MATCHING ONLY READOUTS TELLS INCOMPLETE STORIES: TEMPORAL CONTEXT SHAPES DECODING GAINS IN AN INVASIVE BCI FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen neural foundation models are often evaluated with readouts, which are small networks trained to predict behavior from a temporal window of neural signals or frozen features. The backbone can integrate activity outside the raw-signal readout's window, so matching readout architectures does not ensure the same temporal context. We study how much of the frozen-feature decoding gain depends on the readout's temporal context. We use a model pretrained by masked reconstruction on intracortical spikes and surface electrocorticography (ECoG) without behavioral supervision. On MC-RTT and MC-Maze, two datasets from macaque reaching tasks, we test whether additional temporal context reduces the decoding gain from frozen features. Each readout's input window is widened at fixed parameter count, while both receive inputs matched in time and the same explicit information about each prediction's position within the trial. We also examine whether Gaussian smoothing, which averages neighboring spike counts before decoding, reduces the gap without pretraining, selecting the kernel width on validation data. Widening reduces the R^2 gap between frozen-feature and raw-spike readouts by 89% on MC-RTT and 68% on MC-Maze, mainly through improved raw-spike decoding. At the widest window tested, the gap on MC-Maze averages 0.05 R^2 and is positive in every replicate. At that window, readouts using Gaussian-smoothed spikes perform nearly as well as those using frozen features on MC-Maze and better on MC-RTT. The results support temporal integration as a substantial source of the model's decoding gain on both tasks and leave open whether its frozen features also encode more abstract behavioral structure.

1 INTRODUCTION

To evaluate pretrained neural representations, studies compare behavioral decoding from frozen features with that from raw recordings, using readouts with matched architectures (Wang et al., 2023; Zahorodnii et al., 2025). Better performance from frozen features can suggest that pretraining has learned a more abstract representation of behavior. Yet the two readouts may base their predictions on different spans of the recording. A backbone with trial-wide attention can integrate activity across a trial before passing features to a local readout (Ye & Pandarinath, 2021). The raw-spike readout, however, receives only activity inside its input window (Wessberg et al., 2000; Glaser et al., 2020). We call the backbone's use of activity outside the readout window temporal integration. A decoding gain due to this additional context need not imply a more abstract representation of behavior. We therefore ask how the measured gain changes as the input windows of both readouts widen.

We study BCIM, a brain–computer interface (BCI) foundation model pretrained by masked reconstruction on intracortical spikes and surface electrocorticography (ECoG), without behavioral supervision.[1] Its frozen features reach the highest listed point estimates on MC-Maze and Perich-CO (Table 4), and access to its backbone (Appendix A) lets us control how input tokens are assigned to output times (§3.2). If the gain comes mainly from context outside the readout window, widen-

[1] Code: `https://anonymous.4open.science/r/Mortis-25BC/`

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.

ing both windows should shrink it. On MC-RTT and MC-Maze,[2] we do so with time-aligned input tokens, a shared output-position code and a fixed parameter count within each representation. Smoothing gives a narrow window access to neighboring spike counts without pretraining, so we also measure the gain over Gaussian-smoothed spikes (Pei et al., 2021). We further randomize the context tokens within each trial to test whether the gain from widening depends on ordered local context (Khandelwal et al., 2018; O'Connor & Andreas, 2021). Under a separate protocol, we linearly predict the projected frozen features from raw spikes and trial position. We substitute these predictions into the head trained on frozen features to test whether decoding is preserved without retraining it (Lenc & Vedaldi, 2015; Bansal et al., 2021).

Widening the readout window reduces the frozen-feature gain by 89% on MC-RTT and 68% on MC-Maze, mainly through improved raw-spike decoding (Figure 1). Randomizing the context tokens lowers raw-spike decoding more than frozen-feature decoding. At the widest window, the mean gain on MC-Maze is 0.052 R^2 over raw spikes and 0.005 over Gaussian-smoothed spikes. On MC-RTT, smoothed spikes outperform frozen features by 0.050 on average, with this ordering in all five replicates (§4.2). Linear predictions preserve substantial decoding through the fixed head (§4.3). These findings are consistent with temporal integration as a substantial source of BCIM's decoding gain. Whether BCIM also learns more abstract behavioral structure remains unresolved (§4.4).

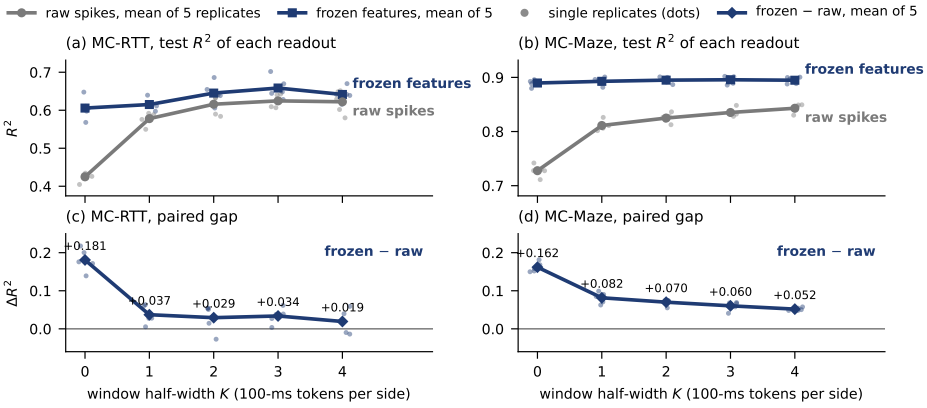

Figure 1: The frozen-feature advantage decreases as the readout window widens. Under the primary protocol, input tokens are aligned to output times and both readouts receive the same output-position code. For each channel, the readout receives a central token and K neighbors on each side, spaced 100 ms apart. Unused slots in a nine-slot layout are zeroed, keeping parameter count fixed within each representation. (a, b) Held-out R^2 on MC-RTT and MC-Maze. (c, d) Paired frozen-minus-raw gaps ΔR^2. Dots show five paired split-and-seed replicates, connected markers show means, and labels give mean gaps. See Table 9 (Appendix C) for values and §3.2 for the protocol.

Related work. Neural pretraining aims at transferable representations (Ortega Caro et al., 2024; Antoniades et al., 2024; Chau et al., 2025) from intracortical spikes (Ye et al., 2023; Zhang et al., 2024; Ryoo et al., 2025), scalp electroencephalography (EEG) (Jiang et al., 2024; Yang et al., 2023; Wang et al., 2024b) and intracranial field potentials (Wang et al., 2023; Zhang et al., 2023). Recent benchmarks find that spectrogram, classical or supervised baselines can match these foundation models (FMs) (Zahorodnii et al., 2025; Yang et al., 2026). Supervised baselines are also competitive with FMs in other specialized domains (Xu et al., 2025). Probe capacity also affects the interpretation of probing scores (Hewitt & Liang, 2019). Changing the probing head changes the ranking of speech representations (Zaiem et al., 2025). Replacing mean pooling with token flattening improves EEG classification with a frozen backbone (Širca et al., 2026). Kommineni et al. (2026) report a bias toward aperiodic and low-frequency components in representations from reconstruction-based EEG models. We build on classical temporal filtering of spike counts (Wiener, 1949; Wessberg et al., 2000). BCIM shares its masked-reconstruction objective (Devlin et al., 2019; He et al., 2022) with

[2]MC-Maze, Perich-CO, and MC-RTT are macaque intracortical neural-recording datasets used to evaluate how well neural signals predict reaching behavior.

NDT2 and MtM (Ye et al., 2023; Zhang et al., 2024), whereas POYO and NDT3 use behavioral supervision (Azabou et al., 2023; Ye et al., 2025). The objective matters because the filtering argument of §2 starts from reconstruction, whereas the readout comparisons also apply to models trained with other objectives, such as behavioral prediction (Appendix D).

2 BACKGROUND

We examine how widening readout windows and smoothing raw spikes affect the decoding gap between frozen features and raw spikes. An unrestricted readout given the full raw trial can implement the encoder followed by the downstream readout. The frozen representation therefore has no positive decoding advantage at the population optimum (Proposition 4, Appendix G). Because this argument applies to any deterministic encoder, it motivates a test of the readout window without identifying the encoder as a filter. Even the widest tested window covers only part of each trial, so a remaining advantage at that window can still reflect temporal integration.

2.1 RECONSTRUCTION AND A FILTERING BASELINE

The window argument holds for any encoder. For a model pretrained by masked reconstruction, the objective itself gives a further reason to expect temporal integration.

Proposition 1 (Population optimum of masked reconstruction)**.** *Let c denote the visible tokens and x a masked continuous target with $\mathbb{E}\|x\|^2 < \infty$. Under squared loss, the population-risk minimizer is $g^\star(c) = \mathbb{E}[x \mid c]$ (Banerjee et al., 2005). For spike counts Y with finite, positive conditional mean, Poisson negative log-likelihood is minimized by a log-rate prediction $\xi^\star(c)$ satisfying*

$$\exp \xi^\star(c) = \mathbb{E}[Y \mid c]. \tag{1}$$

Proofs, the latent-rate interpretation and boundary cases are in Appendix E.

For Poisson counts this conditional mean is the posterior-mean firing rate (Appendix E), so at its optimum reconstruction estimates the rate at a masked time from the visible counts, which include counts at other times. The proposition constrains this output and does not determine the intermediate features or imply a linear filter. A local model shows the form of such an estimate in a simple case. We assume small, stationary log-rate fluctuations z_t in a Poisson-lognormal model, with counts independent across bins given the latent trajectory, so that counts centered and scaled by their baseline mean $\bar{\lambda}$ are approximately z_t plus white noise with spectral density $1/\bar{\lambda}$. The best linear estimator of z is the non-causal Wiener filter $H_W(\omega) = S_z(\omega)/\big(S_z(\omega) + 1/\bar{\lambda}\big)$, where S_z is the power spectral density of z_t, and H_W is low-pass when S_z does not increase with $|\omega|$ (Equation 5, Appendix E). In this idealized case the best linear estimate of z is a temporal smoothing of the spike counts, which requires no pretraining. Smoothing the spikes before the readout gives even a narrow window access to neighboring counts, so we expect smoothed spikes to decode better than raw spikes, and frozen features to gain less over smoothed spikes than over raw spikes (§4.2). We test this with a Gaussian kernel whose only free parameter, the width, is selected on validation data.

2.2 FIXED-FILTER DISTORTION AND NOISE REDUCTION

Temporal estimation can reduce noise that remains after pooling across neurons when signal and noise have different spectra (Proposition 2, Appendix F). But a fixed smoothing kernel can also discard signal, so its benefit depends on what is being estimated. For an observation $y = b + \eta$ of a target b with spectrum S_b and noise η with spectrum S_η, assume both processes are zero-mean and wide-sense stationary, and are mutually uncorrelated at all lags. A fixed linear filter H changes the mean-square risk relative to the unfiltered estimate $\widehat{b}_{\mathrm{raw}} = y$ by

$$\mathcal{E}_H - \mathcal{E}_{\mathrm{raw}} = \int_\Omega \Big[|1 - H(\omega)|^2 S_b(\omega) - \big(1 - |H(\omega)|^2\big) S_\eta(\omega) \Big] \frac{d\omega}{2\pi}, \tag{2}$$

where Ω is the frequency domain. The two terms in the above integral are signal distortion and noise reduction, respectively (Proposition 3, Appendix F). Their balance depends on both the signal and noise spectra. In the cross-task comparisons (§3.3), the frozen-minus-raw gap is positive on the MC datasets and negative on Perich-CO. We therefore test whether changing the bandwidth of the behavioral target can reverse the sign of the gap (§4.4). The theoretical results concern estimation under the stated signal and noise models and do not predict the decoding gain of a fitted readout.

3 METHOD

All raw-spike and frozen-feature readouts are multilayer perceptrons (MLPs) with two hidden layers. The window sweep, the randomized-context control and the Gaussian comparison share one primary protocol, which matches time alignment and output-position information (§3.2). The remaining analyses use other protocols. Fixed-head substitution uses index alignment (Table 10), and the cross-task comparisons and the target-spectrum tests use their own partitions and readout settings.

3.1 MODEL, PRETRAINING AND DATA

The BCIM backbone uses a criss-cross Transformer architecture based on CBraMod (Wang et al., 2025). It shares its spike-and-ECoG pretraining setup with iBrain (Chen et al., 2026) and has 7.7 million parameters (6.5 million on the spike path). On the spike branch, we count spikes in 20-ms bins for each channel and group five bins into each 100-ms patch. We call a recorded spike channel a unit, and the theory models these units as neurons. A modality-specific two-layer MLP maps each patch to one token. We add learned time and modality embeddings before passing the tokens through six Transformer blocks. Each block applies attention across channels and then across time, with hidden width 256 and 8 attention heads. A final LayerNorm yields the hidden states Z, one 256-dimensional vector for each patch in each channel (Appendix B).

We pretrain BCIM by masked reconstruction, randomly selecting 15% of the token positions in each sample and denoting their index set by $\mathcal{I}$. We zero the input patches at those positions and use a modality-specific decoder Dec to reconstruct the corresponding signal patches from the visible context. The loss is Poisson negative log-likelihood (NLL) with a log link on the spike branch and squared loss on the ECoG branch,

$$\mathcal{J} = \sum_{i \in \mathcal{I}} \mathcal{J}_{\mathrm{mod}(i)}\big(\mathsf{Dec}(Z_i), x_i\big), \qquad \mathcal{J}_{\mathrm{spike}} = \text{Poisson NLL}, \quad \mathcal{J}_{\mathrm{cont}} = \tfrac{1}{2}\|\cdot\|^2, \tag{3}$$

where $Z_i \in \mathbb{R}^{256}$ is the hidden state at token index i; x_i is its reconstruction target (five 20-ms bins on the spike branch); and $\mathrm{mod}(i)$ is the token's modality. Equation 3 is the sole pretraining objective, with no contrastive, alignment or behaviorally supervised term. Pretraining uses the primate subset of the Neural Pile (Orhan & Wang, 2025) for intracortical spikes and AJILE12 (Peterson et al., 2022) for surface ECoG. We cap each source dataset at 1,000 hours, and only AJILE12 reaches this limit. The pretraining corpus includes the source datasets of MC-RTT and the Perich tasks, with session-level overlap unverified (Appendix D). Our comparisons examine how readout conditions affect decoding for this checkpoint. For downstream decoding, we run the frozen encoder with observed tokens unmasked and padded channels masked. Downstream spike analyses use MC-RTT and MC-Maze from the Neural Latents Benchmark (Makin et al., 2018; Churchland et al., 2012; Pei et al., 2021) and the Perich center-out (CO) and random-target (RT) tasks (Perich et al., 2018), with neural counts binned at 20 ms. Perich-RT uses fixed-duration windows from continuous recordings, and Perich-CO uses reach-aligned windows, one window per reach.

3.2 READOUT EXPERIMENTS

We refit readouts using frozen features, raw patches and targets cached from the downstream *training* split of MC-RTT and MC-Maze. We split these trials into 70% training, 15% validation and 15% test data in each replicate. With one recording per dataset, replicates capture split and optimization variability, not subject or session variability. Conditions share each partition, giving paired score differences. Absolute scores are not directly comparable to those from the cross-task comparisons (§3.3) or the benchmark entries of Table 4.

Downstream readout-window sweep. If frozen features mainly supply temporal context from outside the readout window, widening both readout windows should reduce the frozen-minus-raw gap. The readout first projects each channel-token to d_{tok} dimensions with one linear map shared across channels and positions. At each of $T = 10$ output positions, it concatenates projected tokens from all channels within the window. The window includes a central position and $K \in \{0, \ldots, 4\}$ neighbors per side, spaced 100 ms apart, and K is the half-width in token steps. To vary context at fixed parameter count within each representation, we keep a nine-slot layout at $d_{\mathrm{tok}} = 8$ and zero the

slots outside $\pm K$. Raw and frozen readouts differ in parameter count because their input dimensions differ. For each sweep condition and replicate, we select the best of three fits by validation R^2 and measure the test-set gap $\Delta_{\mathrm{FM-raw}}(K) := R^2_{\mathrm{FM}}(K) - R^2_{\mathrm{raw}}(K)$, where the subscript FM denotes the frozen-feature condition (Appendix C).

Temporal correspondence and output position. The readout predicts velocity at ten output times (knots) s_k, $k = 0, \ldots, 9$, within the scored time span and interpolates these predictions linearly. Input token j is centered at time t_j on the same clock (Appendix C). The primary protocol (time-aligned + position) centers the window for knot k on the nearest token, $c(k) = \arg\min_j |t_j - (s_k + \tau_{\mathrm{lag}})|$ with prespecified $\tau_{\mathrm{lag}} = 0$ ms, and gives both readouts the same one-hot code for the output-knot index, $U_k = e_k$. Token centers are 100 ms apart and knots are 60 or 70 ms apart, so some adjacent knots receive the same input window. Without the code, a readout with weights shared across knots makes identical predictions for them. Under index alignment, the window for knot k is centered on token $c(k) = k$. Because the neural window extends beyond the scored span to include forward context, the input-token center then occurs up to 383 ms (MC-RTT) or 288 ms (MC-Maze) after the output time s_k. A third protocol, time alignment without the position code, uses the nearest-token correspondence of the primary protocol and leaves the knots, targets and interpolation unchanged (offsets and quantization errors in Appendix C). All three protocols share the masked nine-slot layout at $d_{\mathrm{tok}} = 8$, the splits and the selection rule. Parameter count is fixed across K within each representation and changes across protocols only through the shared one-hot output-position code U_k of the primary protocol.

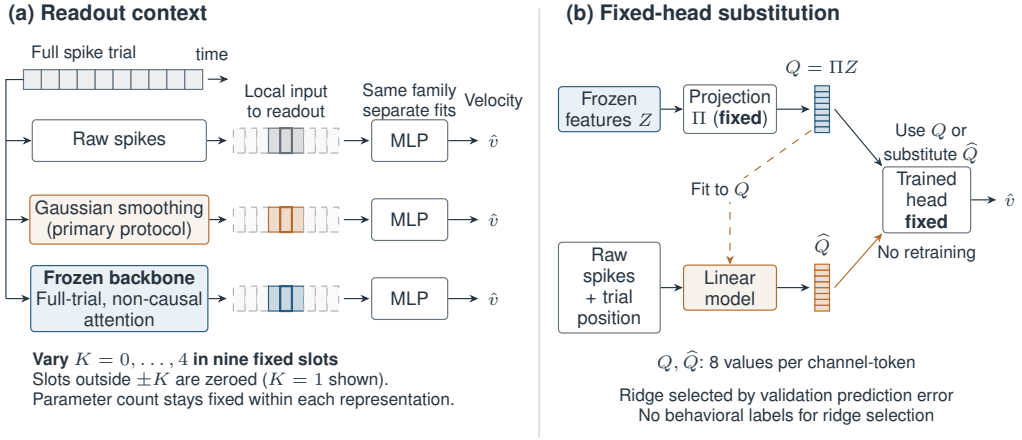

Figure 2: (a) The frozen backbone attends over the full trial before the readout receives a local feature window. The raw readout directly accesses only its local spike window. The primary sweep varies these windows at fixed parameter count within each representation. Gaussian-smoothed spikes are evaluated under the same primary readout protocol. (b) Linear models predict the projected frozen features Q (eight values per channel-token) from raw spikes and trial position. The dashed arrow indicates fitting to Q on training trials. Their predictions replace Q with the projection and trained head held fixed (§4.3). Ridge strength is selected by prediction error on validation trials.

Gaussian baselines. We fit separate readouts to spikes smoothed with a two-sided (non-causal) Gaussian kernel of width σ or a bank of kernels at several widths, using prespecified grids. We smooth each cached trial before assembling readout windows, so even at $K = 0$ the inputs depend on spikes outside the readout window (Figure 2). An earlier exploratory comparison selected σ on test data and then held it fixed. In the validation-selected rerun, we randomly partition trials into training, validation and test sets for each replicate, using the same partition across conditions. Within each replicate, we fit three readouts per candidate σ, select width and readout by validation R^2, and score that fit once on the test split. The primary-protocol comparison repeats this selection at $K = 0$ and $K = 4$ over five paired replicates. Intervals describe split and optimization variation within one recording, and Appendix C gives paired-t intervals over the same replicates.

Context and readout controls. To test whether the gain from widening depends on ordered local context, we replace neighboring tokens with tokens sampled from within the same trial. At $K = 4$, we keep the time-aligned center token, output-position code, slot mask and parameter count fixed. We sample positions for the other eight slots once per trial, use the same positions for both representations, and retrain both readouts. Supporting sweeps under index alignment either keep $d_{\text{tok}} = 8$ as the input window grows with K (fixed width) or set $d_{\text{tok}} = \max\{1, \text{round}(8/(2K + 1))\}$ to keep parameter count approximately constant (approximate matching). The common width $d_{\text{tok}} = 8$ compresses the 256-dimensional frozen tokens and expands the five-dimensional raw patches, so we also test the decoding gap's sensitivity to projection width in the index-aligned sweeps and, at $d_{\text{tok}} = 32$, under the primary protocol (§4.1, Appendix C). Within the index-aligned sweeps, we treat the frozen-versus-Gaussian ordering as secondary because it depends on projection width and the parameter-count setting. On MC-RTT at $K = 3$, we also test whether feature width or backbone architecture alone explains the gain, using raw tokens randomly projected to 256 dimensions and features from an untrained backbone.

Target-spectrum controls. We filter velocity within trials, refit the readouts to each filtered target, and test whether changing the target spectrum can change the sign of the decoding gain. We measure gains separately in the disjoint $[0, 2), [2, 4), [4, 6), [6, 8)$ Hz bands. A complementary low-pass sweep retains frequencies below cutoffs selected on training data to preserve prespecified fractions of within-trial target variance. On the MC datasets, we compare raw spikes, spikes smoothed with a fixed Gaussian kernel of width σ, and frozen features. Because filtering changes the R^2 denominator, we also report a normalized error reduction D_n. We subtract the Gaussian or frozen-feature mean-squared error (MSE) from the raw-spike MSE and divide by the unfiltered target variance. Here n indexes the target manipulation, and both MSEs use the same filtered target. On Perich-RT, we compare input smoothing at $\sigma = 6$ bins with the benchmark-matched $\sigma = 2$ baseline, without a frozen-feature condition, using the approximation to D_n described in Appendix C.

3.3 Cross-task comparisons

In the cross-task comparisons, we decode velocity from frozen features and raw spikes using MLPs with two hidden layers, matched temporal windows and unequal parameter counts. Each window includes three frozen-feature tokens or 15 raw 20-ms bins per side, about ± 300 ms. These offline comparisons use symmetric context. We retain all fits and measure uncertainty across three readout seeds per MC task and six paired sessions per Perich task.

