution of each is measured exactly. The local code carries temporal structure the global one cannot; the global code carries code identity the local one cannot; neither alone suffices. The t-centroid column is a negative result and is reported as one.

Table S15 separates what each code contributes. Convergence is not assumed: the mean of the last 20 epochs matches the final epoch to four decimal places on every reported quantity. The base NLL against which Δ is measured is 6.2482 nats on the flat alphabet and 4.8164 on the texton basis.

Multinomial likelihood against histogram R^2. Scoring a per-clip code histogram with R^2 has a hard ceiling that has nothing to do with the model. Writing the target as a true rate plus multinomial

noise from drawing $n \approx 86$ active tokens gives $R^2_{\max} \approx n\,\mathrm{CV}^2/(V + n\,\mathrm{CV}^2)$, which is 0.082 at $V = 961$. More training approaches that ceiling and cannot move it. A multinomial likelihood has no such denominator, because each clip contributes n real draws, so a 961-way alphabet is exactly as estimable as a 32-way one, and the score is also the objective the prior itself optimizes.

basis, #textons	base NLL	Δlct	Δgct	Δboth
z_1, 32	3.4221	0.0736	0.0716	0.0826
z_1, 128	4.7935	0.0836	0.0808	0.0941
$z_1{+}z_2$, 32	3.4162	0.0666	0.0628	0.0737
$z_1{+}z_2$, 128	4.6878	0.0712	0.0683	0.0800
flat, 32	3.4108	0.0607	0.0580	0.0672
flat, 128	4.6948	0.0675	0.0643	0.0756

Table S16: Which basis the textons are clustered on, and how many. The ordering $z_1 > z_1{+}z_2 > \mathrm{flat}$ is monotone for all three arms, and 128 textons beat 32 throughout. Descriptor space and decoder space serve different roles: level 1 carries reusable motif identity, while levels 2 and 3 carry fine-grained details that the nine clip-level statistics cannot predict.

Table S16 reports the sweep over basis and texton count.

U Stage 1: the global-context pretrain

The per-assay code is pretrained against that assay's spatial support map (which electrodes are ever active) so that the code acquires the association before any prior consumes it. Per-assay support maps and adjacency diagnostics for all 31 assays are in the code release.

The conditioning interface. All assay-level conditioning enters as one vector per assay. The motif prior reads it through the frozen Stage-1 mapper; the activity prior projects it directly with a linear layer. Substituting a measured descriptor of a preparation (unit map, ISI distribution, days *in vitro*, cell type, stimulation protocol) for the seeded random code is therefore a change to those two input projections and nothing downstream. We state this as a design property, not as a result: no experiment in this paper substitutes anything, and the random code licenses no claim about a preparation the model has not trained on.

V Reproducibility

The anonymized release contains the model, all training stages, all six evaluation arms including ours, the evaluation harness, and the generators that produce every table and figure in this paper. Its checkpoint manifest lists every checkpoint, marks which is shipped for each stage, and records how it was selected. The README gives the ordered sequence of commands the generators are driven by: the four extraction steps that write the provenance, preprocessing, hyperparameter and evaluation-budget artifacts, then the table and figure generators, then the prose check. The anonymized supplement is that release with absolute paths replaced by a placeholder. It carries every checkpoint an evaluation or generation run loads: the ten the manifest marks as shipped, plus the Stage-1 pretrain holding the global-context mapper and its spatial memory banks, the context bank the conditioning ladder reads, and the model-free reference banks. It also carries the report artifacts the table and figure generators read, and both generators were run from a fresh unpack of the supplement to confirm it. What is omitted is the exploratory and superseded checkpoints and the per-clip sample dumps, which the manifest lists.

Figures are byte-stable across runs. Every table is a generated file. Editing one by hand would be overwritten and, worse, would silently disagree with the run that produced it, so the generator or source artifact should be modified instead. Numeric values in prose can become stale without being detected, and the last step addresses this risk: it extracts every numeric literal from the manuscript and fails if one no longer appears in any committed artifact. The evaluation protocol is pinned as described in Appendix I.