4 Results

4.1 Readout-window intervention

Under the primary protocol (time alignment with the shared output-position code), widening the readout window from $K = 0$ to $K = 4$ reduces the mean frozen-minus-raw R^2 gap from $+0.181$ to $+0.019$ on MC-RTT and from $+0.162$ to $+0.052$ on MC-Maze (Figure 1; Table 9, Appendix C). These are reductions of 89% and 68%, with a decrease in all five paired replicates per dataset. On MC-RTT, where the neural windows of adjacent trials overlap, a chain-grouped split that keeps adjacent trials in the same set gives a mean gap of $+0.176$ at $K = 0$ and -0.009 at $K = 4$, again with a decline in all five replicates (Appendix D). The gap also declines in every replicate on both datasets at $d_{\text{tok}} = 32$ under the primary protocol. This pattern persists when velocity is predicted directly at each 5-ms bin without knot interpolation (Appendix C).

In the primary sweep, most of the gap reduction comes from improved raw-spike decoding. On MC-RTT one token of context per side already brings the gap to $+0.037$, and further widening moves it by at most 0.018. On MC-Maze raw decoding improves and the gap shrinks at every step, whereas frozen-feature decoding stays within 0.006 of its narrow-window value. At the widest window the gap has mixed signs across MC-RTT replicates (-0.014 to $+0.059$), whereas all five MC-Maze differences remain positive ($+0.048$ to $+0.058$). The 89% and 68% reductions quantify changes in the decoding gap, not the causal contribution of temporal integration. Randomizing context under the primary protocol lowers both decoding scores, with a larger raw drop in every replicate. On average the raw drop exceeds the frozen-feature drop by 0.075 on MC-RTT and 0.057 on MC-Maze

(Table 1). Ordered local context therefore raises raw decoding more than frozen-feature decoding once time correspondence and output position are matched. This is consistent with the backbone integrating temporal context before the readout. The raw drop also exceeds the frozen-feature drop under index alignment at $K = 4$ with fixed projection width (Figure 7b, Appendix C). Random tokens may also act as distractors.

Table 1: Ordered and randomized context at $K = 4$ under the primary protocol. Entries are means over five paired replicates (Table 9). Drops are ordered minus randomized R^2, the gap column is frozen minus raw R^2 under randomized context, and the last column is raw drop minus frozen drop. Parentheses give the number of replicates out of five with a positive value.

	raw			frozen				
Dataset	ordered	random	drop	ordered	random	drop	gap (random)	raw drop − frozen drop
MC-RTT	0.622	0.411	+0.211	0.642	0.505	+0.136	+0.094 (5/5)	+0.075 (5/5)
MC-Maze	0.843	0.686	+0.157	0.895	0.795	+0.100	+0.109 (5/5)	+0.057 (5/5)

The narrow-window gap also depends on temporal alignment and output-position information (Figure 6, Appendix C). Switching from index alignment to time alignment reduces the gap by 0.123 on MC-RTT, and adding the position code then increases it by 0.010. On MC-Maze, these changes reduce the gap by 0.035 and a further 0.059. The gap also declines with window width in the index-aligned sweeps under all three readout settings. In the masked sweeps, the decline holds in every replicate in all seven dataset-by-width conditions, and its magnitude varies with projection width (Appendix C). In these sweeps only the backbone receives a learned time embedding. A wider raw window may let the readout infer trial position, so the gap could decline even if trial position were the only extra information supplied by the frozen features. Under index alignment, the one-hot output-position code U_k (§3.2) also identifies the central token. Giving both readouts this code preserves the decline in all twelve paired replicates per dataset (Table 17). At $K = 0$, adding the code reduces the gap by 26% on MC-Maze and increases it by 6% on MC-RTT.

4.2 Gaussian smoothing under the primary protocol

Under the primary protocol, the mean frozen-minus-Gaussian gap at $K = 0$ is 0.005 on both datasets (Table 2). At $K = 4$, smoothed spikes outperform frozen features on MC-RTT in all five replicates, by 0.050 on average. Mean decoding R^2 from smoothed spikes on MC-RTT rises from 0.601 at $K = 0$ to 0.691 at $K = 4$, so smoothing does not remove the benefit of wider context. On MC-Maze at $K = 4$, frozen features retain a mean advantage of 0.005 over smoothed spikes, compared with 0.052 over raw spikes in the full window sweep. Under the chain-grouped split on MC-RTT, the mean frozen-minus-Gaussian gap is $+0.029$ at $K = 0$ and -0.051 at $K = 4$ (Appendix D). The readouts are fitted separately, and their score differences do not quantify how much additional behavioral information the frozen features contain beyond smoothed spikes.

Table 2: Gaussian smoothing and frozen features under the primary protocol. Entries are mean test R^2 over five paired replicates, with kernel width selected on validation data in each replicate. The last column gives the mean paired frozen-minus-Gaussian gap, followed by its 95% percentile-bootstrap interval in brackets and, in parentheses, the number of replicates out of five in which the gap is positive. Raw and frozen readouts are refitted on shared splits and seeds. At $K = 4$, their means differ from those in Table 1 by at most 0.003. The 89% and 68% reductions use only the fits from the full window sweep (Figure 1).

Dataset	K	raw	Gaussian	frozen	frozen − Gaussian
MC-RTT	0	0.425	0.601	0.606	$+0.005$ $[-0.003, +0.012]$ (3/5)
MC-RTT	4	0.625	0.691	0.641	-0.050 $[-0.069, -0.037]$ (0/5)
MC-Maze	0	0.728	0.885	0.890	$+0.005$ $[+0.001, +0.009]$ (4/5)
MC-Maze	4	0.843	0.890	0.895	$+0.005$ $[-0.001, +0.009]$ (4/5)

Table 3: Linear substitution into the fixed post-projection head at $K = 0$, with index alignment and no output-position input (means over twelve paired replicates per dataset). Decoding R^2 is measured without retraining the head after substitution, and EV is held-out explained variance. The last column gives the paired $M_1 - M_{1\mathrm{ch}}$ decoding increment and its 95% percentile-bootstrap interval.

	decoding R^2			$M_{1\mathrm{ch}}$ EV		M_1 EV		increment
Dataset	frozen	$M_{1\mathrm{ch}}$	M_1	EV(Q)	EV(pre)	EV(Q)	EV(pre)	$M_1 - M_{1\mathrm{ch}}$
MC-RTT	0.491	0.201	0.355	0.323	0.310	0.282	0.442	+0.154 [0.132, 0.171]
MC-Maze	0.883	0.720	0.852	0.445	0.415	0.454	0.618	+0.132 [0.125, 0.138]

4.3 Linear substitution into a fixed head

The Gaussian comparison fits separate readouts, allowing each to adapt to its inputs. We test whether linear predictions of the projected frozen features can support decoding through the already-trained head. Under index alignment, we substitute these predictions without retraining the head. Cross-channel linear predictions preserve substantial decoding through the fixed head.

The models use raw spikes and trial position to predict the projected frozen features $Q = \Pi Z$, with eight values per channel-token. Here $\Pi \in \mathbb{R}^{8\times 256}$ is the readout projection at $d_{\mathrm{tok}} = 8$, with its bias omitted from the notation. At each position, the head concatenates these features across channels before its first linear layer. We substitute the predictions $\widehat{Q}$ for Q. Since errors in different directions of Q need not affect the head equally, we also compare the resulting first-layer outputs before activation, $\mathrm{pre}(\widehat{Q})$ and $\mathrm{pre}(Q) \in \mathbb{R}^{256}$. We fit ridge regression models on training trials in three prespecified classes, each including a position-dependent template. The template M_0 uses trial position alone. For $M_{1\mathrm{ch}}$, each channel's prediction also uses that channel's spikes at token lags $-2, -1, 0, 1, 2$, with spike coefficients shared across positions, as in Gaussian smoothing. The cross-channel map M_1 uses the same lag support and coefficient-sharing constraint and additionally allows inputs from other channels. We choose ridge strength by minimizing validation MSE between $\widehat{Q}$ and Q, without behavioral targets. We report explained variance and fixed-head decoding at $K = 0$ under index alignment, without an output-position input to the head (Table 3).

Although M_1 explains less than half the variance in Q, decoding after substitution reaches 0.355 R^2 versus 0.491 for frozen features on MC-RTT and 0.852 versus 0.883 on MC-Maze (Table 3). Decoding R^2 is higher with M_1 than with channelwise substitution in every paired replicate. With the same lag support and fixed head, including other channels' spikes therefore improves decoding after substitution. The position-only template M_0 yields negative decoding R^2 on both datasets when substituted into the head (Appendix C). Lower decoding after substitution may reflect a changed input distribution rather than missing behavioral information. Explained variance in Q and decoding performance rank these models differently (Appendix C). The decoding advantage of M_1 over $M_{1\mathrm{ch}}$ also holds for untrained backbones in all 48 combinations of replicate and initialization per dataset. This ordering alone therefore does not identify structure acquired during pretraining (Appendix C).

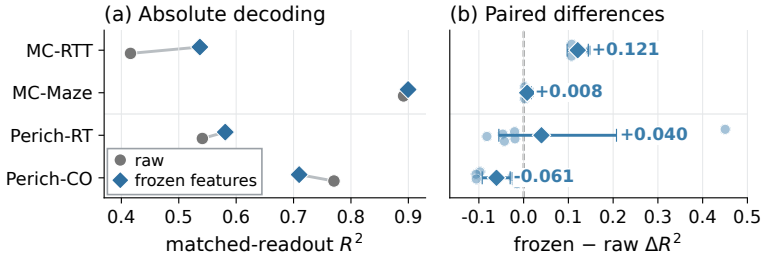

Figure 3: Cross-task comparisons. (a) Mean raw and frozen-feature R^2. (b) Paired frozen-minus-raw differences ΔR^2. In (b), dots show individual MC readout seeds or Perich sessions, and diamonds show their means. Whiskers show the sample standard deviation (SD) over three MC seeds or paired-session bootstrap 95% intervals over six Perich sessions.

4.4 RESULTS THAT BOUND THE ACCOUNT

Cross-task comparisons. With readout family and window matched within each task, frozen features decode better than raw spikes in all three readout seeds on both MC datasets. On Perich-CO, the mean frozen-minus-raw gap is negative and its session-level confidence interval (CI) excludes zero (Figure 3). Retaining all Perich-RT fits gives a mean gap of $+0.040$ (95% interval $[-0.055, +0.208]$), leaving it unclear which representation decodes better on average. Excluding one raw fit based on its test score changes the sign of that mean (Appendix B). The tasks differ in subject, array, behavior, trial structure and readout setting, so their comparison does not isolate the cause of these performance differences across the four tasks.

Target-spectrum tests. On the two MC datasets, restricting velocity to disjoint frequency bands does not make the mean frozen-minus-raw gain negative, so these tests do not reproduce the cross-task sign change. Output interpolation attenuates high frequencies, with -3 dB points at 5.32 Hz on MC-RTT and 4.56 Hz on MC-Maze (Table 11). This limits the test's sensitivity, so a non-negative gain does not exclude attenuation by Gaussian smoothing or the backbone in the higher bands. To interpret a negative gain as loss of decodable signal, the raw readout must first decode that band above $R^2 = 0$. Raw decoding exceeds zero on MC-Maze in the three bands below 6 Hz, but is near zero on MC-RTT above 2 Hz. On MC-Maze, both spikes smoothed with the fixed Gaussian kernel and frozen features yield a positive mean D_n in every band, and a low-pass target sweep gives no sign change (Tables 19 and 18, Figure 8).

Representation controls and probes. In the separate MC-RTT control at $K = 3$, frozen features outperform randomly projected raw spikes by 0.159 R^2 and untrained-backbone features by 0.252 (Table 22). The latter comparison shows a pretraining benefit under this readout. Additional probes examine whether BCIM also learns abstract behavioral structure. The residual-feature probe gives an imprecise estimate of the decoding benefit beyond spikes smoothed at one Gaussian width. Decoding from the Gaussian features F_σ together with the residual $F - \widehat{F}(F_\sigma)$, where the ridge map $\widehat{F}$, fitted on training trials, predicts frozen features F from F_σ (Appendix B), gives a held-out R^2 increment over F_σ alone of $+0.023$ on MC-RTT (CI $[-0.073, +0.143]$) and -0.001 on Perich-RT (CI $[-0.037, +0.036]$). The residual features alone and the rates reconstructed by the pretraining decoder decode worse on average than raw spikes on MC-RTT and Perich-RT. Speed and trial position remain entangled in the features, and zero-shot cross-session transfer gives negative R^2 (Table 5, Appendix B). These probes neither establish nor exclude added abstraction, and decoding gains alone do not resolve the question for this model and these tasks.

5 DISCUSSION AND CONCLUSION

With temporal alignment and output-position information matched, widening the readout window reduces BCIM's decoding advantage by 89% on MC-RTT and 68% on MC-Maze at fixed parameter count within each representation. Most of the gap reduction comes from improved raw decoding. Randomizing context lowers raw decoding more than frozen-feature decoding in all five replicates (§4.1). On MC-Maze, the frozen-minus-raw gap remains positive in every replicate at the widest window, averaging 0.052 R^2, about a third of the variance unexplained by raw decoding ($1 - 0.843$). Relative to Gaussian-smoothed spikes, the mean advantage is 0.005, and on MC-RTT smoothed spikes outperform frozen features by 0.050 on average at this window (§4.2). The remaining advantage therefore depends on the temporal baseline, as well as the token–output mapping (Appendix C).

Readout windows remain constrained by the token grid and trial boundaries, whereas the backbone attends over the full trial and Gaussian smoothing can use spikes outside the readout window. These comparisons use different temporal supports and do not identify the backbone's computation or distinguish band-selective denoising from latent-state estimation. We study one model offline, and added abstraction remains unresolved (§4.4, Appendix D). Within these limits, the results support temporal integration as a substantial source of BCIM's decoding advantage. Evaluations of frozen features should therefore report the window and temporal preprocessing of the raw-spike baseline alongside the readout architecture, since both changed the measured gain in these tasks.

AI USE STATEMENT

In this work, we used generative AI tools, namely Claude (Anthropic; Claude Opus 5 and Claude Fable 5.1 through Claude Code) and ChatGPT (OpenAI; GPT-5.6 Sol and GPT-6 Astra), for the following tasks with required disclosure. For the theoretical results, the authors formulated the claims and produced the proofs, and the tools assisted in formalizing the claims and in writing out the proofs. For the experiments, the authors formulated the central research question, designed the primary readout protocol and made the final experimental decisions. The tools gave feedback on these designs, suggested and refined robustness controls, implemented the analysis and figure-generation code to the authors' specification, and assisted the authors in discussing the interpretation of the experimental results. The authors evaluated every suggestion and verified the resulting analyses. We have not used generative AI tools to generate synthetic data sets, to develop theoretical models or conceptual frameworks, to provide critical ingredients for proving mathematical claims, to assist with translation, or to clean or reformat data sets, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: to create and edit software code, to manage experiment runs, to edit the paper for readability, and to format references. We have reviewed all AI-assisted work: every derivation was checked by the authors against its stated assumptions; all code was reviewed, executed and tested by the authors; every reported number is backed by a retained per-replicate output or source log; and no data were collected or generated by an LLM at any stage. Every AI-assisted derivation, result and passage was reviewed and edited by the authors. Model identifiers and session records are retained by the authors and available on request. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is a secondary analysis of previously collected human intracranial and nonhuman-primate recordings and involved no new human or animal data collection. The original approvals, informed-consent procedures for human recordings and animal oversight are documented by the cited source studies. We followed the applicable dataset access and use conditions.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices B, C and F specify the datasets, preprocessing, readouts, splits, uncertainty units, the validation-based selection rule and every reported experimental and simulation setting, and Appendices E, F and G state the assumptions and give complete proofs. The cross-task comparisons of §3.3 and the readout experiments of §3.2 are analyzed separately, the interpretation of each readout experiment rests on paired comparisons within one protocol, and absolute scores are not compared across the two or against the published entries of Table 4. Applying the readout experiments to another frozen temporal model requires only aligned raw inputs, accessible frozen features and a controllable readout window, with no access to model internals. Numerical claims are traceable to retained per-replicate outputs and source logs, the figure-generation scripts contain the plotted summary values, and Table 10 indexes the readout experiments and their controls. We release the analysis code, the per-replicate outputs behind every reported number, the pretraining scripts and the data identifiers at `https://anonymous.4open.science/r/Mortis-25BC/`. Pretraining with the released scripts takes less than one day on six NVIDIA RTX PRO 6000 GPUs.

REFERENCES

F. J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3–4):246–254, 1948. doi: 10.1093/biomet/35.3-4.246.

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and multitask generative pretraining for brain data. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2311.00136`.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scal-

able framework for neural population decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. doi: 10.52202/075280-1948.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005. doi: 10.1214/009117905000000233.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 225–236, 2021.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011. doi: 10.1016/j.aim.2011.02.007.

Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population Transformer: Learning population-level representations of neural activity. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/pdf?id=FVuqJt3c4L.

Ying Chen, Tiou Wang, and Zhifeng Yue. iBrain: A unified foundation model reading the brain from surface to spikes. arXiv preprint arXiv:2609.06960, 2026. URL https://arxiv.org/abs/2609.06960.

Yuzhi Chen, Wilson S. Geisler, and Eyal Seidemann. Optimal temporal decoding of neural population responses in a reaction-time visual detection task. *Journal of Neurophysiology*, 99(3): 1366–1379, 2008. doi: 10.1152/jn.00698.2007.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Raeed H. Chowdhury, Joshua I. Glaser, and Lee E. Miller. Area 2 of primary somatosensory cortex encodes kinematics of the whole arm. *eLife*, 9:e48198, 2020. doi: 10.7554/eLife.48198.

Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012. doi: 10.1038/nature11129.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171–4186, 2019. URL https://arxiv.org/abs/1810.04805.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Joshua I. Glaser, Ari S. Benjamin, Raeed H. Chowdhury, Matthew G. Perich, Lee E. Miller, and Konrad P. Kording. Machine learning for neural decoding. *eNeuro*, 7(4):ENEURO.0506–19.2020, 2020. doi: 10.1523/ENEURO.0506-19.2020.

Charles Guan, Tyson Aflalo, Carey Y. Zhang, Elena Amoruso, Emily R. Rosario, Nader Pouratian, and Richard A. Andersen. Stability of motor representations after paralysis. *eLife*, 11:e74478, 2022. doi: 10.7554/eLife.74478.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022. URL https://arxiv.org/abs/2111.06377.

David J. Herzfeld and Scott A. Beardsley. Improved multi-unit decoding at the brain-machine interface using population temporal linear filtering. *Journal of Neural Engineering*, 7(4):046012, 2010. doi: 10.1088/1741-2560/7/4/046012.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Binjie Hong, Rui Xiong, Liyuan Han, and Tielin Zhang. UniBCI: Towards a unified pretrained model for invasive brain-computer interfaces. arXiv preprint arXiv:2605.00061, 2026. URL https://arxiv.org/abs/2605.00061.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/pdf?id=QzTpTRVtrP.

Ingmar Kanitscheider, Ruben Coen-Cagli, and Alexandre Pouget. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences*, 112(50):E6973–E6982, 2015. doi: 10.1073/pnas.1508738112.

Brianna M. Karpowicz, Yahia H. Ali, Lahiru N. Wimalasena, Andrew R. Sedler, Mohammad Reza Keshtkaran, Kevin Bodkin, Xuan Ma, Daniel B. Rubin, Ziv M. Williams, Sydney S. Cash, Leigh R. Hochberg, Lee E. Miller, and Chethan Pandarinath. Stabilizing brain-computer interfaces through alignment of latent dynamics. *Nature Communications*, 16(1):4662, 2025. doi: 10.1038/s41467-025-59652-y.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 284–294, 2018. doi: 10.18653/v1/P18-1027.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6114.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Simon Bock Segaard, Jeppe Roden Münster, Andreas Peter Juhl Hansen, Takfarinas Medani, Tiantian Feng, Richard Leahy, and Shrikanth Narayanan. Aperiodic and low-frequency spectral bias in reconstruction based EEG foundation models. arXiv preprint arXiv:2605.26434, 2026. URL https://arxiv.org/abs/2605.26434.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015. doi: 10.1109/CVPR.2015.7298701.

Joseph G. Makin, Joseph E. O'Doherty, Mariana M. B. Cardoso, and Philip N. Sabes. Superior arm-movement decoding from cortex with a new, unsupervised-learning algorithm. *Journal of Neural Engineering*, 15(2):026010, 2018. doi: 10.1088/1741-2552/aa9e95.

Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417, 2014. doi: 10.1038/nn.3807.

Joe O'Connor and Jacob Andreas. What context features can Transformer language models use? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 851–864, 2021. doi: 10.18653/v1/2021.acl-long.70.

A. Emin Orhan and Feiyi Wang. The Neural Pile: 476 billion tokens of broad-coverage spiking neural activity data. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*, 2025. URL https://openreview.net/forum?id=NKFxs7UvN9.

Josue Ortega Caro, Antonio H. de O. Fonseca, Syed A. Rizvi, Matteo Rosati, Christopher Averill, James L. Cross, Prateek Mittal, Emanuele Zappala, Rahul M. Dhodapkar, Chadi G. Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings. In *International Conference on Learning Representations (ICLR)*, 2024.

Leslie C. Osborne, William Bialek, and Stephen G. Lisberger. Time course of information about motion direction in visual area MT of macaque monkeys. *Journal of Neuroscience*, 24(13):3210–3222, 2004. doi: 10.1523/JNEUROSCI.5305-03.2004.

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.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural Latents Benchmark '21: Evaluating latent variable models of neural population activity. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2109.04463.

Matthew G. Perich, Juan A. Gallego, and Lee E. Miller. A neural population mechanism for rapid learning. *Neuron*, 100(4):964–976, 2018. doi: 10.1016/j.neuron.2018.09.030.

Steven M. Peterson, Satpreet H. Singh, Benjamin Dichter, Michael Scheid, Rajesh P. N. Rao, and Bingni W. Brunton. AJILE12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific Data*, 9:184, 2022. Dataset on DANDI: 000055.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 157–163. University of California Press, 1956.

Avery Hee-Woon Ryoo, Nanda H. Krishna, Ximeng Mao, Mehdi Azabou, Eva L. Dyer, Matthew G. Perich, and Guillaume Lajoie. Generalizable, real-time neural decoding with hybrid state-space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-1725.

Urban Širca, Maryam Alimardani, Stefanos Zafeiriou, and Konstantinos Barmpas. Beyond accuracy: Robustness, interpretability and expressiveness of EEG foundation models. arXiv preprint arXiv:2605.17562, 2026. URL https://arxiv.org/abs/2605.17562.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=xmcYx_reUn6.

Christopher Wang, Adam Uri Yaari, Aaditya K. Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph R. Madsen, Scellig Stone, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brain Treebank: Large-scale intracranial recordings from naturalistic language stimuli. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, volume 37, 2024a. doi: 10.52202/079017-3060.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024b. doi: 10.52202/079017-1239.

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

Johan Wessberg, Christopher R. Stambaugh, Jerald D. Kralik, Pamela D. Beck, Mark Laubach, John K. Chapin, Jung Kim, S. James Biggs, Mandayam A. Srinivasan, and Miguel A. L. Nicolelis. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408 (6810):361–365, 2000. doi: 10.1038/35042582.

Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. The MIT Press, Cambridge, MA, 1949.

Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. Specialized foundation models struggle to beat supervised baselines. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=JYTQ6ELUVO.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. doi: 10.52202/075280-3420.

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

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with Neural Data Transformers. *Neurons, Behavior, Data analysis, and Theory*, 5(3):1–18, 2021. doi: 10.51628/001c.27358.

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

Joel Ye, Fabio Rizzoglio, Adam Smoulder, Hongwei Mao, Xuan Ma, Patrick Marino, Raeed Chowdhury, Dalton Moore, Gary Blumenthal, William Hockeimer, Nicolas G. Kunigk, J. Patrick Mayo, Aaron P. Batista, Steven Chase, Adam Rouse, Michael L. Boninger, Charles Greenspon, Andrew B. Schwartz, Nicholas G. Hatsopoulos, Lee E. Miller, Kristofer E. Bouchard, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. A generalist intracortical motor decoder. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-3692.

Andrii Zahorodnii, Christopher Wang, Geeling Chau, Bennett Stankovits, Charikleia Moraitaki, Eli Gross, Alexander Brady, Andrei Barbu, Boris Katz, and Ila R. Fiete. Neuroprobe: Evaluating intracranial brain responses to naturalistic stimuli. arXiv preprint arXiv:2509.21671, 2025. URL https://arxiv.org/abs/2509.21671.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. Speech self-supervised representations benchmarking: A case for larger probing heads. *Computer Speech & Language*, 89:101695, 2025. doi: 10.1016/j.csl.2024.101695.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. doi: 10.52202/075280-1144.

Yizi Zhang, Yanchen Wang, Donato Jiménez-Benetó, Zixuan Wang, Mehdi Azabou, Blake Richards, Renee Tung, Olivier Winter, International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 80495–80521, 2024.

A Related work

This appendix reviews related baselines, neural foundation models and the filtering theory used in §2. *Classical and supervised baselines.* On intracranial recordings, Zahorodnii et al. (2025) find an optimized spectrogram readout competitive with foundation models. On scalp EEG, Yang et al. (2026) report that foundation models often do not outperform compact neural or classical readouts, especially in data-scarce settings. Across other specialized domains, Xu et al. (2025) find supervised baselines competitive with foundation models. Širca et al. (2026) report that token flattening in place of mean pooling improves EEG classification with a frozen backbone. These studies motivate better baselines and readouts. Here we ask why one frozen model improves decoding on intracortical spikes.

Neural foundation models. Self-supervised neural FMs pair Transformers (Vaswani et al., 2017) with masked objectives developed in BERT and masked autoencoders (Devlin et al., 2019; He et al., 2022). Scalp EEG models include LaBraM and BIOT (Jiang et al., 2024; Yang et al., 2023), EEGPT (Wang et al., 2024b), and CBraMod (Wang et al., 2025). Neural pretraining also covers intracranial field potentials (Wang et al., 2023; Zhang et al., 2023), functional magnetic resonance imaging (fMRI) (Ortega Caro et al., 2024), and multimodal or population recordings (Antoniades et al., 2024; Chau et al., 2025). Our spike-decoding comparisons are closest to work on spike-token models, NDT2 (Ye et al., 2023), its generalist successor (Ye et al., 2025), POYO (Azabou et al., 2023), MtM (Zhang et al., 2024), and real-time variants (Ryoo et al., 2025). A separate line infers single-trial population dynamics with sequential autoencoders rather than by pretraining (Pandarinath et al., 2018), and supplies several of the benchmark baselines in Table 4. These studies contribute models; we examine one frozen model's decoding gains and losses. Comparing another model's frozen features with raw spikes requires extracting its features and matching them to output times. The candidate NDT2 checkpoint was pretrained on Indy random-target recordings from the same animal and task as MC-RTT. Readout comparisons using this checkpoint remain possible; claims about unseen-session transfer require checking which recordings were used in pretraining. The candidate MtM model was pretrained on mouse Neuropixels recordings. Applying it to macaque motor cortex requires fitting new unit embeddings, an adaptation we have not validated here. POYO and NDT3 use behavioral supervision, so the reconstruction premise does not apply to them.

Classical ingredients. The established ingredients include the connection between denoising autoencoders and the score (Vincent, 2011); Tweedie and empirical-Bayes identities (Robbins, 1956; Efron, 2011); Bregman means (Banerjee et al., 2005); the Wiener filter and Anscombe stabilization (Wiener, 1949; Anscombe, 1948); information-limiting correlations (Moreno-Bote et al., 2014; Kanitscheider et al., 2015); and the spatial-versus-temporal coding-limit literature (Osborne et al., 2004; Chen et al., 2008; Herzfeld & Beardsley, 2010). We apply these results to a frozen invasive foundation model and test the resulting filtering account empirically.

B Benchmark and protocol details

The empirical pattern is split across tasks. Table 4 summarizes it across heterogeneous protocols: the BCIM backbone is frozen, and the baseline results use the implementations and checkpoints evaluated by Hong et al. (2026). The per-task gains and shortfalls motivate the cross-task comparison but do not isolate feature or head effects. We report per-task numbers and no average.

Abbreviations used in Table 4. WF, Wiener filter; GRU, gated recurrent unit (Cho et al., 2014); MLP, multilayer perceptron; VAE, variational autoencoder (Kingma & Welling, 2014); NoMAD, nonlinear manifold alignment with dynamics (Karpowicz et al., 2025); NDT1 and NDT2, Neural Data Transformer 1 (Ye & Pandarinath, 2021) and 2 (Ye et al., 2023); MtM, multi-task masking (Zhang et al., 2024); POYO, the cross-session spiking Transformer of that name (Azabou et al., 2023); UniBCI, the benchmark's own unified model; BCIM, the frozen backbone studied here. B-Acc, balanced accuracy. Baseline values are as listed by Hong et al. (2026), and the BCIM entries come from our own evaluations of the frozen model under each task's protocol.

BCIM has 7.7 million parameters. Its Perich-RT R^2 is 0.06 below the highest listed estimate. Frozen BCIM has the highest listed point estimate on MC-Maze (0.911) and Perich center-out (0.811). In the exploratory Gaussian sweep, each kernel smooths channels separately and the matched read-

Table 4: Gains that vary by task. Per-task decoding on the UniBCI benchmark datasets (Hong et al., 2026). We report R^2 for regression and balanced accuracy (B-Acc) for PPC-FINGER. Bold marks the highest listed point estimate in each row. Baseline values come from the implementations and checkpoints evaluated by Hong et al. (2026), so this is a descriptive listing without a statistical ranking. Their VAE uses the same self-supervised pretraining corpus as UniBCI; their NDT1 and NDT2 use MtM-provided weights pretrained on 34 sessions (abbreviations below). Figure 3 shows our matched comparison. The BCIM entry for Perich-RT reports the highest score across three output widths of the readout's per-token projection ($0.610/0.623/0.640$ at widths $4/8/16$).

Dataset	Metric	WF	GRU	MLP	NoMAD	VAE	NDT1	NDT2	MtM	POYO	UniBCI	BCIM
		downstream-trained				pretrained						frozen
Area2-Bump	R^2	0.724	0.814	0.778	0.793	0.797	**0.902**	0.888	0.874	0.882	0.890	0.892
MC-Maze	R^2	0.728	0.773	0.804	0.795	0.797	0.882	0.884	0.879	0.883	0.890	**0.911**
Perich-CO	R^2	0.580	0.625	0.682	0.656	0.716	0.737	0.738	0.682	0.753	0.757	**0.811**
Perich-RT	R^2	0.559	0.621	0.613	0.605	0.629	0.687	0.681	0.621	0.692	**0.702**	0.640
PPC-FINGER	B-Acc	0.852	0.894	0.929	0.717	0.881	0.959	0.961	0.934	0.940	**0.967**	0.959

out learns their combination. The selected kernel comes within 0.011 R^2 of frozen-feature decoding in a separate index-aligned analysis, not of the Table 4 entries. On Area2-Bump (Chowdhury et al., 2020), BCIM's listed value is 0.892, versus 0.890 for UniBCI and 0.902 for NDT1; on PPC-FINGER (Guan et al., 2022) it is 0.959, versus the highest listed value of 0.967. Its Perich-RT value is 0.640, below UniBCI's 0.702 and most listed fine-tuned Transformer baselines. Our matched raw-spike readout scores R^2 0.541 on Perich-RT with all fits retained (Table 7). MC-RTT, which is outside the UniBCI suite, is treated in Section 4.1.

BCIM is frozen, with no task fine-tuning. The table combines our BCIM evaluations with the baseline results reported by Hong et al. (2026), without paired uncertainty estimates. We choose BCIM because reconstruction is its sole pretraining objective and we control how features are extracted and aligned in time to the decoded outputs.

Backbone details. Writing $x_{j,i}$ for the patch at channel j and token-time index i, $\mathrm{tok}_{\mathrm{mod}}$ for the per-modality tokenizer, and pos_i, $\mathrm{emb}_{\mathrm{mod}}$ for the learned time and modality embeddings, the token embedding and encoder output are

$$\mathrm{emb}_{j,i} = \mathrm{tok}_{\mathrm{mod}}(x_{j,i}) + \mathrm{pos}_i + \mathrm{emb}_{\mathrm{mod}}, \qquad Z = \mathrm{LayerNorm}\big(\mathrm{Blk}_5 \circ \cdots \circ \mathrm{Blk}_1 \circ \mathrm{Blk}_0^{\mathrm{mod}}(\mathrm{emb})\big),$$

where $\mathrm{Blk}_0^{\mathrm{mod}}$ is the modality-specific first block and $\mathrm{Blk}_1, \ldots, \mathrm{Blk}_5$ are shared, for an encoder depth of six, and $d_{\mathrm{model}} = 256$ is the model width. Instantiating the model code with the training configuration ($d_{\mathrm{model}} = 256$, feed-forward width $1{,}024$, 8 heads, six blocks, 256 time positions, 50-sample field-potential patches and 5-bin spike patches) gives $7{,}666{,}432$ backbone parameters, of which $6{,}455{,}296$ lie on the spike path (the field-potential tokenizers and first block are not used by spikes); the pretraining decoders add $26{,}985$. The retained checkpoint was pretrained for 20 epochs with AdamW, with a $1{,}000$-hour cap applied to each source dataset, masking 15% of each sample's tokens, chosen uniformly at random over the channel-by-patch grid.

Downstream datasets. The three datasets cover four macaque reaching tasks. Neural counts are binned at 20 ms, and τ_{res}, τ_{vel} denote the within-trial $1/e$ autocorrelation-decay lags, in ms on a common 20 ms grid, of the shared population residual and of velocity. The shared population residual is the leading principal component of the standardized population activity after its velocity-predictable part has been regressed out on the training split. Both series are restricted to the behavior-aligned bins of each trial, 5-ms velocity samples are averaged into 20-ms bins, and the velocity timescale is computed on the velocity magnitude. MC-RTT (Neural Latents Benchmark, NLB; Makin et al., 2018) has 98 primary motor cortex (M1) units, random-target reaches, $\sim$ 600 ms trials, a 2-D velocity target, low stereotypy (little trajectory repetition), and $\tau_{\mathrm{res}} = 44 < \tau_{\mathrm{vel}} = 94$ ms. MC-Maze (NLB; Churchland et al., 2012) has 137 M1 and dorsal premotor cortex units, stereotyped center-out reaches with low-dimensional rotational dynamics and $\tau_{\mathrm{res}} = 59 < \tau_{\mathrm{vel}} = 102$ ms. Perich-CO/RT (DANDI:000688; Perich et al., 2018) is stored per recording session in Neurodata Without Borders format, with center-out (stereotyped) and random-target (continuous,

faster) reaches over six RT sessions. Both NLB datasets are accessed through the Neural Latents Benchmark (Pei et al., 2021).

Settings and constants. *Readout.* The frozen-versus-raw and Gaussian comparisons in this appendix use separately fitted instances of the same family of windowed MLPs, with two hidden layers of width 256 and weights shared across time. The physical window is matched within each pair. The UniBCI pipeline check uses about ± 120 ms (± 6 bins at 20 ms). The filter and residual analyses use a three-token half-window for all conditions. The cross-task comparison of §3.3 uses three frozen-feature tokens versus 15 raw 20-ms bins per side, approximately ± 300 ms in both conditions. Analyses use either benchmark-matched per-session 80/20 splits or 70/15/15 trial splits with explicit validation. Inputs are z-scored, and Perich R^2 values are averaged across sessions rather than pooled over trials (session means in Table 7).

Reconstruction and transfer probes. In the exploratory cross-session test, decoding without adaptation gives negative R^2. The inspected embeddings have cross-session representational-similarity correlation ≈ 0.45 (Spearman correlation between sessions' reach-direction dissimilarity matrices) and are dominated by time-in-trial structure. The frozen feature leaves the two tested variables, speed and time-in-trial, entangled.

Table 5: Decoding from residual features and reconstructed firing rates. All listed point estimates are below their matched raw-spike baselines. The residual is $F_{\mathrm{res}} := F - \widehat{F}(F_\sigma)$, where a ridge map predicts frozen features F from the selected single-σ Gaussian features F_σ; the reconstruction probe decodes velocity from the model's reconstructed firing rates. Within each row, the probe and raw spikes use the same readout family and temporal window. All reconstructed-rate fits are retained, and differences are computed before rounding. The MC-RTT residual row and its raw baseline of 0.479 use the cached MC-RTT store of §3.2 (688 trials, 98 units), eight $70/15/15$ trial splits with seeds 42 to 49, and the index-aligned per-token MLP head with three neighbors per side, $d_{\mathrm{tok}} = 8$ and hidden width 256, trained for up to 500 epochs with patience 80 (run log `a1_dense.log` in the released code). These probes are reported as point estimates; bootstrap intervals for the increment $\Delta_{F|F_\sigma}$ appear in §4.4 of the main text.

Probe	Dataset	probe R^2	matched raw R^2	difference
Residual alone	MC-RTT	0.353	0.479	−0.126
	Perich-RT	0.343	0.605	−0.262
Reconstructed rate	MC-RTT	0.382	0.416	−0.034
	Perich-RT	0.485	0.541	−0.057

Residual probe. Writing F for the frozen-model feature and F_σ for the selected single-σ feature, a ridge map $\widehat{F}$ is fitted on training trials to predict F from F_σ, and the residual is $F_{\mathrm{res}} := F - \widehat{F}(F_\sigma)$. With $[F_\sigma, F_{\mathrm{res}}]$ denoting feature concatenation and 8 seeds on MC-RTT and 3 on Perich-RT, the reported increment is $\Delta_{F|F_\sigma} := R^2([F_\sigma, F_{\mathrm{res}}]) - R^2(F_\sigma)$, decoded through the same matched windowed readout.

Seeds and intervals. The filter-sweep comparison (Figure 4) uses 3 seeds and the residual probe uses 8 seeds on MC-RTT; reported $\pm$ values are the standard deviation across seeds.

For MC-RTT and MC-Maze in Table 7, we report the mean and sample SD of the three paired seed-level frozen-minus-raw gaps; no nonparametric interval is formed from $n = 3$. For Perich-RT and Perich-CO, seed results are first averaged within session and the bootstrap resamples paired session-level differences. No time bins are resampled independently. The residual-alone and reconstructed-rate decodes are reported as point estimates; the increment $\Delta_{F|F_\sigma}$ carries the bootstrap intervals quoted in §4.4.

Filter bank. Gaussian smoothing is applied to raw counts without the variance-stabilizing Anscombe transform. The candidate grids were prespecified as $\{2, 4, 6, 8\}$ bins for MC-RTT and $\{1, 2, 4, 6\}$ bins for MC-Maze. The reported $\sigma^\star = 4$ and $\sigma^\star = 2$, respectively, maximize test-set R^2 in an exploratory seed-42 sweep and are then held fixed for the three-seed comparison; validation performance selects the checkpoint within each fit. The result is thus an exploratory best-observed value rather than an unbiased validation-selected estimate. The evaluated multi-σ bank does not outperform this value. In the retained Anscombe run, bank and FM decoding are 0.378 and 0.577

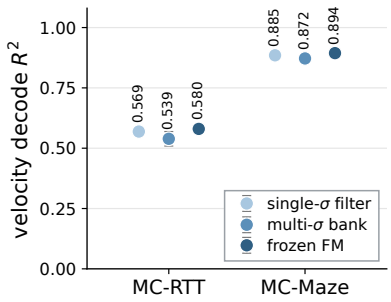

Figure 4: The best-observed member of the prespecified four-scale Gaussian sweep comes within 0.011 R^2 of the frozen model on MC-RTT and 0.009 on MC-Maze, and the multi-σ bank does not outperform it. The scale was selected on the test set and then held fixed, so this is a best-observed value and not an unbiased model-selection estimate. Error bars show SD across three readout seeds.

on MC-RTT (gap $+0.199$), and 0.840 and 0.893 on MC-Maze (gap $+0.053$). Relative to the separate non-Anscombe sweep plotted in Figure 4, the bank point estimate is lower by 0.161 and 0.032, respectively; this cross-run contrast is descriptive and is not used for causal attribution.

Neuron-dropping curve. The law is equation 10. Within each neuron-count level, the bootstrap resamples the recorded random neuron-subset repetitions, forms a mean curve, and refits the raw and FM neuron-dropping curves (NDCs) in every replicate. The resulting conditional interval for the ratio of half-saturation unit counts (Appendix F) is $\kappa_{\mathrm{FM}} \in [0.88, 2.10]$, with $\kappa_{\mathrm{bank}} \approx 1$. Because only three subset repetitions were recorded at each level, this is a descriptive sensitivity interval. It is conditional on the recorded MC-RTT dataset, neuron counts, readout, subset repetitions and NDC family.

Simulation. The two-dimensional Gaussian-process (GP)–Poisson model drives N Poisson units with velocity and injects information-limiting noise along the signal direction in two conditions. In the in-band condition the noise timescale equals the signal's, so the component is spectrally inseparable; in the fast condition the noise is much faster and therefore out-of-band. The large-population extension reaches $N = 2400$ and decodes both the raw and the smoothed condition through a 50-component principal component analysis (PCA) followed by ridge regression. At $N = 2400$ the smoothing-minus-raw gap is $+0.11$ (fast) and -0.013 (in-band).

Table 6: Paired frozen-minus-raw R^2 differences for the three readout seeds used on MC-RTT and MC-Maze. All six paired differences are positive. Because there are only three seed-level observations per task, no nonparametric bootstrap confidence interval is formed.

Task	seed 0	seed 1	seed 2	mean $\pm$ sample SD
MC-RTT	$+0.1074$	$+0.1482$	$+0.1070$	$+0.1209 \pm 0.0237$
MC-Maze	$+0.0024$	$+0.0189$	$+0.0024$	$+0.0079 \pm 0.0095$

Cross-task scores. Table 7 gives the numerical summaries plotted in Figure 3 and shows how the original exclusion rule changes the Perich-RT result. Table 6 lists the seed-level differences behind each MC entry of Table 7.

Perich-RT scores under different protocols. The published MLP trained from scratch on raw spikes scores 0.613 (Table 4). Our target-spectrum baseline scores 0.609, and the matched raw-spike readout scores 0.541 with all fits retained. The $+0.040$ in Section 4.4 is the frozen-minus-raw gap under the latter protocol. The benchmark value 0.640 is a frozen-feature decoding score obtained under a third, separately tuned protocol.

Raw-spike decoding under three readouts. Decoding scores from the same raw-spike recording depend on the readout, so each raw-spike baseline is specific to a dataset and protocol (Table 8).

Table 7: Mean performance in the cross-task comparisons (R^2). MC gaps are reported with the sample SD across three readout seeds. For Perich gaps, we compute 95% bootstrap intervals by resampling paired session differences after averaging seeds within each session. The primary estimates include every fit. The final row reports the original analysis, which retained scores with test $R^2 > -0.5$ separately for each representation before session averaging, removing one raw fit.

Task	FM	raw (matched)	FM−raw uncertainty
MC-RTT	0.537	0.416	$+0.121 \pm 0.024$ (3-seed SD)
MC-Maze	0.900	0.892	$+0.008 \pm 0.010$ (3-seed SD)
Perich-RT, all fits	0.581	0.541	$+0.040$ $[-0.055, +0.208]$ (session CI)
Perich-CO	0.710	0.770	-0.061 $[-0.093, -0.030]$ (session CI)
Sensitivity to the historical test-score exclusion			
Perich-RT, thresholded	0.581	0.617	-0.036 $[-0.057, -0.018]$ (session CI)

Table 8: Raw-spike velocity decoding (R^2) under three readouts. Column 1 is our matched readout, the frozen-versus-raw comparison of Table 7; column 2 is the UniBCI MLP of Table 4. The two differ in window, unit set and protocol. Our matched-readout scores include all fits, including on Perich-RT. Each frozen-versus-raw comparison uses the same readout family and physical window across conditions. A further residual-probe readout (§4.4) gives raw R^2 of 0.479 on MC-RTT and 0.605 on Perich-RT; the values 0.416, 0.479, and 0.680 are MC-RTT results obtained with different unit sets, trial construction or readout settings.

Dataset	Matched windowed MLP (Table 7)	UniBCI MLP (Table 4)	Wide $\pm$400 ms, 130 units
MC-RTT	0.416	–	0.680
MC-Maze	0.892	0.804	–
Perich-RT	0.541	0.613	–
Perich-CO	0.770	0.682	–

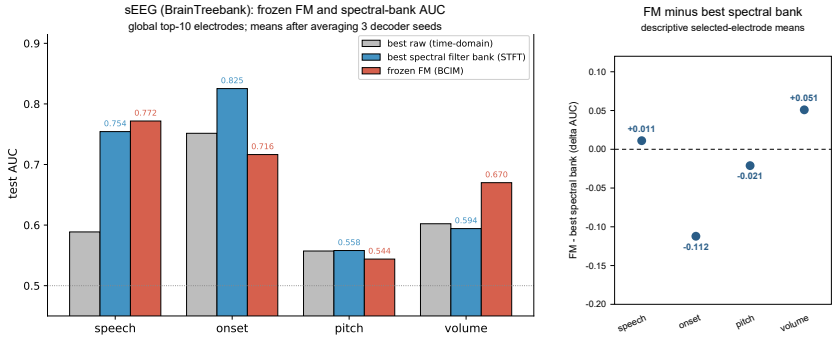

Figure 5: In a descriptive cross-modal check on BrainTreebank stereo-electroencephalography (sEEG) (Wang et al., 2024a), which was absent from pretraining, the FM-minus-spectral-bank differences in area under the receiver operating characteristic curve (AUC) are speech $+0.011$, onset -0.112, pitch -0.021, and volume $+0.051$, so the FM does not consistently outperform the spectral bank across the four tasks. The FM uses a two-layer readout. In the left panel, the bank value is the better of a linear and a two-layer readout on short-time Fourier transform (STFT) features, chosen per task on the test AUC. In the right panel, the better bank readout is chosen per electrode before averaging, so each plotted difference is at most the difference of the two left-panel means. Both choices favor the spectral bank. Points in the right panel show mean AUC differences across the selected electrodes after averaging three readout seeds. The comparison is conditional on the selected electrodes and matched readout, and no population-level significance claim is made.

Cross-modal sEEG check. Proposition 1 extends to continuous modalities in the limited sense that squared-loss masked reconstruction targets a conditional mean, which under a stationary linear-Gaussian approximation motivates a spectral-bank baseline. On BrainTreebank sEEG (Wang et al., 2024a), which was absent from pretraining, the frozen model does not consistently outperform that spectral bank (Figure 5). The bank's largest margin is on onset and the FM's on volume, with the values, readouts and electrode selection given in the figure caption. With electrodes concentrated in one subject per task, the check provides a descriptive comparison with a spectral baseline. It supports no cross-modal sign-change claim and does not test the link between the spatial cap and the temporal residual reduction of Appendix F.

Controls against artifacts. *The downstream window is matched across conditions.* The $+0.121$ MC-RTT gap uses half-widths of three FM tokens and 15 raw 20-ms bins (approximately ± 300 ms). A separate wider comparison uses ± 400 ms for both inputs. There, raw reaches 0.680 and the frozen model 0.654 on the continuous MC-RTT recording (DANDI:000129) with all 130 units. Because window, unit set, dataset construction and evaluation protocol all differ between those two comparisons, that contrast does not isolate window width. The dedicated sweeps of §4.1 vary it under three readout settings; the masked setting fixes parameter count, projection width and input layout within each representation. Through the backbone's full-trial attention, frozen features may already contain information from outside the readout window. This difference in available context must be considered when interpreting the decoding gap. *Cross-validation is trial-held-out.* Trials are assigned to disjoint splits, which avoids leakage through within-trial temporal autocorrelation, although on MC-RTT their extended neural input windows may nevertheless overlap in time (Appendix D). *Readout family is matched within each comparison.* A Perich-CO result obtained with a convolutional readout head is excluded because pairing it with the windowed-MLP conditions would confound readout family with representation.

Sensitivity to test-score exclusion. The Perich-RT comparison retains all 18 fits per representation, averaging three seeds within each of six sessions. Raw and frozen-feature means are 0.541 and 0.581. The frozen-minus-raw mean is $+0.040$, with a paired-session 95% interval of $[-0.055, +0.208]$. The original aggregation retained test scores strictly greater than -0.5 separately for each representation. It removed one raw fit (test $R^2 = -0.7734$, session RT-20130820, seed 2), raising the raw mean to 0.617 and changing the mean gap to -0.036. No independent diagnosis of a failed fit supports this exclusion, so we report it only as a sensitivity comparison (Table 7). With all fits retained, the mean gap is negative in five of six sessions, but the confidence interval for the across-session mean includes zero. The reconstructed-rate comparison uses the same corrected raw baseline (Table 5). The threshold removed no Perich-CO fits. The MC matched-readout analyses do not use this rule. Each matched-readout fit is a single run per seed or session with early stopping on validation performance and no restarts. The primary window analyses include all five replicates at each width, with restarts selected by validation performance. We recomputed the Perich summaries from scores logged to four decimal places, using the original 2,000-resample session bootstrap (seed 0). The scores and the checks of the aggregation code are documented with the released analysis code.

Reconstruction and raw spikes. The Pearson correlation between reconstructed rate and raw spikes is 0.378 on MC-RTT. Under the tested readouts, decoding from reconstructed rates yields lower mean scores than decoding from raw spikes (Table 5).

C Readout experiments in detail

For the primary protocol, see the reproduction details (p. 22), randomized-context control (p. 34) and Gaussian comparison (p. 28). The separate index-aligned Gaussian rerun and fixed-head analysis begin on pp. 28 and 26, respectively.

These analyses use the frozen features, raw patches, targets and valid-channel masks listed in §3.2. Available examples come only from the original downstream training split; each analysis constructs a trial-level $70/15/15$ train/validation/test partition within it, shared across conditions and paired seeds. The analysis set contains 688 MC-RTT trials over 98 units and 1463 MC-Maze trials over 137 units. In both cases a trial is represented by ten tokens per channel, each containing 5 raw 20-ms bins or 256 frozen-feature values. The targets contain 120 bins on MC-RTT and 140 on MC-Maze

at 5 ms. The neural window of 1,000 ms is longer than the scored behavior span of 600 or 700 ms because it includes forward bins. The output knots and the input tokens do not share a grid. The interpolation places knot k at behavior time approximately $(k + \frac{1}{2})\, T_b/10$ of the scored span T_b, and token k covers neural-window time $[100k,\ 100(k{+}1))$ ms. In the dataset construction the spike window and the scored span share their start. The trial-aligned spikes are followed by 200 ms of forward bins and summed from 5 to 20 ms bins with no shift, then zero-padded to 50 bins, so the live window covers the scored span plus 200 ms of forward context and the final 200 ms (MC-RTT) or 100 ms (MC-Maze) of the window are padding. Under index alignment on that common clock, at $K = 0$ the token behind output position k is centered $40k + 23$ ms (MC-RTT) or $30k + 18$ ms (MC-Maze) after the behavior instant it predicts, and the centers of tokens $k \geq 6$ (MC-RTT) or $k \geq 7$ (MC-Maze) fall beyond the scored span. A per-bin cross-trial regression on the cached MC-RTT store confirms this layout (best-correlated behavior time $\approx 18.6\, b + 28$ ms for neural bin b, matching the implied $20\, b$ plus a spike-to-velocity lag). The offset is the same for both conditions, but the two conditions are not equally exposed to it. Full-trial attention can route correctly timed content to every output position, and a narrow raw window cannot, so the offset can itself contribute to the measured gap at $K = 0$ and to its decline as K widens. Under index alignment, the parameter-count, position-code and randomized-context controls of §4.1 do not separate local temporal integration from compensation for the token–target offset. A constant parameter count leaves the physical window centers unchanged. A position code tells the readout which knot it predicts without supplying correctly timed input from outside its window. Randomizing the context destroys ordered neighbors, degrading temporal integration and offset compensation alike. The time-aligned + position protocol addresses the coarse offset by centering the readout window on the token nearest each output time. Token quantization and truncation at trial boundaries remain.

Temporal correspondence, output position and window width. Input tokens and output knots have different spacings, so index alignment and time alignment have different limitations. Index alignment creates the physical offsets above. Re-centering each output on the token nearest its behavior time removes them, with a prespecified lag of 0 ms and a quantization error of at most 43 ms (MC-RTT) or 48 ms (MC-Maze). With 100-ms tokens and 60–70-ms knot spacing it assigns the same center to adjacent knots (6 distinct centers of 10 on MC-RTT, 7 on MC-Maze), and a readout shared across positions with no position input then emits identical predictions for knots that share a center. Supplying both inputs the same one-hot code for the output-knot index, $U_k = e_k$, as in the position-code equalization control, removes this forced equality without adding neural observations. Table 9 reports the three protocols at the narrowest and widest windows under the layout with fixed parameter count, with everything else held fixed apart from the ten position-code inputs of the third protocol, and Figure 6 plots them.

Under time alignment with the shared output-position code, the mean $K = 0$ gap is $+0.181$ on MC-RTT and $+0.162$ on MC-Maze, about 0.6 of its index-aligned value. The gap is positive in all five replicates per dataset and falls to means of $+0.019$ and $+0.052$ at $K = 4$. It declines in all five replicates on both datasets. Across the intermediate widths the mean gap is $+0.037$, $+0.029$ and $+0.034$ on MC-RTT and $+0.082$, $+0.070$ and $+0.060$ on MC-Maze, so most of the MC-RTT reduction occurs at $K = 1$ and the MC-Maze reduction continues across the sweep (Figure 1). On MC-RTT, raw decoding rises from 0.425 at $K = 0$ to 0.578 at $K = 1$ and then stays between 0.616 and 0.625. On MC-Maze, it rises at every step, from 0.728 to 0.843. Frozen-feature decoding ranges from 0.606 to 0.658 on MC-RTT and from 0.890 to 0.896 on MC-Maze.

The two tasks differ in sensitivity to alignment and output-position information. On MC-RTT time alignment alone raises raw decoding by 0.224 and lowers the gap by 0.123 at $K = 0$ (5/5 each), and the position code then changes the gap by $+0.010$. On MC-Maze, alignment alone lowers both decoding scores and reduces the gap by 0.035. The position code then raises raw decoding by 0.116 and lowers the gap by a further 0.059 (5/5 each). These are paired effects of the implemented protocol changes, not an additive decomposition of the backbone's computation, and the recovery of the frozen condition's score under the position code does not identify the source of its earlier loss. At $K = 4$, the ± 400 ms window already covers the offset. On MC-Maze, raw decoding under index alignment is 0.024 higher than under the time-aligned + position protocol, with frozen-feature decoding nearly unchanged. Index alignment assigns a distinct central token to each knot, whereas time alignment can assign the same token to adjacent knots. We have not isolated how this repetition affects the final gap. The $+0.052$ that remains on MC-Maze under the time-aligned protocol with the position code is stable across replicates (0.048–0.058, against -0.014 to $+0.059$ on MC-RTT)

and about twice the index-aligned value. The time-aligned + position protocol removes the coarse index offset and the forced equality of outputs. It retains token quantization, window truncation at trial boundaries and the difference in temporal support between the backbone and the readout.

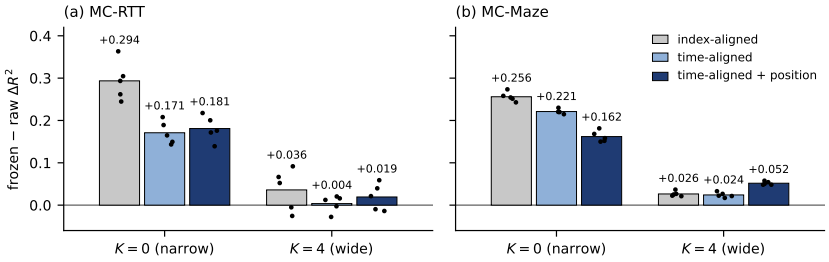

Figure 6: Frozen-minus-raw gap at the narrowest and widest windows under the three window protocols of §3.2, on (a) MC-RTT and (b) MC-Maze. The time-aligned + position bars are the primary protocol, whose full sweep is Figure 1. Bars and labels give means over five paired replicates. Dots show individual replicates. Within a dataset the three bars at one K share splits and seeds, so their differences are paired effects, not an additive decomposition.

Table 9: Frozen-minus-raw gap under three window protocols, at the narrowest and widest windows for the first two and at every half-width for the primary protocol. All use the masked layout at $d_{\text{tok}} = 8$, with parameter count fixed across windows within each representation and five paired split-and-seed replicates following the restart rule. *Index-aligned* centers the window of output knot k on token k; *time-aligned* centers it on the token nearest the knot's behavior time; *time-aligned + position* additionally supplies both inputs the same one-hot output-knot index. The "live slots" column gives the mean number of non-zero window slots per output position and differs between protocols, not between the two conditions within a protocol. The twelve narrowest- and widest-window conditions were run in one batch on one machine, and the primary protocol's $K = 1$–3 conditions in a later batch on the same machine with the same seeds; all conditions share splits and seeds. The index-aligned rows are a separate run batch from Table 13. Differences are unrounded, and the reductions quoted in the text use the unrounded gaps.

Dataset	Protocol	K	raw	frozen	frozen−raw (pos.)	live slots
MC-RTT	index-aligned	0	0.200	0.494	+0.294 (5/5)	1.00
		4	0.543	0.579	+0.036 (3/5)	7.00
	time-aligned	0	0.424	0.595	+0.171 (5/5)	1.00
		4	0.631	0.634	+0.004 (3/5)	7.30
	time-aligned + position	0	0.425	0.606	+0.181 (5/5)	1.00
		1	0.578	0.615	+0.037 (5/5)	2.80
		2	0.616	0.645	+0.029 (4/5)	4.50
		3	0.625	0.658	+0.034 (5/5)	6.00
		4	0.622	0.642	+0.019 (3/5)	7.30
MC-Maze	index-aligned	0	0.629	0.885	+0.256 (5/5)	1.00
		4	0.867	0.893	+0.026 (5/5)	7.00
	time-aligned	0	0.612	0.833	+0.221 (5/5)	1.00
		4	0.820	0.844	+0.024 (5/5)	7.50
	time-aligned + position	0	0.728	0.890	+0.162 (5/5)	1.00
		1	0.811	0.893	+0.082 (5/5)	2.90
		2	0.825	0.895	+0.070 (5/5)	4.60
		3	0.835	0.896	+0.060 (5/5)	6.20
		4	0.843	0.895	+0.052 (5/5)	7.50

Reproducing the primary comparison. The comparisons reported in Figure 1 and the time-aligned + position rows of Table 9 use the same frozen BCIM backbone. For each dataset, we cache the frozen features (trials × channels ×10 × 256), raw patches (trials × channels ×10 × 5), 5-ms velocity targets and valid-channel mask from the original downstream training split. For repli-

Table 10: Protocols for the readout experiments. Each analysis uses paired splits and training seeds on the cached data, selecting the best of three fits on validation data for each readout-fitting condition and replicate (the fixed-head substitution selects ridge strength by validation prediction error instead). The statistical unit is the replicate (Reps), and absolute scores are not comparable across rows.

Analysis	Token–output mapping	Position code	K	Setting (d_{tok})	Reps
Primary sweep (Figure 1, Table 9)	time-aligned + position	both	0–4	masked, 8	5
Randomized context, primary (Table 1)	time-aligned + position, randomized	both	4	masked, 8	5
Window sweeps (Table 12)	index-aligned	none	0–4	fixed 8 or approximate	5
Constant-parameter sweep (Table 13)	index-aligned	none	0–4	masked, 2–32	5
Position-code equalization (Table 17)	index-aligned	both vs. zeros	0–4	fixed 8	12
Randomized context, index-aligned (Table 20)	index-aligned, randomized	none	1–4	fixed 8 or approximate	5
Narrowest and widest windows (Table 9)	all three protocols	none/one-hot	0, 4	masked, 8	5
Gaussian, primary (Table 16)	time-aligned + position	both	0, 4	masked, 8	5
Direct per-bin output (p. 25)	time-aligned + per-bin code	both	0, 4	masked, 8	5
Gaussian, validation rerun (p. 28)	index-aligned	none	3	fixed 8	15
Training-set size (Table 21)	index-aligned	none	0, 3	fixed 8	5
Fixed-head substitution (Tables 14 and 15)	index-aligned	none	0	fixed 8	12

cate $r \in \{0, \ldots, 4\}$, we permute trials with seed $42 + r$ and assign the first 70%, next 15% and last 15% to training, validation and test, respectively.

Knot k is placed at behavior time $((k + \frac{1}{2})\, n_b/10 - \frac{1}{2}) \times 5$ ms, where n_b is the number of 5-ms target bins (120 or 140). Token j is centered at neural-window time $100j + 50$ ms. After expressing both times on the same clock, we set $c(k)$ to the index of the token nearest to knot k, with $\tau_{\mathrm{lag}} = 0$. The readout projects each channel-token to $d_{\mathrm{tok}} = 8$ and places the projected tokens in nine slots centered on $c(k)$. It zeros slots more than K token steps from $c(k)$ or beyond the trial boundaries, and appends the one-hot code U_k. A two-hidden-layer MLP of width 256 with dropout 0.2 predicts velocity. We train the readout with AdamW (learning rate 10^{-3}, weight decay 10^{-2}, batch size 128) for up to 500 epochs, stopping after 80 epochs without improvement in validation R^2. Restart $q \in \{1, 2, 3\}$ uses initialization seed $1000q + 42 + r$, and the restart with the highest validation R^2 supplies the reported test score for that replicate.

Repeated single-seed fits reproduce the reported values to within about 0.015 in R^2, and the raw condition's across-replicate SD is about 0.05. Comparisons use paired splits and initializations. The repository provides checkpoint identifiers, scripts and per-replicate outputs.

In the readout experiments, rerunning the published filter comparison recovers it within seed noise (Gaussian 0.554 against the published 0.569, frozen model 0.570 against 0.580, raw 0.455; single seed). These reruns verify reproducibility of the cached-feature pipeline, not physical alignment between input tokens and output knots. The effects of temporal mapping and output-position coding are evaluated separately above. The untrained control uses BCIM's architecture, tokenizer and feature-extraction procedure; only checkpoint loading is skipped.

Readout output bandwidth. The interpolation of the readout's output could itself suppress the high target bands, so its response is recorded here. The readout emits $T = 10$ velocity knots per trial and interpolates linearly to the full target length, a first-order hold. Its -3 dB point lies well below

the knot-grid Nyquist frequency, and its gain is still 0.77–0.82 at 4 Hz (Table 11). The knot update rate, the knot-grid Nyquist frequency and the -3 dB point of the first-order hold differ.

Table 11: Output bandwidth of the knot-and-interpolate readout. Most power in the original velocity targets lies below 4 Hz, and the interpolation gain at 4 Hz is 0.77–0.82. Interpolation attenuates high-frequency output components, limiting sensitivity in the higher target bands (§4.4).

	MC-RTT	MC-Maze
Knot update rate	16.67 Hz	14.29 Hz
Knot-grid Nyquist frequency	8.33 Hz	7.14 Hz
Analytic interpolation gain at Nyquist	0.405	0.405
-3 dB point of the first-order hold	5.32 Hz	4.56 Hz
Cumulative velocity power below the -3 dB point	99.6%	98.2%
Interpolation gain at 4 Hz	0.82	0.77
Cumulative velocity power below 4 Hz	97.9%	96.5%

Window sweeps with fixed or varying projection width. Table 12 gives the absolute per-condition values for the two ways of setting projection width as the window widens (§4.1); the masked sweep in Figure 7a is reported in Table 13. Each neural token contains five 20-ms bins, so one step in K adds 100 ms on each side. The 16.67/14.29 Hz knot rates of Table 11, that is 60/70 ms knot spacings, are output quantities and not neural-token spacings. Approximate parameter matching sets $d_{\mathrm{tok}}(2K+1) \approx 8$, holding frozen-condition head parameters within 0.24–0.32 M on MC-RTT and 0.31–0.42 M on MC-Maze. At fixed $d_{\mathrm{tok}} = 8$, they grow from 0.27 M to 1.87 M and from 0.35 M to 2.59 M. Raw and Gaussian have the same architecture but a five-dimensional input instead of the frozen condition's 256-dimensional input, so only their small per-token input-projection parameter count differs. Frozen decoding increases less than raw under both settings (fixed width, MC-RTT: 0.489 to 0.587, raw 0.193 to 0.547). It is nearly flat on MC-Maze under both settings and rises only 0.058 on MC-RTT under approximate matching.

Table 12: Velocity decoding (R^2, mean over five paired split-and-seed replicates) against readout half-width K, for raw spike counts, the fixed Gaussian and the frozen model. Within each replicate, validation R^2 selects the best of three random restarts. The frozen-minus-raw gap falls from $+0.292$ to -0.026 on MC-RTT and from $+0.252$ to $+0.005$ on MC-Maze between $K = 0$ and $K = 4$ under approximate parameter matching, and by 86% and 90% at fixed projection width. The frozen-minus-Gaussian mean turns negative at $K = 3$ on both datasets under approximate matching; at fixed width it is negative by at most 0.02 at $K = 3$–4 on MC-RTT and stays nonnegative on MC-Maze at the displayed precision. Figure 7a plots the masked setting of Table 13; because the Gaussian ordering differs by setting, it is secondary to the frozen-minus-raw decline.

		MC-RTT			MC-Maze		
K	d_{tok}	raw	Gaussian	frozen	raw	Gaussian	frozen
approximately parameter-matched							
0	8	0.192	0.335	0.484	0.633	0.805	0.885
1	3	0.283	0.411	0.495	0.786	0.860	0.885
2	2	0.366	0.507	0.514	0.838	0.882	0.887
3	1	0.514	0.579	0.541	0.877	0.893	0.882
4	1	0.568	0.598	0.542	0.882	0.897	0.887
fixed projection width							
0	8	0.193	0.335	0.489	0.633	0.805	0.885
1	8	0.298	0.419	0.523	0.769	0.857	0.891
2	8	0.402	0.505	0.552	0.824	0.878	0.892
3	8	0.504	0.578	0.572	0.853	0.886	0.894
4	8	0.547	0.607	0.587	0.867	0.891	0.891

Window width at fixed parameter count. The two settings above either reduce d_{tok} or increase head size as K grows, so neither isolates window width. We instead keep $2K_{\max}+1 = 9$ input slots and vary the window by *zeroing* slots outside $\pm K$. Parameter count, projection width and input

layout stay fixed across the five window widths within each representation. Table 13 reports this masked setting at $d_{\text{tok}} = 8$. The frozen-minus-raw decline reproduces at 86.9% on MC-RTT and 89.5% on MC-Maze, against 86.3% and 90.4% under the fixed-width setting.

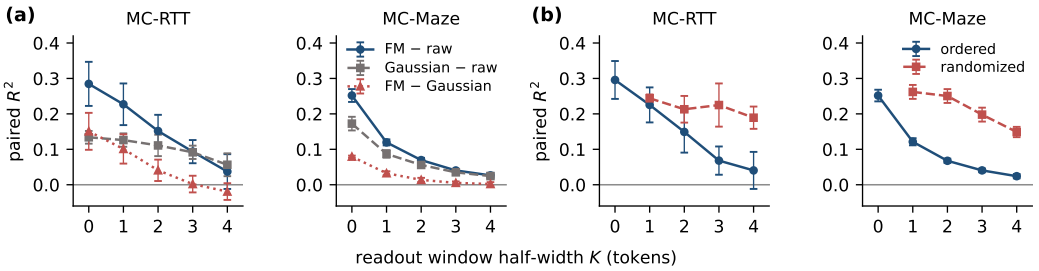

Figure 7: Window interventions at $d_{\text{tok}} = 8$. (a) Masked setting with parameter count exactly constant, the nine-slot head varying K only by zeroing slots outside $\pm K$. Curves show paired differences (mean and sample SD over five split-and-seed replicates, each validation-selected over three restarts). (b) Fixed-width setting with ordered or randomized within-trial context. Curves show frozen-minus-raw differences (mean and sample SD over five paired replicates). Randomized windows preserve slot count and edge padding, and $K = 0$ has no randomized counterpart.

The reduction is not monotone in the projection width on MC-RTT, at $75.4/86.9/90.5/80.0\%$ for $d_{\text{tok}} = 2/8/16/32$, against $86.7/89.5/92.6\%$ at $2/8/16$ on MC-Maze, so each percentage should be quoted with its projection width. $\Delta_{\text{FM}-\text{raw}}$ is smaller at $K = 4$ than at $K = 0$ in $5/5$ replicates in all seven projection-width blocks, four on MC-RTT and three on MC-Maze, so the direction is stable across widths even where the percentage is not.

Under the primary protocol at $d_{\text{tok}} = 32$, the MC-RTT gap is $+0.189$ at $K = 0$ and $+0.017$ at $K = 4$ (91% reduction), against $+0.181$ and $+0.019$ at $d_{\text{tok}} = 8$. On MC-Maze it is $+0.210$ at $K = 0$ and $+0.061$ at $K = 4$ (71%), against $+0.162$ and $+0.052$. The decline holds in $5/5$ replicates on both datasets. The wider projection leaves the frozen scores within 0.022 of their $d_{\text{tok}} = 8$ values and lowers the MC-Maze raw scores (0.728 to 0.677 at $K = 0$, 0.843 to 0.830 at $K = 4$). At $K = 4$, the paired-t intervals for MC-RTT include zero at both $d_{\text{tok}} = 32$ ($[-0.028, +0.061]$) and $d_{\text{tok}} = 8$ ($[-0.020, +0.058]$). The corresponding MC-Maze interval at $d_{\text{tok}} = 32$ is $[+0.055, +0.067]$. The gap reduction therefore persists at both projection widths under the primary protocol.

Direct per-bin output. To check whether the window effect persists without knot interpolation, we refit the raw-spike and frozen-feature readouts to predict velocity directly at each 5-ms behavior bin (120 per trial on MC-RTT and 140 on MC-Maze). Each readout window is centered on the token nearest to the output time. Both readouts receive a one-hot code identifying that output bin in place of U_k. The masked nine-slot layout, $d_{\text{tok}} = 8$, the restart rule and the seeds are unchanged. The higher-dimensional position code adds $28{,}160$ parameters to each head on MC-RTT and $33{,}280$ on MC-Maze. MC-RTT uses the chain-grouped split and MC-Maze uses the random split of Table 9; each comparison uses the ten-knot results from the same split.

On MC-RTT, widening from $K = 0$ to $K = 4$ raises raw decoding from 0.406 to 0.619 and frozen-feature decoding from 0.567 to 0.602. The mean frozen-minus-raw gap falls from $+0.161$ to -0.017, compared with a decline from $+0.176$ to -0.009 with ten knots. The gap is positive in $5/5$ replicates at $K = 0$ and $1/5$ at $K = 4$, where its paired-t interval is $[-0.065, +0.031]$. On MC-Maze, widening raises raw decoding from 0.714 to 0.842 and frozen-feature decoding from 0.893 to 0.896. The gap falls from $+0.179$ to $+0.053$, compared with a decline from $+0.162$ to $+0.052$ with ten knots. It is positive in $5/5$ replicates at both window sizes, with a paired-t interval of $[+0.046, +0.061]$ at $K = 4$. The gap declines in every replicate on both datasets, and the absolute scores differ from the ten-knot values by at most 0.032.

Because direct prediction also changes the number of output positions and the dimension of the position code, this check does not isolate the effect of interpolation. It covers two window sizes on the same two datasets; the script and per-replicate outputs are in the released code.

Rerunning all four sweeps of Table 12 from the same cached store on different hardware and library versions reproduces the $K = 0$ to $K = 4$ decline in $5/5$ replicates in every case, with per-K gaps within 0.016 of the reported values on MC-RTT and within 0.004 on MC-Maze. At fixed width the reductions are 83.2% against the reported 86.3% on MC-RTT and 89.9% against 90.4% on MC-Maze, and under approximate matching the $K = 4$ gaps are -0.014 against -0.026 and $+0.007$ against $+0.005$. The gap reduction persists across these reruns.

Table 13: Window sweep with masked inputs, keeping parameter count fixed across window widths within each representation (R^2, mean over five paired split-and-seed replicates, each the best of three restarts selected on validation R^2; $d_{\mathrm{tok}} = 8$). The Gaussian condition uses one width per dataset ($\sigma = 4$ on MC-RTT, $\sigma = 2$ on MC-Maze), selected on test in the earlier exploratory sweep and held fixed before this sweep. The decline seen in Table 12 persists when parameter count, projection width and input layout remain fixed across windows. The gap row is computed before rounding.

	MC-RTT			MC-Maze		
K	raw	Gaussian	frozen	raw	Gaussian	frozen
0	0.201	0.335	0.486	0.632	0.805	0.884
1	0.296	0.422	0.523	0.771	0.858	0.891
2	0.402	0.513	0.553	0.823	0.879	0.892
3	0.485	0.576	0.578	0.854	0.889	0.894
4	0.554	0.610	0.591	0.867	0.892	0.894
$\Delta_{\mathrm{FM-raw}}$	$+0.285 \to +0.037$ (86.9%)			$+0.252 \to +0.026$ (89.5%)		

Linear predictions in a fixed head. The readout uses its per-token projection Π to map each 256-dimensional channel-token to $Q = \Pi Z \in \mathbb{R}^8$. At $K = 0$, it concatenates these vectors across channels and passes the resulting vector to the head's first linear layer. The rest of the fixed head receives the first-layer output before activation, $\mathrm{pre}(Q)$. The feature prediction $\widehat{Q}$ from raw spikes and trial position is therefore scored with $\mathrm{pre}(\widehat{Q})$; reapplying the projection Π before the head's first layer would duplicate it. Ridge strength minimizes mean-squared feature-prediction error on validation trials, without using behavioral targets. The decoding column reports performance after feature substitution without retraining the head, which was previously trained on behavioral targets.

We save the head weights at the epoch with the best validation score. At run time we verify the decomposition by checking that the remainder of the head, applied to $\mathrm{pre}(Q)$, reproduces that head's own test R^2; the check held to 10^{-8} in every replicate. M_{1ch} constrains each output channel to depend only on the same input channel, which is a block correspondence rather than a column subset, so it is fitted as one independent ridge per channel. M_{1ch} is a channelwise, lag-shared, linear family of the same structural type as Gaussian smoothing. Because every linear model includes a position-dependent template term while the index-aligned substitution head receives no output-position input, this experiment tests feature substitution into a fixed head and does not show that the substituted mapping belongs to the raw-readout family of §4.1. It is not a superset of the swept Gaussian conditions, since its support of two tokens on each side does not cover the widest Gaussian kernels. Repeating edge values into the kernel support makes a wide Gaussian filter depend on position within a ten-token trial, so it is not strictly shift-invariant.

The four reported columns differ in scope. Explained variance on pre weights error directions by the Gram matrix of the head's first linear layer. Explained variance on Q is measured in the basis learned by the readout's own downstream projection Π. Neither quantity is invariant to an invertible reparameterization that leaves the network's function unchanged, and Π is fitted separately for each backbone. Both quantities are intermediate between the feature and the output, and because neither is invariant they support comparisons within a condition but not across conditions. Decoding through the fixed head is our primary measure and depends only on its input–output function. On MC-RTT, M_{1ch} explains more variance in Q, whereas M_1 explains more variance in $\mathrm{pre}(Q)$ and gives higher decoding R^2, so unweighted explained variance in Q and decoding performance rank these models differently. On MC-Maze, the two have nearly equal explained variance in Q, and M_1 again gives higher decoding R^2. These orderings depend on the tested linear classes, window, lag support and recordings.

Table 14: Linear predictions in a fixed head, evaluated over twelve paired split-and-seed replicates at $K = 0$. M_1 and $M_{1\mathrm{ch}}$ use token lags $-2, -1, 0, 1, 2$. EV columns are held-out explained variance on the projected feature, on the head-visible preactivation, and on the fixed head's own output; the last column reports decoding after substituting predictions without retraining the head. M_2 is a position-dependent full-trial linear model, an exploratory higher-capacity control described below. Frozen-feature scores come from the best of three readout restarts, selected on validation (details below). They are not directly comparable with Gaussian scores from other protocols (§4.2).

	MC-RTT				MC-Maze			
model	EV(Q)	EV(pre)	EV(out)	decoding	EV(Q)	EV(pre)	EV(out)	decoding
frozen feature	–	–	–	0.491	–	–	–	0.883
M_0	−0.013	−0.014	−0.028	−0.010	−0.006	−0.006	−0.038	−0.014
$M_{1\mathrm{ch}}$	0.323	0.310	0.383	0.201	0.445	0.415	0.796	0.720
M_1	0.282	0.442	0.547	0.355	0.454	0.618	0.921	0.852
M_2	0.122	0.328	0.321	0.250	0.183	0.465	0.861	0.818

Table 15: Paired decoding increments between the linear models of Table 14, over the same twelve replicates. The first three rows report the mean increment in fixed-head decoding R^2, its 95% percentile-bootstrap interval and the number of replicates in which the increment is positive. The last three rows give the relative loss in fixed-head decoding when each model is refitted on half of the training set.

	MC-RTT	MC-Maze
$M_1 - M_0$	+0.365 [+0.344, +0.385], 12/12	+0.866 [+0.851, +0.882], 12/12
$M_1 - M_{1\mathrm{ch}}$	+0.154 [+0.132, +0.171], 12/12	+0.132 [+0.125, +0.138], 12/12
$M_2 - M_1$	−0.105 [−0.124, −0.081], 0/12	−0.034 [−0.046, −0.018], 2/12
decoding loss when the training set is halved		
$M_{1\mathrm{ch}}$	1.1%	0.5%
M_1	11.4%	1.7%
M_2	47.2%	−1.4%

Cross-channel advantage on untrained backbones. We test the same linear model classes on four untrained initializations of the same architecture, each with its own downstream projection and head. Cross-channel predictions again yield better decoding than channelwise predictions, but explained variance differs across backbones. Across the four untrained backbones, M_1 explains 0.884–0.905 of the variance of Q on MC-RTT and 0.938–0.950 on MC-Maze, against pretrained values of 0.282 and 0.454. The measured object here is $Q = \Pi Z$, with Π and the head trained downstream for each backbone separately, so it describes the task-conditioned projection only.

Linear prediction before and after the readout projection. Q is expressed in a task-selected basis Π that differs per backbone, so the pretrained-versus-untrained ordering at Q may be a property of Π. The M_1 fit was therefore repeated against the backbone output Z in place of the projected feature Q, with the same split, the same two-token lag support at $K = 0$, and the same ridge grid selected by validation prediction error, over three replicates. On MC-RTT the pretrained backbone gives 0.420 ± 0.008 against 0.247–0.262 over the four untrained initializations; on MC-Maze, 0.506 ± 0.006 against 0.577–0.611. Relative to the projected feature, untrained levels fall from 0.88–0.95 to 0.25–0.61, and the pretrained-versus-untrained ordering reverses between datasets.

Each measurement point carries a distinct confound. A Π fitted on an untrained backbone can select a usable subspace whose linear predictability from raw spikes differs from that of the full backbone output (0.88–0.95 against 0.25–0.61 here). That difference makes EV(Q) unsuitable for cross-backbone comparison. The fit at Z regresses 256 feature dimensions onto 98 (MC-RTT) or 137 (MC-Maze) channels, so ridge conditioning, effective sample size and input statistics all vary by dataset. The linear predictability of an untrained transformer's output also depends on how far its nonlinearities are driven and on the scale after the closing LayerNorm. It therefore need not reflect a representational property. Consistent with this, the untrained conditions span 0.25 to 0.61 across the two datasets, a spread exceeding the pretrained-minus-untrained difference on either. Explained

variance therefore depends on where the features are measured, and we make no cross-backbone comparison of them in either direction. The comparisons supporting §4.3 are made within one condition, one head and one fixed Π, where the confound does not arise.

The expressive restriction behind the channelwise deficit is architectural, since each input channel is confined to the column space of its own projected token block. Its empirical cost is already visible without pretraining. $M_1 - M_{1\mathrm{ch}}$ is positive in $48/48$ combinations of replicate and initialization on the untrained backbones. Across the four initializations, the mean increment ranges from $+0.149$ to $+0.170$ on MC-RTT and from $+0.230$ to $+0.246$ on MC-Maze.

Rerunning one initialization end to end moves these quantities by 0.001 to 0.016 across the twelve comparable summaries, four of them by more than 0.008, since every replicate retrains its own head using a nondeterministic fitting procedure. The third decimal is therefore not a reproducible digit and the second moves in some summaries, so ranges are quoted instead of point values.

The M_2 ordering also flips. On the untrained backbone M_2 exceeds M_1 through the head (0.313 against 0.307 on MC-RTT, 0.844 against 0.789 on MC-Maze), the opposite of the pretrained case. M_2 was fitted on one of the four initializations only, so that ordering, unlike the $M_1 - M_{1\mathrm{ch}}$ comparison above, has been checked for only one backbone initialization.

M_2 is an exploratory higher-capacity control. Its comparison with M_1 jointly changes temporal support, position dependence and estimation variance, since it relaxes a two-token lag-shared kernel to a position-dependent full-trial one, so it isolates none of the three. Separating support from position dependence would need the intermediate lag-shared full-support class, which we did not fit, since the main claim rests on M_0, $M_{1\mathrm{ch}}$ and M_1 alone.

The halved-training-set rows of Table 15 show that $M_2 - M_1 < 0$ does not establish an expressivity difference. On MC-RTT the loss is monotone in parameter count and M_2 loses nearly half its score, so its deficit there is consistent with estimation failure. On MC-Maze, with roughly twice the trials, M_2 scores 1.4% higher with half the training set but remains below M_1. These two training sizes do not establish whether additional data would close the deficit. Ridge-selection variability is one possible explanation for the small increase on halving. Reproducing the head's output is also not the same as matching its decode, since on MC-RTT M_2 has lower EV(out) than $M_{1\mathrm{ch}}$ but decodes better.

Selecting the Gaussian scale on validation. The $\sigma^\star$ in the earlier exploratory comparison (Figure 4) was selected on test, introducing an optimistic bias into that comparison. We rerun it with σ chosen on validation, nested inside the restart rule, over fifteen paired split-and-seed replicates on both datasets; that rerun quantifies the bias. It uses the restart rule, whereas the Gaussian/frozen values behind Figure 4, $0.569/0.580$ on MC-RTT and $0.885/0.894$ on MC-Maze, use the three-readout-seed protocol, so the absolute values are not comparable. In the rerun the frozen model scores 0.566 on MC-RTT and exceeds the multi-σ bank by $+0.03$ there and $+0.02$ on MC-Maze. These index-aligned values now support the primary-protocol comparison reported below.

Test-selected minus validation-selected Gaussian is $+0.0027$ (CI $[+0.0003, +0.0060]$) on MC-RTT and $+0.0005$ (CI $[+0.0002, +0.0010]$) on MC-Maze, under half the published difference on the first and a twentieth of it on the second. Under validation selection the two datasets give different orderings. MC-Maze keeps a frozen-model advantage of $+0.0068$ (CI $[+0.0043, +0.0093]$, positive in $14/15$), which excludes zero, so the Gaussian comes within about 0.007 rather than matching. MC-RTT gives -0.0032 (CI $[-0.0196, +0.0137]$, positive in $7/15$), a null supporting no ordering in either direction. Holding σ at its published value instead of reselecting gives $+0.0025$ (CI $[-0.0139, +0.0194]$), so even the sign there depends on the selection variant. The validation-chosen scale is $\sigma = 2$ in $10/15$ MC-RTT replicates and $\sigma = 2$ in $12/15$ MC-Maze replicates.

A null difference test cannot establish that the Gaussian matches the frozen model; that conclusion requires an equivalence bound. Both intervals fall inside the $\delta = 0.02$ margin of practical equivalence fixed before this rerun. Because the margin was set after the published standard deviations were known, the paired difference and its interval remain primary.

Gaussian comparison under the primary protocol. We repeat validation selection under the time-aligned + position protocol at $K = 0$ and $K = 4$ (§3.2, Table 16). The readout uses the masked nine-slot layout at $d_{\mathrm{tok}} = 8$, with five paired split-and-seed replicates and the best of three restarts

selected on validation data. Each replicate selects σ from $\{2, 4, 6, 8\}$ bins on MC-RTT or $\{1, 2, 4, 6\}$ on MC-Maze. The bank concatenates raw spikes with inputs smoothed at widths $\{2, 4, 8, 12\}$ bins. The bank's wider input is read out with the same MLP width and regularization, which were not tuned for it. We smooth each channel's cached trial window at 20-ms resolution before assembling readout windows. The Gaussian kernel is truncated at a radius of 4σ bins. Beyond each boundary of the cached window, we repeat the edge value. The filter uses no samples from other trials. Smoothing includes the cached window's zero-padded tail, so padded zeros contribute to filtered values near the end of the recorded segment. The bank uses the same boundary rule.

The chosen width is $\sigma = 4$ bins (80 ms) in 16 of 20 replicate-by-window cases; this width has a 320-ms kernel radius. Three MC-RTT replicates at $K = 4$ select $\sigma = 2$, and one MC-Maze replicate at $K = 0$ selects $\sigma = 6$. At $K = 0$, the selected single scale recovers all but about 0.005 of the mean frozen-feature gain on both datasets. At $K = 4$, smoothed spikes outperform frozen features on MC-RTT in all five replicates, with a mean advantage of 0.050. Paired-t intervals over the five replicate differences (mean $\pm\, 2.776\,\mathrm{SD}/\sqrt{5}$, paired SDs 0.010, 0.021, 0.005 and 0.006) are $[-0.008, +0.018]$ for MC-RTT at $K = 0$, $[-0.076, -0.024]$ for MC-RTT at $K = 4$, $[-0.001, +0.011]$ for MC-Maze at $K = 0$ and $[-0.003, +0.012]$ for MC-Maze at $K = 4$. For MC-RTT at $K = 4$ both intervals lie below zero, whereas for the two MC-Maze rows the paired-t intervals include zero. All intervals describe split and optimization variation within one recording. On MC-Maze, frozen features retain a mean advantage of 0.005, with a positive difference in four of five replicates. The bank has a lower mean score than the selected single scale in all four dataset-by-window conditions. On MC-RTT, mean Gaussian decoding rises from 0.601 at $K = 0$ to 0.691 at $K = 4$, so smoothing does not eliminate the benefit of a wider readout window at these endpoints.

Raw and frozen readouts are refitted in this batch using the splits and seeds of the primary sweep (Table 9), and we do not reuse its fitted readouts. MPS operations are not bit-deterministic, so the rerun scores need not be identical. For MC-RTT at $K = 4$, mean scores are 0.625 for raw and 0.641 for frozen features in this batch, compared with 0.622 and 0.642 in the full sweep.

Table 16: Gaussian smoothing under the primary protocol. Entries are mean test R^2 over five paired replicates, with σ chosen on validation data in each replicate. Differences are computed before rounding and reported with 95% percentile-bootstrap intervals from $10{,}000$ resamples of the five paired replicate differences, and the number of positive differences. Raw and frozen scores are refitted controls for this comparison. The 89% and 68% window reductions use only the fitted readouts from the full sweep in Figure 1.

Dataset	K	raw	Gaussian	bank	frozen	frozen − Gaussian	Gaussian − raw
MC-RTT	0	0.425	0.601	0.567	0.606	$+0.005$ $[-0.003, +0.012]$ (3/5)	$+0.176$ (5/5)
MC-RTT	4	0.625	0.691	0.664	0.641	-0.050 $[-0.069, -0.037]$ (0/5)	$+0.066$ (5/5)
MC-Maze	0	0.728	0.885	0.866	0.890	$+0.005$ $[+0.001, +0.009]$ (4/5)	$+0.157$ (5/5)
MC-Maze	4	0.843	0.890	0.875	0.895	$+0.005$ $[-0.001, +0.009]$ (4/5)	$+0.047$ (5/5)

Predictability from trial position alone. On a stereotyped task, a readout might predict velocity from trial position alone. We measure this possibility before interpreting the time-embedding control. A template fitted on training trials uses only position within the trial, without neural inputs. It achieves held-out R^2 of -0.006 on MC-RTT and $+0.025$ on MC-Maze, compared with neural decoding scores near 0.5 and 0.89. Direction averaging makes this score unsuitable for measuring trajectory stereotypy. A held-out direction-template measure avoids reusing the scored data. Each trial receives a label computed from its first quarter alone, and the label's fitted velocity template is scored on the trial's second half, so scored bins never inform their own trial's label. Across five paired replicates, this measure gives $+0.066$ and $+0.167$, with condition-mean upper bounds of $+0.240$ and $+0.548$ from a labeling that summarizes each trial's own target; both quantities are larger on the stereotyped task. These results show that trial position *alone* does not predict velocity and that a direction label plus position predicts a modest fraction on the stereotyped task. Whether time is informative jointly with the neural input is tested by the position-code equalization below.

Gaussian ordering under index alignment. The frozen-minus-Gaussian gap also declines with K under all three readout settings. Whether it crosses zero depends on both the setting and projection

width, so we treat the ordering as secondary within these index-aligned sweeps. The primary-protocol comparison is in §4.2 (Table 2). Under the masked setting on MC-RTT the contrast crosses zero at $d_{\text{tok}} = 2, 8, 16$ but not at 32; there it stays positive, $+0.021$ at $K = 3$ and $+0.027$ at $K = 4$. On MC-Maze the only negative values, -0.006 at $d_{\text{tok}} = 2$ and -0.001 at $d_{\text{tok}} = 16$, both at $K = 4$, are within about one per-replicate SD (near 0.005) of zero, so we make no ordering claim for those conditions at either projection width.

Sources of shifted-window decodability in the untrained backbone. Shifted-window decoding tests whether a token carries information about other positions in the trial, by circularly shifting the readout's window within the trial and comparing with unshifted decoding. An untrained backbone of BCIM's architecture, with randomly initialized weights, already retains part of its own unshifted decoding score under that shift (the comparison is within the untrained backbone), so pretraining is not required for this retention. Three same-seed ablations on the untrained backbone narrow the attribution, at $K = 0$ with three paired initializations and five replicates each.

We ablate the learned absolute time embedding by either replacing every position's embedding with the zero padding row or flattening the embedding to its mean over positions. Both *raise* shifted-window decoding in $30/30$ paired comparisons across the two datasets. The increases are 0.027–0.028 on MC-RTT and 0.008–0.009 on MC-Maze. The same embedding does contribute to *unshifted* decoding, but only on the stereotyped task, $+0.063$ on MC-Maze in $15/15$ against -0.002 on MC-RTT. Feeding every trial the training-mean spike pattern collapses both conditions to approximately zero, excluding a position-only or template-only bypass. Both ablations point to the content-dependent part of the untrained backbone as the source of the effect. Zeroing removes a common offset together with the between-position differences, whereas flattening removes only those differences; their agreement therefore isolates the differences between positions. This analysis is limited to the untrained backbone. The pretrained model's own time embedding was not ablated, and slot-preserving residual paths and the ordered readout remain as sources of temporal identity.

Position-code equalization. Under index alignment, the backbone adds a learned absolute time embedding before its attention blocks, whereas the readout applies the same weights at every output position and receives no token index. Write X_K and Z_K for the raw and frozen readout windows at half-width K, each a concatenation of $2K+1$ tokens; U for the position code, a one-hot vector over the T output positions; and $\Delta_{\text{FM}-\text{raw}}(K)$ for the measured frozen-minus-raw gap in this analysis. Suppose the frozen representation supplied only the raw window plus trial position, $Z_K \simeq (X_K, U)$. Then $\Delta_{\text{FM}-\text{raw}}(K) = R^2(X_K, U) - R^2(X_K)$, the added value of U given the raw window. That quantity could decline if position becomes inferable from a wider raw window, even without added neural information. The trial-position control above measures marginal predictability from position alone and therefore does not test their joint informativeness.

To test that possibility, both conditions receive an identical one-hot code of the center output position, appended to the concatenated tokens in the readout window. Every condition keeps this input path, and the conditions without the position code feed it zeros, so parameter counts match exactly. On a common test set, target, and R^2 denominator, the decomposition

$$\begin{aligned}\Delta_{\text{FM}-\text{raw}}(K) &= \Xi_U(K) + \Delta_{\text{FM}-\text{raw}|U}(K),\\ \Xi_U(K) &:= \big[R^2(X_K, U) - R^2(X_K)\big] - \big[R^2(Z_K, U) - R^2(Z_K)\big],\\ \Delta_{\text{FM}-\text{raw}|U}(K) &:= R^2(Z_K, U) - R^2(X_K, U)\end{aligned}$$

is an exact accounting identity, not a causal decomposition, and held to 0.00 to floating-point precision in all 12 replicates at every K on both datasets. We verified on a fixed untrained head that the input path is connected, since under an all-zero input its weights receive no gradient and the zero-fed control alone would not show the path is present in the forward pass.

The decline is reproduced under equalization on both datasets. $\Delta_{\text{FM}-\text{raw}|U}$ falls by 89.5% on MC-RTT and 86.9% on MC-Maze, in $12/12$ replicates on each, against 88.8% and 90.2% for $\Delta_{\text{FM}-\text{raw}}$ itself. The accounting share assigned to the position code behaves differently by task. On MC-Maze it is nonzero but partial at $K = 0$, $\Xi_U/\Delta_{\text{FM}-\text{raw}} = 0.26$ with $\Xi_U > 0$ in $12/12$ replicates, and falls to 0.01 by $K = 4$; on MC-RTT it is negative at $K = 0$ (-0.06, positive in $3/12$ replicates) and nonnegative at $K \geq 1$, at most 0.18 at $K = 3$. On the stereotyped task, then, differential access to trial position is assigned part of the frozen model's narrow-window advantage and not most of it, in this accounting sense. Table 17 gives the full decomposition at every width.

Ξ_U also absorbs how a finite regularized readout exploits U and any redundancy or synergy between U and the spikes, so the identity remains an accounting one, as noted above. Giving the readout a one-hot index also does not reproduce pos_i, which enters before attention and interacts nonlinearly with content, so $\Delta_{\mathrm{FM-raw}|U} > 0$ supports only the statement that a gap remains once both conditions receive the same position code index. Ratios are quoted as ratios of means. At $K = 4$, $\Delta_{\mathrm{FM-raw}}$ is small enough that individual per-replicate ratios diverge and are uninformative.

Table 17: Position-code equalization, twelve paired split-and-seed replicates. $\Delta_{\mathrm{FM-raw}}$ is the frozen-minus-raw gap, $\Delta_{\mathrm{FM-raw}|U}$ is the same gap once both conditions receive an identical one-hot position code. The difference $\Xi_U = \Delta_{\mathrm{FM-raw}} - \Delta_{\mathrm{FM-raw}|U}$ measures how much adding the code changes the gap; it is not a causal decomposition. The final row is the reduction from $K = 0$ to $K = 4$; it is not reported for Ξ_U, whose $K = 4$ value is within noise of zero on either dataset.

	MC-RTT			MC-Maze		
K	$\Delta_{\mathrm{FM-raw}}$	Ξ_U	$\Delta_{\mathrm{FM-raw}\|U}$	$\Delta_{\mathrm{FM-raw}}$	Ξ_U	$\Delta_{\mathrm{FM-raw}\|U}$
0	+0.267	−0.017	+0.284	+0.253	+0.066	+0.188
1	+0.212	+0.002	+0.210	+0.118	+0.020	+0.098
2	+0.139	+0.002	+0.137	+0.067	+0.008	+0.058
3	+0.076	+0.014	+0.062	+0.039	+0.002	+0.038
4	+0.030	+0.000	+0.030	+0.025	+0.000	+0.025
reduction K=0 → 4	88.8%	–	89.5%	90.2%	–	86.9%

Low-pass target sweep. Cutoffs are chosen to retain a fixed fraction of within-trial target variance, solved on the training split. That fraction defines the grid but does not fix the pooled held-out variance used by R^2. The latter still changes between the narrowest-band and unfiltered targets, by factors of about 1.5 on MC-RTT and 1.8 on MC-Maze. Table 18 therefore also gives the fixed-denominator mean-squared-error gain $D_{n,\mathrm{g}} := (\mathrm{MSE}_{\mathrm{raw}} - \mathrm{MSE}_{\mathrm{g}})/\mathrm{Var}(v_{\mathrm{unfiltered}})$ for the fixed Gaussian, and likewise $D_{n,\mathrm{FM}}$ for the frozen model, at target manipulation n, a low-pass level here and a frequency band in the paragraph below, abbreviated D_n when the condition is clear.

Under ΔR^2, the MC-RTT frozen-model narrowest-minus-unfiltered contrast is $+0.063$, with that direction in $8/8$ paired reruns. Under D_n the same difference between these targets falls to $+0.020$ ($5/8$ positive), and on MC-Maze it is -0.018 ($0/8$ positive), meaning the absolute advantage is larger for the unfiltered target. The Gaussian contrasts are $+0.036$ and $+0.006$ on MC-RTT and -0.006 and -0.020 on MC-Maze, and these reruns describe split and optimization stability.

Both metrics retain confounds, since D_n grows with how much raw error is available to remove and the relative version divides by a target variance that also moves across the sweep. We therefore treat this design as auxiliary and use the disjoint-band design as the primary test.

Table 18: Low-pass sweep of the velocity target, eight paired split-and-seed replicates. "Level" is the retained fraction of within-trial target variance and ΔR^2 is measured against the raw condition at the same level. D_n uses the unfiltered target variance as a fixed denominator. Subscript g denotes the fixed Gaussian.

Level	Cutoff	Retained var.	raw R^2	ΔR^2_{g}	ΔR^2_{FM}	$D_{n,\mathrm{g}}$	$D_{n,\mathrm{FM}}$
MC-RTT ($\sigma = 4$)							
25%	1.32 Hz	25.2%	0.4288	+0.0880	+0.1304	+0.0587	+0.0875
50%	1.86 Hz	50.1%	0.4801	+0.0799	+0.1049	+0.0635	+0.0836
75%	2.57 Hz	75.3%	0.5012	+0.0670	+0.0955	+0.0594	+0.0846
90%	3.55 Hz	90.2%	0.5003	+0.0616	+0.0797	+0.0585	+0.0757
100%	—	100.0%	0.4890	+0.0524	+0.0672	+0.0524	+0.0672
MC-Maze ($\sigma = 2$)							
25%	1.70 Hz	25.1%	0.8889	+0.0316	+0.0457	+0.0173	+0.0249
50%	2.08 Hz	50.3%	0.8709	+0.0418	+0.0590	+0.0276	+0.0389
75%	2.81 Hz	75.1%	0.8759	+0.0383	+0.0472	+0.0306	+0.0378
90%	3.90 Hz	90.0%	0.8678	+0.0392	+0.0465	+0.0360	+0.0427
100%	—	100.0%	0.8478	+0.0372	+0.0426	+0.0372	+0.0426

Disjoint-band target manipulation. Bands are formed by masking the per-trial discrete Fourier transform (DFT) of the target and are exactly disjoint within the evaluated range. They do not span the full spectrum, since the MC analyses cover $[0, 8)$ Hz, whereas Perich-RT additionally evaluates $[8, 12)$ Hz and has a 50-Hz target Nyquist frequency. Table 19 gives the per-band values summarized in §4.4, and Figure 8 plots the baseline R^2 per band and the gains. The per-trial DFT treats each trial as periodic, so filtering can introduce temporal boundary artifacts.

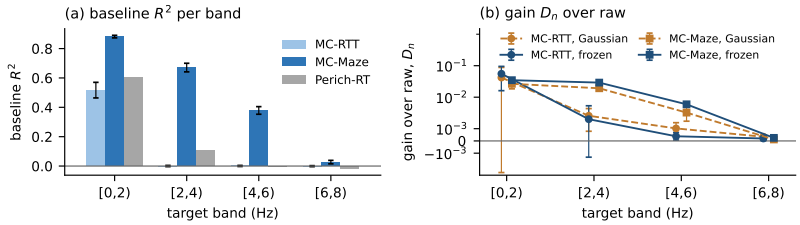

Figure 8: Disjoint-band target manipulation. (a) Baseline R^2 per band, where Perich-RT compares stronger smoothing with its smoothed baseline. Beyond 2 Hz, raw decoding remains near $R^2 = 0$ on MC-RTT but remains positive on MC-Maze. (b) Gain D_n over raw for the two MC datasets. All means are positive or zero at reported precision. Whiskers on the MC bars are sample SDs over eight replicates, not CIs, and the Perich-RT bars carry no whiskers. Values, baselines and the Perich-RT approximation to D_n are in Table 19. Perich-RT has no frozen-feature condition.

On MC-RTT, the small positive gains above 2 Hz do not indicate high-frequency recovery, since raw decodes at chance there ($R^2 \approx 0$), so the gains are consistent with regularization or floor effects. On MC-Maze that precondition is met, since the baseline decodes the target above 2 Hz ($R^2 = 0.67$ at 2–4 Hz). There the gains remain positive, and the frozen model's largest high-band gain, $+0.029$ at 2–4 Hz, coincides with the best-decoded band above 2 Hz.

Perich-RT has no corresponding frozen-model condition. Its canonical benchmark-matched input uses $\sigma = 2$-bin smoothing and scores 0.609 across six sessions against the benchmark's 0.613; the comparison condition uses $\sigma = 6$ bins. Its D_n values are approximations obtained by scaling each session's ΔR^2 by the band's share of unfiltered target variance, then averaging across sessions. An earlier custom ridge readout that fell 0.24 below that reference value was discarded, and with it a collapse to negative R^2 in the high bands that the benchmark-matched readout did not reproduce.

Table 19: Disjoint-band target manipulation. For the MC datasets, D_n is the fixed-denominator gain over raw at the same band; Perich-RT uses the variance-scaled approximation described above. Every dataset-level mean is positive or zero at reported precision, including where the baseline decodes the target above 2 Hz, so no evaluated condition reproduces the sign change of Table 7.

Band	Share of target var.	baseline R^2	$D_{n,\mathrm{g}}$	$D_{n,\mathrm{FM}}$
MC-RTT, eight replicates				
$[0, 2)$ Hz	85.7%	$+0.5170$	$+0.0420$	$+0.0555$
$[2, 4)$ Hz	8.7%	$+0.0002$	$+0.0026$	$+0.0020$
$[4, 6)$ Hz	2.2%	$+0.0007$	$+0.0010$	$+0.0004$
$[6, 8)$ Hz	0.9%	-0.0007	$+0.0003$	$+0.0002$
MC-Maze, eight replicates				
$[0, 2)$ Hz	73.1%	$+0.8816$	$+0.0266$	$+0.0347$
$[2, 4)$ Hz	21.9%	$+0.6707$	$+0.0192$	$+0.0291$
$[4, 6)$ Hz	4.3%	$+0.3787$	$+0.0032$	$+0.0059$
$[6, 8)$ Hz	0.3%	$+0.0269$	$+0.0001$	$+0.0003$
Perich-RT, six sessions, benchmark $\sigma = 2$ baseline versus $\sigma = 6$, no frozen-model condition				
$[0, 2)$ Hz	75.0%	$+0.6027$	$+0.0233$	—
$[2, 4)$ Hz	13.0%	$+0.1080$	$+0.0020$	—
$[4, 6)$ Hz	1.6%	-0.0021	$+0.0000$	—
$[6, 8)$ Hz	1.0%	-0.0169	$+0.0001$	—
$[8, 12)$ Hz	2.3%	-0.0165	$+0.0007$	—

Restart rule for the window sweep. An earlier single-fit version of this sweep contained isolated optimization failures. One split-and-seed replicate of the *raw* condition collapsed at $K \geq 3$ under approximate parameter matching on both datasets. The same seed determined both the trial split and the readout's initial weights on both datasets, so the poor draw recurred across datasets.

For each combination of dataset, readout setting, K, condition and replicate, we fit the readout from three independent random initializations and keep the one with the highest *validation* R^2, reporting that fit's test R^2. The trial split is derived from the replicate and held fixed across restarts, so selection ranges over restarts, never over test splits. The rule uses no threshold estimated from the results and applies identically to all conditions. Tables 12 and 13 and Figure 7a report the sweep.

The rule acts mainly where fits fail. The spread across the three fits is larger for the raw condition than for the filtered conditions. We compute the maximum-minus-minimum test R^2 for each combination and report the largest spread. On MC-RTT it is 0.515 for raw against 0.218 for the frozen condition and 0.102 for the Gaussian under approximate matching, and 0.218, 0.101 and 0.058 at fixed projection width. On MC-Maze all three stay below 0.04. Averaged over combinations, selecting on validation relative to the first restart moves the reported value by only $+0.018$ and $+0.013$ on MC-RTT and $+0.003$ and $+0.006$ on MC-Maze. The rule mainly corrects unusually poor fits.

Under this rule no single replicate drives the gap reduction from $K = 0$ to $K = 4$. Leaving each out in turn gives a reduction of 101–118% on MC-RTT and 98–99% on MC-Maze under approximate matching, where values above 100% mean the gap crosses zero at $K = 4$. At fixed projection width the corresponding ranges are 81–91% and 90–91%, and replicate medians give 115% and 98% under approximate matching and 93% and 89% at fixed width (MC-RTT/MC-Maze in each pair). Relative to the earlier single-fit sweep, the decline is larger because the failed raw fits had inflated the remaining frozen-minus-raw gap at large K.

Randomized-context control. The fixed-width and approximately matched settings do not isolate window width, so we add a control that holds the readout architecture fixed and replaces ordered local neighbors with tokens sampled from other within-trial positions. At each output position, the head retains its $2K + 1$ slots, parameter count and input width. The $2K$ context slots use the sampled positions in place of the local neighbors around i. This changes temporal distance and token content as well as local order. The center slot is always the true token, so the $K = 0$ information is untouched. The stock head zero-pads its window at trial edges.

With ten tokens and $K = 4$, an edge position has four empty slots out of nine; the randomized version reinstates exactly that padding pattern through a slot mask, so the two conditions do not differ in boundary structure. Without that reinstatement the contrast below is inflated on MC-Maze, from $+0.125$ to $+0.230$ at $K = 4$ and fixed width, because unpadded random windows cost raw more.

The position assignment is drawn once per trial and reused at train and test, since redrawing it each epoch would turn the condition into noise injection. Each trial receives its own assignment because one shared across trials would give every position a consistent partner that is informative in stereotyped trials. Everything else follows §4.1, including the restart rule. The quantity estimated is a difference in differences. Write $\Delta^{\text{real}}_{\text{FM}-\text{raw}}(K)$ and $\Delta^{\text{rand}}_{\text{FM}-\text{raw}}(K)$ for the frozen-minus-raw gap under the two context conditions. The quantity is then

$$\Gamma(K) := \Delta^{\text{rand}}_{\text{FM}-\text{raw}}(K) - \Delta^{\text{real}}_{\text{FM}-\text{raw}}(K) = \big(R^2_{\text{raw,real}} - R^2_{\text{raw,rand}}\big) - \big(R^2_{\text{FM,real}} - R^2_{\text{FM,rand}}\big).$$

Thus $\Gamma(K) > 0$ means ordered context benefits raw decoding more.

The measured values are $\Gamma(4) = +0.149$ and $+0.125$ at fixed projection width on MC-RTT and MC-Maze, and $+0.108$ and $+0.100$ under approximate matching. Γ is positive in $5/5$ paired replicates at every $K \geq 3$ in all four conditions, and in $4/5$ in three of the eight conditions at $K \leq 2$, all on MC-RTT. The absolute values support the same interpretation, since at fixed width ordered context raises raw by 0.333 on MC-RTT against the frozen condition's 0.184, and by 0.181 on MC-Maze against 0.057. At $K \geq 2$ randomization preserves a larger frozen-minus-raw gap than ordered context does in all four conditions, and the randomized gap declines part of the way with K (Table 20).

Identical architecture and parameter count within each comparison exclude head size as an explanation for Γ. The control does not yield an exact causal partition of the window effect. A random within-trial token retains trial identity, global state and the model's own position encoding, so ran-

domization removes local order while preserving some information. Randomized slots may also act as distractors; any difference in that cost between conditions enters Γ directly. The pattern of empty trial-edge slots is matched by construction, whereas the randomized fills follow a different distribution from ordered neighbors. Ratios of $\Gamma(K)$ to the total decline (58% and 55% at $K = 4$ and fixed width) describe relative score changes and do not partition the window effect.

Table 20: Frozen-minus-raw difference against readout half-width K, with ordered context and with context gathered from randomized within-trial positions. Means over five paired split-and-seed replicates under the restart rule; $K = 0$ has no context and so no randomized counterpart. Randomized windows reinstate the ordered windows' trial-edge padding pattern. Within a column pair the two conditions have identical architecture and parameter count and differ in how the context tokens are selected. The paired contrast is positive in $5/5$ replicates at every $K \geq 3$.

	fixed projection width				approximately parameter-matched			
	MC-RTT		MC-Maze		MC-RTT		MC-Maze	
K	ordered	random	ordered	random	ordered	random	ordered	random
0	$+0.296$	—	$+0.252$	—	$+0.292$	—	$+0.252$	—
1	$+0.225$	$+0.244$	$+0.121$	$+0.262$	$+0.212$	$+0.272$	$+0.100$	$+0.237$
2	$+0.149$	$+0.213$	$+0.068$	$+0.250$	$+0.148$	$+0.201$	$+0.049$	$+0.208$
3	$+0.068$	$+0.225$	$+0.041$	$+0.198$	$+0.027$	$+0.100$	$+0.004$	$+0.146$
4	$+0.040$	$+0.189$	$+0.024$	$+0.149$	-0.026	$+0.083$	$+0.005$	$+0.104$

Randomized context under the primary protocol. The control above was run under index alignment, where randomizing the neighbors can remove local temporal order and the compensation of the index-alignment offset at once. We therefore repeat it at $K = 4$ under the primary protocol, keeping the time-aligned center token, the shared output-position code, the ordered protocol's slot mask (so the trial-edge zeros and the live-slot count are unchanged) and the constant-parameter layout, and filling the eight non-center slots with tokens drawn once per trial from random positions of the same trial, identically for both representations. Heads are retrained under the randomized input. Table 1 in the main text places the randomized scores beside the ordered scores of Table 9, which share splits and seeds. Randomization lowers raw decoding by 0.211 on MC-RTT and 0.157 on MC-Maze and frozen-feature decoding by 0.136 and 0.100, with the raw drop larger in every replicate (raw minus frozen drop $+0.075$ and $+0.057$, $5/5$ each; paired-t intervals $[+0.034, +0.116]$ and $[+0.038, +0.077]$). The frozen-minus-raw gap under randomized context is $+0.094$ and $+0.109$. We report the two drops and their paired difference, with no recovered fraction. Randomized slots may act as distractors, and any asymmetry in that cost enters the paired difference, so the control does not identify the backbone's computation.

Training-set size under index alignment. This experiment uses index alignment without output-position input and holds Gaussian width fixed at $\sigma = 4$ on MC-RTT and $\sigma = 2$ on MC-Maze. A representation that mainly reduces variance would be expected to show a shrinking advantage as training data grow, at fixed window width. We use 25%, 50%, 75% or 100% of the training split and retain the full validation and test sets within each replicate, so all training fractions are evaluated on the same trials. With $d_{\text{tok}} = 8$, parameter count stays fixed across training fractions within each representation and window width. Table 21 reports the resulting paired differences at the narrowest and at a wide window on both datasets. These are descriptive learning curves. They do not test the residual term of Appendix G, which is not identified, and no asymptote is fitted or claimed. All statements below are about the change in each condition's advantage over raw between 25% and 100% training data. Individual Gaussian-minus-raw curves decrease monotonically in only $1/5$ and $3/5$ replicates at $K = 0$ and $K = 3$ on MC-RTT, and $1/5$ at each width on MC-Maze.

The Gaussian-minus-raw gap decreased between 25% and 100% training data in all four conditions and in $5/5$ paired replicates in each. On MC-RTT it fell from $+0.171$ to $+0.145$ ($K = 0$) and from $+0.176$ to $+0.066$ ($K = 3$), and on MC-Maze from $+0.196$ to $+0.171$ and from $+0.048$ to $+0.034$.

The frozen-minus-raw gap did not decrease in every condition. It decreased between the same training sizes on MC-Maze at both widths, from $+0.321$ to $+0.250$ at $K = 0$ and from $+0.060$ to $+0.042$ at $K = 3$, in $5/5$ replicates each, and on MC-RTT at $K = 3$ in $4/5$. On MC-RTT at $K = 0$

the mean curve instead increased monotonically, from $+0.217$ to $+0.295$, and the gap increased between these training sizes in all five paired replicates. That condition is the narrowest readout on the dataset where raw decoding is weakest (0.095 at the smallest training set), and there the frozen condition's R^2 rises faster with added training data than the raw condition's does.

The Gaussian result is consistent with a sample-efficiency reading without identifying one, since at fixed K a smoothed token still carries raw temporal information from outside the readout's own window, leaving its declining advantage with more than one available explanation. The frozen condition's curve tracks how the gap scales with training-set size rather than what the features make accessible to a narrow-window readout, and so does not test the filtering account.

The prespecified directional prediction was not supported in one of four conditions, providing another reason to avoid decomposing the window effect into approximation and estimation parts. The separate index-aligned validation rerun finds a mean frozen-minus-Gaussian gap below 0.007 on MC-Maze (p. 28). That small gap at one training-set size does not imply that the two inputs yield similar learning curves. The primary-protocol Gaussian comparison in §4.2 holds training-set size fixed and does not retest these curves.

Table 21: Paired differences against training-set size under index alignment without output-position input, at fixed projection width and Gaussian width ($\sigma = 4$ on MC-RTT, $\sigma = 2$ on MC-Maze). Validation and test splits are held at their full replicate values. Means over five paired split-and-seed replicates under the restart rule. The Gaussian-minus-raw gap decreases between 25% and 100% training data in every condition and in $5/5$ replicates. The frozen-minus-raw gap decreases except at $K = 0$ on MC-RTT, where the mean curve rises monotonically and the gap increases between these training sizes in $5/5$ replicates. Individual curves are mostly not monotone throughout. These are descriptive learning curves with no fitted asymptote, and n_{tr} is the number of training trials; $K = 0$ is the narrowest window and $K = 3$ a wide readout. FM and Gauss denote the frozen-minus-raw and Gaussian-minus-raw gaps at the same training size.

	MC-RTT					MC-Maze				
		$K = 0$		$K = 3$			$K = 0$		$K = 3$	
train	n_{tr}	FM	Gauss	FM	Gauss	n_{tr}	FM	Gauss	FM	Gauss
25%	120	+0.217	+0.171	+0.087	+0.176	256	+0.321	+0.196	+0.060	+0.048
50%	240	+0.266	+0.177	+0.066	+0.122	512	+0.288	+0.187	+0.044	+0.034
75%	361	+0.281	+0.155	+0.075	+0.098	768	+0.258	+0.174	+0.045	+0.038
100%	481	+0.295	+0.145	+0.061	+0.066	1024	+0.250	+0.171	+0.042	+0.034

Random-projection and untrained-backbone controls. All conditions use the same MC-RTT examples in the readout experiments at $K = 3$ with five paired split-and-seed replicates, and the untrained-backbone condition additionally uses three independent initializations. Per-initialization means of the unmodified untrained backbone are 0.213–0.385, those of the centered sensitivity variant 0.098–0.473, and within-initialization sample SD across split-and-seed fits is 0.12–0.31. Only the pretrained-minus-untrained gap ($+0.25$) is interpretable. The architecture ends in LayerNorm, which already centers each token, so the unmodified untrained backbone is the primary architecture control. Explicit per-token centering, reported as a sensitivity analysis, gives 0.2252 ± 0.2143 against the unmodified condition's 0.3047 ± 0.0866. Both 256-dimensional conditions pass through a per-token projection of the same narrow width, so part of the pretrained-minus-untrained gap describes pretraining placing task signal in directions that the projection retains. The raw condition's 5-dimensional input is compressed almost losslessly, an asymmetry that affects only the raw comparisons.

Table 22: Random-projection and untrained-backbone controls on MC-RTT at $K = 3$, means $\pm$ sample SD across five paired split-and-seed replicates (last row: across three backbone initializations). That condition predates the restart rule, so it is not comparable with Table 7 or Table 12.

Condition	R^2	vs. raw
raw	0.4952 ± 0.0245	—
fixed Gaussian ($\sigma = 4$)	0.5648 ± 0.0169	$+0.0696$
pretrained frozen model	0.5565 ± 0.0242	$+0.0613$
random 256-d projection of raw tokens	0.3978 ± 0.0190	-0.0974
untrained backbone, unmodified	0.3047 ± 0.0866	-0.1905

D LIMITATIONS

Theory. The Wiener derivation is a local linear minimum mean-square error (LMMSE) approximation rather than the exact Poisson-lognormal posterior mean (§2, Appendix F). The spatial cap assumes a locally parameter-independent covariance with a rank-one signal-aligned component, and more general geometry gives a structured floor instead of the scalar $1/\epsilon$. The temporal result is a second-moment residual result (Appendix F), and the recorded NDC, with three subset repetitions per level, fixes no asymptote.

Data. Pretraining read `neural-pile-primate`, the public release of the primate subset of the Neural Pile (test split excluded), under a cap of $1{,}000$ hours per source dataset, with the same cap applied to AJILE12 as one source. That release does not contain the Neural Latents Benchmark files for MC-Maze and MC-RTT (DANDI 000128 and 000129). It does contain their parent datasets: the reaching recordings of Churchland et al. (2012) (DANDI 000070), the source study of MC-Maze; the release of Makin et al. (2018) (Zenodo 3854034), which includes the MC-RTT source session `indy_20160627_01`; and DANDI 000688 (Perich et al., 2018), from which the Perich-CO and Perich-RT sessions are taken. These three datasets form a small share of the release by token count, roughly 3%, 0.5% and 2%. No spike source reached the $1{,}000$-hour cap, so the release was used in full apart from its held-out test split. Whether the evaluated MC-RTT and Perich sessions were included in pretraining has not been verified. The Churchland recordings also include the MC-Maze animal. The downstream sessions' combined duration is below 0.1% of the corpus. These proportions describe pretraining composition, not the fraction of evaluation data previously seen. The window comparisons hold the checkpoint and evaluation data fixed. Widening improves raw-spike decoding without using the backbone. Prior exposure may affect frozen-feature performance, and these comparisons do not establish transfer to sessions unseen during pretraining.

On MC-RTT, a continuous recording cut into 600-ms trials, the 200-ms forward context of one trial is the first 200 ms of the next, so windows of different splits can overlap. A readout-level check removes this overlap by recovering adjacent trials from their identical 200-ms blocks (250 chains) and assigning whole chains to train, validation and test (five permutations, otherwise the primary protocol). The gap is then $+0.176$ at $K = 0$ and -0.009 at $K = 4$ ($5/5$ decline), against $+0.181$ and $+0.019$ under the random trial split, with the wide-window raw score similar (0.627 against 0.622), so the overlap does not appear to inflate the $K = 0$ gap or the raw context gain, although the grouped split also alters the trial composition. The validation-selected Gaussian comparison of §4.2, repeated under the same grouped split with the same width grid and restart rule, gives frozen-minus-Gaussian differences of $+0.029$ at $K = 0$ ($4/5$ positive, percentile-bootstrap interval $[+0.001, +0.057]$, paired-t interval $[-0.014, +0.073]$) and -0.051 at $K = 4$ ($0/5$, $[-0.062, -0.038]$ and $[-0.069, -0.033]$), against $+0.005$ and -0.050 under the random split. The mean Gaussian score falls from 0.601 to 0.576 at $K = 0$ and from 0.691 to 0.669 at $K = 4$, whereas the frozen score changes by less than 0.001 at $K = 0$ and falls from 0.641 to 0.618 at $K = 4$ (five replicates, selected widths 2 or 4 bins).

Cross-task identification. The positive gaps on the MC tasks and the negative gap on Perich-CO do not identify a causal effect of target bandwidth. With all Perich-RT fits retained, the interval does not establish that either representation decodes better on average. Frozen-feature target-spectrum manipulations cover only the two MC datasets used in the readout experiments. On Perich-RT,

these tests compare smoothing widths without a frozen-feature condition. The tests do not support bandwidth as the sole factor, and signal-to-noise ratio and other task differences are unmanipulated.

Protocols, controls and scope. Absolute scores cannot be compared directly across evaluation protocols, so the evidence is the paired change within each analysis. Fixed width co-varies window width with head size, approximate matching narrows the projection and may penalize frozen features, and the randomized-context control shows differential dependence on ordered context without partitioning the window effect. In the index-aligned sensitivity sweeps, the frozen-minus-raw decline reproduces under all three projection schedules, whereas the frozen-minus-Gaussian ordering is not consistent across them. Under the primary protocol, the decline also reproduces at $d_{\mathrm{tok}} = 32$ and with direct per-bin output (Appendix C), which leaves intermediate output resolutions, other position codes, nonzero lags and $K = 1$ to 3 untested there. The $K = 3$ untrained-backbone control uses three initializations and supports a qualitative gap only. The decoding advantage of cross-channel over channelwise models persists across four untrained initializations. Explained variance varies across backbones and measurement points. The absence of decisive evidence for added abstraction is conditional on the tested readouts and diagnostics. The spatial-to-temporal link is tested only for spike decoding. Downstream decoding on surface ECoG remains untested, and the descriptive sEEG comparison in Appendix B does not establish general cross-modal transfer. We evaluate one self-supervised invasive model, so cross-model generality remains untested (Appendix A).

E Reconstruction optimum and local Wiener approximation

Masked reconstruction targets a conditional mean at its population optimum. The following proofs also state the assumptions needed for the local expansion that yields a Wiener filter.

E.1 Masked reconstruction and the conditional mean

Proposition (Population optimum of masked reconstruction, restated from §2)**.** *Let c denote the visible tokens and x a masked continuous target with $\mathbb{E}\|x\|^2 < \infty$. Under squared loss the population-risk minimizer is $g^\star(c) = \mathbb{E}[x \mid c]$. For Poisson negative log-likelihood with log-rate prediction $\xi(c)$, let the spike count Y have finite conditional mean. Wherever $\mathbb{E}[Y \mid c] > 0$, the unique finite minimizer satisfies*

$$\exp \xi^\star(c) = \mathbb{E}[Y \mid c]. \tag{4}$$

For a latent firing intensity Λ, measured in expected spikes per bin, with $Y \mid \Lambda, c \sim \mathrm{Poisson}(\Lambda)$, this is the posterior-mean rate $\mathbb{E}[\Lambda \mid c]$. The boundary case and latent-rate interpretation follow.

Let c denote the visible context and let x or Y denote the masked token. Assume $\mathbb{E}\|x\|^2 < \infty$ and that the conditional mean of Y is finite almost surely.

Continuous token under squared loss. For the squared-loss case of Proposition 1, let $g(c)$ be a measurable predictor and define

$$m(c) = \mathbb{E}[x \mid c].$$

Conditioning on c and expanding around $m(c)$ gives

$$\begin{aligned}\mathbb{E}\big[\|g(c) - x\|^2 \mid c\big] &= \mathbb{E}\big[\|g(c) - m(c) + m(c) - x\|^2 \mid c\big] \\ &= \|g(c) - m(c)\|^2 + \mathbb{E}\big[\|m(c) - x\|^2 \mid c\big]\,,\end{aligned}$$

because

$$\mathbb{E}[m(c) - x \mid c] = 0.$$

The second term does not depend on g, and the first is minimized uniquely almost surely by

$$g^\star(c) = m(c) = \mathbb{E}[x \mid c].$$

Poisson count under a log link. For the Poisson case, let the decoder predict a log rate $\xi(c)$ and use the Poisson negative log-likelihood, omitting the Y-only term,

$$\mathcal{J}(\xi, Y) = e^\xi - Y\xi.$$

The conditional risk is

$$\mathcal{E}_c(\xi) = \mathbb{E}[e^{\xi} - Y\xi \mid c] = e^{\xi} - \xi\,\mathbb{E}[Y \mid c].$$

Its first and second derivatives are

$$\mathcal{E}_c'(\xi) = e^{\xi} - \mathbb{E}[Y \mid c], \qquad \mathcal{E}_c''(\xi) = e^{\xi} > 0.$$

Hence, whenever $\mathbb{E}[Y \mid c] > 0$, the unique finite minimizer satisfies

$$e^{\xi^\star(c)} = \mathbb{E}[Y \mid c].$$

If $Y \mid \Lambda, c \sim \mathrm{Poisson}(\Lambda)$, the tower property gives

$$\mathbb{E}[Y \mid c] = \mathbb{E}[\mathbb{E}[Y \mid \Lambda, c] \mid c] = \mathbb{E}[\Lambda \mid c].$$

Thus the optimal predicted rate is the posterior-mean firing rate. If $\mathbb{E}[Y \mid c] = 0$, the infimum is approached as $\xi \to -\infty$.

Scope. This proposition characterizes the population optimum of the reconstructed output $\mathsf{Dec} \circ \mathrm{enc}$, where Dec is the reconstruction decoder and $\mathrm{enc}(c)$ is the encoder feature. It leaves open whether the intermediate representation $\mathrm{enc}(c)$ is minimal, in one-to-one correspondence with the conditional mean, or carries additional task-relevant structure; the probes in the main text address those questions. The additive-Gaussian Tweedie identity is not used here.

E.2 The local Wiener approximation

Assume, conditional on the full latent trajectory, that counts are independent across bins and

$$Y_t \mid \{z_{t'}\}_{t'} \sim \mathrm{Poisson}(\bar{\lambda} e^{z_t}),$$

where $\bar{\lambda} > 0$ is the baseline mean count per bin and z_t is zero-mean, wide-sense stationary, and small in amplitude. The exponential admits the first-order expansion

$$\bar{\lambda} e^{z_t} = \bar{\lambda}(1 + z_t) + O(z_t^2).$$

Introducing the mean-zero counting fluctuation δ_t, write

$$Y_t = \bar{\lambda} e^{z_t} + \delta_t, \qquad \mathbb{E}[\delta_t \mid \{z_{t'}\}_{t'}] = 0, \qquad \mathrm{Var}(\delta_t \mid z_t) = \bar{\lambda} e^{z_t}.$$

Define

$$\tilde{Y}_t = \frac{Y_t - \bar{\lambda}}{\bar{\lambda}}.$$

Then

$$\begin{aligned} \tilde{Y}_t &= \frac{\bar{\lambda}(e^{z_t} - 1)}{\bar{\lambda}} + \frac{\delta_t}{\bar{\lambda}} \\ &= z_t + \tilde{\delta}_t + O(z_t^2), \end{aligned}$$

where $\tilde{\delta}_t = \delta_t / \bar{\lambda}$. Around $z_t = 0$,

$$\mathrm{Var}(\tilde{\delta}_t \mid z_t) = \frac{e^{z_t}}{\bar{\lambda}} \approx \frac{1}{\bar{\lambda}}.$$

Conditional independence also gives $\mathrm{Cov}(\delta_t, \delta_{t'} \mid \{z_{t'}\}_{t'}) = 0$ for $t \neq t'$. The conditional zero mean makes $\tilde{\delta}_t$ orthogonal to $z_{t'}$ for every lag, and the counting noise is temporally white with variance $\mathbb{E}[e^{z_t}]/\bar{\lambda} = 1/\bar{\lambda} + O(\mathbb{E}[z_t^2])$. Thus, to first order,

$$S_{\tilde{\delta}}(\omega) \approx \frac{1}{\bar{\lambda}}.$$

The local observation model is therefore

$$\tilde{Y}_t = z_t + \tilde{\delta}_t,$$

with signal spectrum S_z and white-noise spectrum $1/\bar{\lambda}$. For a linear time-invariant (LTI) estimator $\widehat{z} = H * \tilde{Y}$, the error spectrum is

$$S_{\mathrm{err}}(\omega) = |1 - H(\omega)|^2 S_z(\omega) + |H(\omega)|^2 \frac{1}{\bar{\lambda}}.$$

Minimizing this expression pointwise in frequency gives

$$H_W(\omega) = \frac{S_z(\omega)}{S_z(\omega) + 1/\bar{\lambda}}. \tag{5}$$

This is the linear minimum mean-square error (LMMSE) Wiener response. If the approximating processes are jointly Gaussian, it is also the minimum mean-square error (MMSE) estimator in the approximating model. If $S_z(\omega)$ is non-increasing in $|\omega|$, then $H_W(\omega)$ is likewise non-increasing. The result is local, since the exact posterior mean in the original Poisson-lognormal model is generally nonlinear. In a comparison using the variance-stabilizing Anscombe transform on counts, frozen-feature decoding remains higher than decoding from the multi-σ Gaussian bank (Appendix B).

F Spatial limits and temporal residual reduction

In the idealized model below, adding neurons cannot remove signal-aligned noise, which caps the information spatial pooling can supply. The pooled residual is a time series, so temporal estimation can still reduce it when its spectrum differs from the signal's. These results concern residual risk and do not specify a Fisher-information ratio, a downstream R^2, or the neuron-dropping ratio κ.

F.1 Setup

Fix a reference value s_0 and let a scalar parameter s change the population mean locally along

$$f' = \left.\frac{\partial \mu}{\partial s}\right|_{s_0} \in \mathbb{R}^N.$$

Here $\mu(s)$ is the mean response vector, f' is its sensitivity at s_0, and N is the number of recorded units. Evaluate the covariance at the same reference point and assume

$$\Sigma(s_0) = \Sigma_0 + \epsilon f' f'^{\top}, \qquad \Sigma_0 \succ 0, \quad \epsilon > 0,$$

with $\partial\Sigma(s)/\partial s|_{s_0} = 0$, where the second term is a signal-aligned information-limiting component. The local covariance-derivative assumption is required because otherwise the full Gaussian Fisher information contains the additional term $\frac{1}{2}\mathrm{tr}[\Sigma^{-1}\Sigma'\Sigma^{-1}\Sigma']$ with $\Sigma' = \partial\Sigma/\partial s$.

Write

$$I_0 = f'^{\top}\Sigma_0^{-1} f'.$$

Proposition 2 needs a scalar time series, so the population vector is pooled by one fixed linear projection. For a sensitivity-normalized projection, $w^{\top} f' = 1$, the minimum baseline variance is attained by $w_0 = \Sigma_0^{-1} f'/I_0$ and equals $w_0^{\top}\Sigma_0 w_0 = 1/I_0$. Under the full covariance, the same projection has noise variance $1/I_0 + \epsilon$. We evaluate w_0 at s_0 and hold it fixed.

To extend the local model through time, write $s(t) = s_0 + b(t)$ and assume, to first order,

$$\mu(s_0 + b(t)) - \mu(s_0) = f' b(t) + o(|b(t)|).$$

Represent the remaining signal-aligned noise component as $f'\eta(t)$ with $\mathrm{Var}(\eta) = \epsilon$. The spatial model fixes this zero-lag variance but not the autocovariance or spectrum of η.

At finite I_0, the sensitivity-normalized projection also contains a baseline term $\nu_0(t)$ with zero-lag variance $1/I_0$. In the idealized post-pooling limit $I_0 \to \infty$, this term vanishes in mean square. Dropping the $o(|b(t)|)$ remainder and taking this limit defines the leading-order local model; after subtracting the reference mean, its scalar observation is

$$y(t) = w_0^{\top}\big(\psi(t) - \mu(s_0)\big) = b(t) + \eta(t),$$

where $\psi(t) \in \mathbb{R}^N$ is the population response, $b(t)$ is the task-relevant scalar component, and $\eta(t)$ is the projected residual noise. Assume that b and η are zero-mean, wide-sense stationary, and mutually uncorrelated at all lags (equivalently, their cross-spectrum is zero).

Under the sensitivity normalization, the rank-one component $\epsilon f' f'^{\top}$ contributes temporal variance $\epsilon(w_0^{\top} f')^2 = \epsilon$ to the projected observation. Let the spectra of b and η be $S_b(\omega)$ and $S_\eta(\omega)$ on frequency domain Ω ($[-\pi, \pi]$ in discrete time). We therefore use the convention

$$\mathrm{Var}(\eta) = \int_{\Omega} S_\eta(\omega)\frac{d\omega}{2\pi} = \epsilon.$$

F.2 Derivation of the population spatial cap

Proposition (Spatial cap; classical, after Moreno-Bote et al. (2014) and Kanitscheider et al. (2015))**.** *Let a decoded scalar parameter s determine the response mean $\mu(s)$, define its local sensitivity vector as $f' := \partial\mu/\partial s$, and let the response covariance be $\Sigma = \Sigma_0 + \epsilon f' f'^{\top}$, where $\Sigma_0 \succ 0$ is the baseline covariance, $\epsilon > 0$ is the signal-aligned variance coefficient, and the covariance is locally independent of the decoded scalar parameter. Define the Fisher information in the absence of the limiting component by $I_0 = f'^{\top}\Sigma_0^{-1} f'$. Then the Gaussian Fisher information satisfies*

$$I = \frac{I_0}{1 + \epsilon I_0} \longrightarrow \frac{1}{\epsilon} \qquad \text{as} \qquad I_0 \to \infty. \tag{6}$$

The identity is classical. By the Sherman–Morrison formula, $\Sigma^{-1} = \Sigma_0^{-1} - \epsilon\,\Sigma_0^{-1} f' f'^{\top}\Sigma_0^{-1}/(1 + \epsilon I_0)$, and contracting with f' gives $I = I_0 - \epsilon I_0^2/(1 + \epsilon I_0) = I_0/(1 + \epsilon I_0)$, which tends to $1/\epsilon$ as $I_0 \to \infty$. In the isotropic special case $\Sigma_0 = \beta_0 \,\mathrm{Id}_N$ for a population of N neurons, where β_0 is the baseline variance of each neuron, define the mean baseline information per neuron as $\bar{I} := \|f'\|^2/(N\beta_0)$, so $I_0 = N\bar{I}$. Then

$$I(N) = \frac{N\bar{I}}{1 + \epsilon N\bar{I}},$$

the spatial information-limiting cap of Moreno-Bote et al. (2014) and Kanitscheider et al. (2015).

F.2.1 Finite-sample recovery of the rank-one component

The population Fisher cap above is distinct from the finite-sample threshold for recovering the rank-one direction from a sample covariance. The left panel of Figure 9 plots the finite-sample result, which we state separately. In the normalized spiked-Wishart model

$$\Sigma = \beta_0\big(\mathrm{Id}_p + \theta\zeta\zeta^{\top}\big), \qquad \|\zeta\| = 1,$$

ζ is the population direction of the rank-one component and its normalized strength θ is dimensionless. If the dimension p, here $p = N$, and the number of independent samples n_{s} tend to infinity with aspect ratio $p/n_{\mathrm{s}} \to \gamma$, the leading sample eigenvector $\widehat{\zeta}$ has nonzero asymptotic overlap with ζ only when

$$\theta > \sqrt{\gamma},$$

and, above threshold,

$$|\langle\widehat{\zeta}, \zeta\rangle|^2 \to \frac{1 - \gamma/\theta^2}{1 + \gamma/\theta}.$$

For the rank-one term $\epsilon f' f'^{\top}$, the corresponding normalized strength is

$$\theta = \frac{\epsilon\|f'\|^2}{\beta_0}.$$

The eigenvalue threshold is due to Baik et al. (2005) and the eigenvector-overlap law to Paul (2007), both recovered by Benaych-Georges & Nadakuditi (2011). This finite-sample result is not the origin of the $1/\epsilon$ cap, and it leaves open whether adding neurons worsens recovery, because both γ and the normalized strength may change with population size. The numerical check uses $\theta = 2$, $p = 400$, and $n_{\mathrm{s}} = 1600, 800, 400$ for $\gamma = 0.25, 0.5, 1$ respectively, with eight Monte Carlo repetitions per reported value. A single global pseudorandom stream is initialized once at seed 0. The check gives empirical versus theory squared overlaps 0.841 versus 0.833 at $\gamma = 0.25$, 0.694 versus 0.700 at $\gamma = 0.5$, and 0.461 versus 0.500 at $\gamma = 1$. Each empirical value is an eight-repetition mean rather than a converged estimate, so the larger discrepancy at $\gamma = 1$ may be sampling variability.

F.3 Temporal residual and band-separation bound

Proposition 2 (Temporal residual reduction after spatial pooling)**.** *Suppose that after spatial pooling the remaining scalar observation is $y(t) = b(t) + \eta(t)$, where b is the task-relevant scalar signal and η is the projected residual noise. Assume that they are zero-mean, wide-sense stationary, and mutually uncorrelated at all lags (equivalently, their cross-spectrum is zero), with integrable spectra*

S_b and S_η on frequency domain Ω ($[-\pi, \pi]$ in discrete time). The non-causal Wiener LMMSE filter is $H_W(\omega) = S_b(\omega)/\big(S_b(\omega) + S_\eta(\omega)\big)$, with $H_W = 0$ wherever $S_b + S_\eta = 0$, and its mean-square residual risk is

$$\mathcal{E}_W = \int_\Omega \frac{S_b(\omega) S_\eta(\omega)}{S_b(\omega) + S_\eta(\omega)} \frac{d\omega}{2\pi},$$

where the integrand is set to zero wherever $S_b + S_\eta = 0$. If S_b is supported on a task-relevant band B, define

$$\epsilon = \int_\Omega S_\eta(\omega) \frac{d\omega}{2\pi}, \qquad \epsilon_{\text{in}} = \int_B S_\eta(\omega) \frac{d\omega}{2\pi},$$

assume $0 < \epsilon < \infty$, and let $\phi := 1 - \epsilon_{\text{in}}/\epsilon \in [0, 1]$ denote the fraction of residual noise power outside B. Then $\mathcal{E}_W \le \epsilon_{\text{in}} = (1 - \phi)\epsilon$, so the residual-reduction factor relative to the unfiltered identity estimate obeys $\rho_{\text{temp}} := \epsilon/\mathcal{E}_W \ge 1/(1 - \phi)$ for $\phi < 1$, with the limiting interpretation used when $\mathcal{E}_W = 0$ or $\phi = 1$. Under the sensitivity-normalized pooling of the setup in the idealized limit $I_0 \to \infty$, ϵ equals the signal-aligned variance coefficient in equation 6.

For an LTI filter H, the error $\text{err} = b - H * (b + \eta)$ has spectrum

$$S_{\text{err}}(\omega) = |1 - H(\omega)|^2 S_b(\omega) + |H(\omega)|^2 S_\eta(\omega).$$

Minimization is pointwise in ω. Differentiating with respect to the complex conjugate $H^*(\omega)$, or applying the orthogonality principle, yields

$$H_W(\omega) = \frac{S_b(\omega)}{S_b(\omega) + S_\eta(\omega)},$$

with $H_W = 0$ wherever the denominator vanishes. Substitution gives

$$\begin{aligned} S_{\text{err},W}(\omega) &= \frac{S_b(\omega) S_\eta(\omega)}{S_b(\omega) + S_\eta(\omega)}, \\ \mathcal{E}_W &= \int_\Omega S_{\text{err},W}(\omega) \frac{d\omega}{2\pi}, \end{aligned} \tag{7}$$

where the ratio is defined as zero when $S_b + S_\eta = 0$.

If S_b vanishes on B^c, the residual integral is supported on B. Since

$$0 \le \frac{S_b(\omega)}{S_b(\omega) + S_\eta(\omega)} \le 1,$$

we have

$$\begin{aligned} \mathcal{E}_W &= \int_B \frac{S_b(\omega) S_\eta(\omega)}{S_b(\omega) + S_\eta(\omega)} \frac{d\omega}{2\pi} \\ &\le \int_B S_\eta(\omega) \frac{d\omega}{2\pi} = \epsilon_{\text{in}}. \end{aligned}$$

Writing $0 < \epsilon = \int_\Omega S_\eta(\omega) d\omega/(2\pi) < \infty$ and $\phi = 1 - \epsilon_{\text{in}}/\epsilon$, the ratio bound for $\phi < 1$ is

$$\frac{\epsilon}{\mathcal{E}_W} \ge \frac{1}{1 - \phi}, \tag{8}$$

with the limiting interpretation when $\mathcal{E}_W = 0$ or $\phi = 1$, and equality is approached when $S_b \gg S_\eta$.

Interpretation and scope. The baseline residual ϵ is the error of the unfiltered identity estimate $\widehat{b} = y$, not the optimal memoryless estimator. Thus $\epsilon/\mathcal{E}_W$ is a residual-reduction ratio, as stated at the start of this appendix. Temporal observations can reduce the limiting residual left by spatial pooling without increasing the information in a fixed full time series.

Comparison with the optimal memoryless linear estimator. The identity baseline admits a further decomposition within the linear estimator class. For $0 < \beta_b := \mathrm{Var}(b) < \infty$ and $\mathcal{E}_W > 0$, the best zero-delay scalar linear estimator $\widehat{b} = ay$ uses $a^\star = \beta_b/(\beta_b + \epsilon)$ and has risk $\mathcal{E}_{\mathrm{scalar}} = \beta_b\epsilon/(\beta_b + \epsilon)$, so

$$\frac{\epsilon}{\mathcal{E}_W} = \left(1 + \frac{\epsilon}{\beta_b}\right) \frac{\mathcal{E}_{\mathrm{scalar}}}{\mathcal{E}_W}.$$

Within the linear class, the second factor measures the additional gain available from frequency-dependent temporal filtering, whereas the first measures the gain available from scalar shrinkage alone. The first factor tends to one only when $\beta_b/\epsilon \to \infty$, which is not implied by in-band dominance $S_b \gg S_\eta$ on B.

Proposition 2 assumes only zero-mean wide-sense-stationary processes with zero cross-spectrum, so the unrestricted memoryless MMSE optimum is $\mathbb{E}[b_t \mid y_t]$, which need not be linear or equal to $a^\star y_t$. We retain the identity normalization because, under $w_0^\top f' = 1$, its residual is exactly the projected rank-one variance ϵ, whereas $a^\star$ relaxes that gain-one normalization. Accordingly, the right panel of Figure 9 fits an optimal scalar gain in both conditions and is an illustrative fixed-filter comparison rather than a check of $\epsilon/\mathcal{E}_W$. Figure 10 checks that ratio itself.

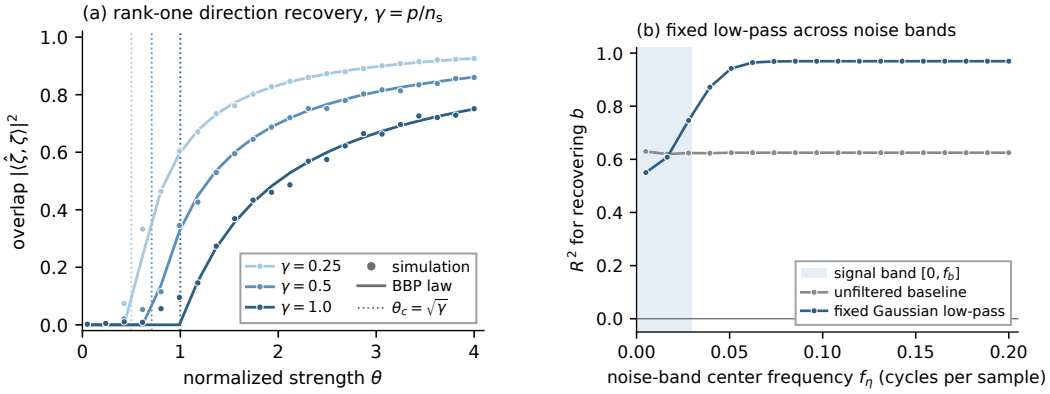

Figure 9: Numerical checks of two distinct results. On the left, finite-sample recovery of the rank-one covariance direction follows the spiked-Wishart Baik–Ben Arous–Péché (BBP) overlap law, where θ is the rank-one component's normalized strength and $\gamma = p/n_s$ is the dimension-to-sample aspect ratio; this is separate from the population $1/\epsilon$ Fisher cap. On the right, in the temporal simulation, the fixed Gaussian low-pass improves the chosen R^2 metric by a mean of $+0.344$ when the limiting noise is spectrally separated ($f_\eta > 3f_b$); in-band ($f_\eta < f_b$) it is approximately neutral on average (mean $+0.011$), with a deepest illustrated point of about -0.08. In that panel, f_b is the signal-band cutoff and f_η the center frequency of the noise spectrum. Both the unfiltered and fixed-low-pass conditions are followed by their optimal scalar gain, so the plotted R^2 difference is an illustrative fixed-filter comparison rather than the ratio $\epsilon/\mathcal{E}_W$.

F.4 Fixed-filter distortion–denoising tradeoff

Proposition 3 (Fixed-filter distortion–denoising criterion)**.** *Under the assumptions of Proposition 2, let $\widehat{b}_H = H * y$ and let the unfiltered identity estimate be $\widehat{b}_{\mathrm{raw}} = y$. Define their mean-square risks by $\mathcal{E}_H := \mathbb{E}|b - \widehat{b}_H|^2$ and $\mathcal{E}_{\mathrm{raw}} := \mathbb{E}|b - \widehat{b}_{\mathrm{raw}}|^2$. Then*

$$\mathcal{E}_H - \mathcal{E}_{\mathrm{raw}} = \int_\Omega \Big[|1 - H(\omega)|^2 S_b(\omega) - \big(1 - |H(\omega)|^2\big) S_\eta(\omega)\Big] \frac{d\omega}{2\pi}. \tag{9}$$

Thus a fixed filter increases risk exactly when its integrated distortion exceeds its noise reduction.

Proof. The filtered risk is

$$\mathcal{E}_H = \int_\Omega \left[|1 - H(\omega)|^2 S_b(\omega) + |H(\omega)|^2 S_\eta(\omega)\right] \frac{d\omega}{2\pi}.$$

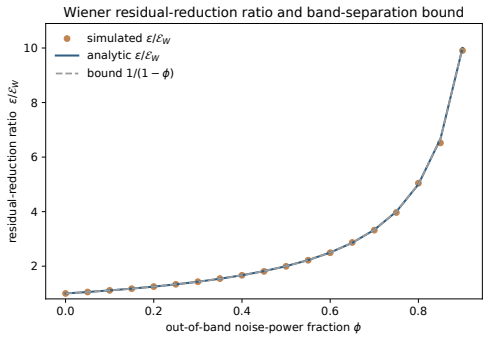

Figure 10: Numerical check of the Wiener residual and band-separation bound. The analytic finite-grid ratio satisfies the lower bound $1/(1-\phi)$; at $\phi = 0.9$ it is 10.00 against the 10.0 bound. The 30-realization Monte Carlo estimate is 9.91, an absolute sampling discrepancy of 0.09 (0.9%). The residual formula gives 0.500 versus 0.502 in simulation.

For the identity estimate $H \equiv 1$,

$$\mathcal{E}_{\mathrm{raw}} = \int_{\Omega} S_\eta(\omega) \frac{d\omega}{2\pi}.$$

Subtracting gives the stated formula. For an ideal passband B, $H = 1$ on B and $H = 0$ on B^c, so

$$\mathcal{E}_H - \mathcal{E}_{\mathrm{raw}} = \int_{B^c} \big(S_b(\omega) - S_\eta(\omega)\big) \frac{d\omega}{2\pi}. \qquad \square$$

Gaussian kernels. For an idealized Gaussian response $H_\sigma(\omega) = e^{-\sigma^2\omega^2/2}$, the distortion term $|1-H_\sigma|^2 S_b$ and the noise-reduction term $(1-H_\sigma^2)S_\eta$ are both nondecreasing in σ at each frequency, so equation 9 is the difference of two nondecreasing functions of the width. Near $\sigma = 0$ the distortion is fourth order, $|1-H_\sigma|^2 = \sigma^4\omega^4/4 + O(\sigma^6)$, whereas the noise reduction is second order, $1-H_\sigma^2 = \sigma^2\omega^2 + O(\sigma^4)$. If both spectra have finite fourth moments,

$$\mathcal{E}_{H_\sigma} - \mathcal{E}_{\mathrm{raw}} = -\sigma^2 \int_{\Omega} \omega^2 S_\eta(\omega) \frac{d\omega}{2\pi} + O(\sigma^4),$$

so a sufficiently narrow Gaussian lowers the risk of the identity estimate whenever the noise spectrum has a nonzero second moment. Any interior stationary width $\sigma > 0$ satisfies

$$\int_{\Omega} \omega^2 H_\sigma(\omega)\big[(1 - H_\sigma(\omega))S_b(\omega) - H_\sigma(\omega)S_\eta(\omega)\big] \frac{d\omega}{2\pi} = 0.$$

This calculation motivates a Gaussian baseline; its width is selected by validation decoding performance (§3.2).

Connection to the empirical sign change. The exact Wiener filter for estimating the same latent b minimizes $\mathcal{E}_H$ and therefore cannot hurt that estimation problem. The empirical sign change concerns a different setting, in which the frozen representation is shaped by firing-rate reconstruction, the downstream target is behavior, and the readout has finite temporal window and capacity. The proposition gives a sufficient condition for filtering to increase mean-squared error but does not determine downstream R^2. Figure 10 checks the residual formula equation 7 and the bound equation 8.

Non-isotropic numerical check. A finite-population simulation compares isotropic white noise with first-order autoregressive (AR(1)) noise, using temporal coefficient $\rho_{\mathrm{AR}} = 0.8$. The AR(1) noise is either independent across channels or includes a shared component with spatial loading proportional to f'. We measure the ratio of estimation MSE before Gaussian smoothing to that after smoothing, fitting a scalar coefficient to each estimate. Relative to white noise, the absolute relative changes in this MSE ratio average approximately 2% and 25%, respectively, over the three tested fractions of out-of-band noise. The ratio increases with this fraction in

all three conditions. These measurements test sensitivity to noise structure rather than the exact Wiener residual formula. Whitening $\Sigma_0 + \epsilon f' f'^\top$ gives normalized spiked-covariance strength $\theta_{\text{eff}} = \epsilon \|\Sigma_0^{-1/2} f'\|^2 = \epsilon f'^\top \Sigma_0^{-1} f'$. The whitened and isotropic overlap checks differ by at most 0.016.

F.5 Implications for neuron-dropping curves

The temporal result establishes that the large-population residual can be lower when temporal structure is used. It does not determine the finite-N curve, where N denotes population size. Consider the generic residual model

$$\mathcal{E}_N = \mathcal{E}_\infty + \frac{\mathcal{E}_{\text{priv}}}{N},$$

where $\mathcal{E}_\infty$ is the limiting component and $\mathcal{E}_{\text{priv}}/N$ is finite-population private noise, both depending on the representation. The corresponding inverse-error quantity is

$$\frac{1}{\mathcal{E}_N} = \frac{1}{\mathcal{E}_\infty} \frac{N}{N + \mathcal{E}_{\text{priv}}/\mathcal{E}_\infty}.$$

Both the asymptote $1/\mathcal{E}_\infty$ and the half-saturation count $\mathcal{E}_{\text{priv}}/\mathcal{E}_\infty$ can change with the representation.

A denoiser that acted like recording more units would shift the neuron-dropping curve horizontally, so the ratio of half-saturation unit counts was fitted to test that reading. Empirically, the curves are obtained by subsampling N units, refitting the matched readout, recording test R^2, and fitting

$$R^2(N) = \frac{R^2_\infty N}{N + N_{50}}, \qquad \kappa = \frac{N_{50}^{\text{raw}}}{N_{50}^{\text{denoised}}}, \tag{10}$$

with N_{50} the fitted unit count at half of the asymptote and R^2_∞ the fitted $N \to \infty$ asymptote. Because R^2_∞ is fitted separately per condition, κ compares half-saturation scales after each curve is normalized by its own asymptote, so it does not measure a pure horizontal translation. Its interval is conditional on the recorded dataset, readout and fitted family.

Neuron-dropping and large-population behavior. In the fitted MC-RTT neuron-dropping curves (Figure 11), the multi-σ bank and the frozen model decode above raw at similar N_{50}. A GP–Poisson simulation is consistent with temporal residual reduction persisting at the largest evaluated population size, and the ordering $\tau_{\text{res}} < \tau_{\text{vel}}$ in both datasets is consistent with a low-pass filtering account (values in Appendix B). Both readings are descriptive. The resampling interval for κ_{FM} includes one, so the curves are compatible with $\kappa = 1$ and identify neither an asymptote nor a vertical shift.

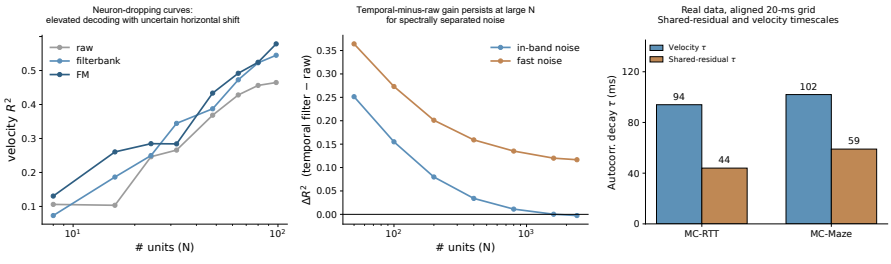

Figure 11: Evidence consistent with temporal residual reduction that persists after spatial pooling (§F.5). On the recorded MC-RTT neuron-dropping curve (left), multi-σ-bank and frozen-model decoding lie above raw at similar fitted N_{50}, but the κ interval $[0.88, 2.10]$ is too wide to exclude a change of up to about a factor of two. In the GP–Poisson simulation (middle), the temporal-minus-raw advantage persists through $N = 2400$ for fast noise but not for in-band noise. In both recorded datasets (right), the estimated shared-residual timescale is shorter than the velocity timescale. These finite-population results are consistent with a band-separation account but establish neither an asymptote nor the behavior of the half-saturation count.

F.6 CONNECTION TO BCIM

Proposition 1 constrains the reconstruction output at the population optimum, and the local Wiener approximation of equation 5 supplies a linearized model for spike-rate denoising. Neither result identifies the intermediate feature with a Wiener filter. The Gaussian-sweep comparison, the readout-window sweep, the random-projection and untrained-backbone controls, the residual probe, the reconstruction comparison and the transfer diagnostics are the empirical tests of whether the frozen representation behaves as the filtering account predicts. The temporal residual result supplies a mechanism by which temporal context can reduce the signal-aligned residual left after spatial pooling, and the fixed-filter decomposition shows why a representation suited to reconstruction can mismatch a task. The model is multimodal. Proposition 1's squared-loss result also applies to its continuous reconstruction head, but the Poisson linearization, analysis of noise remaining after spatial pooling and fixed-filter criterion are developed for the spike branch. The sEEG experiment is therefore a descriptive test of the broader filtering account only.

G BOUND ON THE FROZEN-FEATURE ADVANTAGE

Appendix F concerns what a denoiser can recover from a population of noisy spike trains. The window sweep of §4.1 instead measures how the frozen representation's downstream advantage changes as the *readout* receives more temporal context. We relate the two by deriving a requirement of any finite-support filtering account: the positive advantage must vanish with window width.

G.1 THE BOUND

Fix a velocity target v_i at token position i. Let $\mathcal{W}_K$ be the set of all measurable functions of the raw window $(x_{i-K}, \dots, x_{i+K})$, and write $R^{2,\mathrm{pop}}_{\mathrm{raw}}(K) := \sup_{\alpha \in \mathcal{W}_K} R^2(\alpha)$ for the population optimum over that class; define $R^{2,\mathrm{pop}}_{\mathrm{FM}}(K)$ the same way with the frozen tokens Z in place of x. Quantities marked pop are population optima; the unmarked R^2 and $\Delta_{\mathrm{FM-raw}}$ of §4.1 are measured test-set values.

Proposition 4 (Readout-window bound)**.** *Suppose the frozen representation has temporal half-support L, that is, $Z_i = \chi_i(x_{i-L}, \dots, x_{i+L})$ for a measurable χ_i at every i. Then for every $K \geq 0$*

$$\Delta^{\mathrm{pop}}_{\mathrm{FM-raw}}(K) \;:=\; R^{2,\mathrm{pop}}_{\mathrm{FM}}(K) - R^{2,\mathrm{pop}}_{\mathrm{raw}}(K) \;\;\leq\;\; R^{2,\mathrm{pop}}_{\mathrm{raw}}(K+L) - R^{2,\mathrm{pop}}_{\mathrm{raw}}(K).$$

Proof. Take any measurable readout α acting on $(Z_{i-K}, \dots, Z_{i+K})$. Each Z_{i+r} with $|r| \leq K$ is by hypothesis a measurable function of $(x_{i+r-L}, \dots, x_{i+r+L})$, and every index appearing there lies in $[i-K-L,\, i+K+L]$. Hence α composed with the per-position maps $\chi_{i-K}, \dots, \chi_{i+K}$ is a measurable function of the raw window of half-width $K+L$, so it lies in $\mathcal{W}_{K+L}$ and $R^{2,\mathrm{pop}}_{\mathrm{FM}}(K) \leq R^{2,\mathrm{pop}}_{\mathrm{raw}}(K+L)$. Subtracting $R^{2,\mathrm{pop}}_{\mathrm{raw}}(K)$ gives the claim. □

Fixed filters as a special case. The proof uses only that each frozen token is a function of a raw window of half-support L. For a fixed temporal filter, let S be a token half-width that covers the raw inputs needed to compute each filtered token. The truncated Gaussian of §3.2 has this finite-support property, so the proposition applies with $L = S$ and gives, for every K,

$$\Delta^{\mathrm{pop}}_{\mathrm{Gauss-raw}}(K) \;\leq\; R^{2,\mathrm{pop}}_{\mathrm{raw}}(K+S) - R^{2,\mathrm{pop}}_{\mathrm{raw}}(K).$$

Appendix G.2 discusses the finite-sample qualifications; Table 16 reports the empirical raw-spike and Gaussian decoding scores at the two tested windows.

(i) The positive part of the advantage. $K \mapsto R^{2,\mathrm{pop}}_{\mathrm{raw}}(K)$ is nondecreasing and bounded above, so its increments tend to zero and the bound gives $\limsup_{K\to\infty} \Delta^{\mathrm{pop}}_{\mathrm{FM-raw}}(K) \leq 0$, equivalently $[\Delta^{\mathrm{pop}}_{\mathrm{FM-raw}}(K)]_+ \to 0$. A lossy representation may leave $\Delta^{\mathrm{pop}}_{\mathrm{FM-raw}}$ at a strictly negative value, so the one-sided bound does not imply $\Delta^{\mathrm{pop}}_{\mathrm{FM-raw}}(K) \to 0$. The bound does not require monotonic decline at finite K or specify a finite window at which the positive advantage vanishes. Its application to BCIM depends on the temporal-support premise discussed in (iii).

(ii) Implication for a negative advantage. At large K the one-sided bound permits $R^{2,\mathrm{pop}}_{\mathrm{FM}}(K) < R^{2,\mathrm{pop}}_{\mathrm{raw}}(K)$ when the maps χ_i are not invertible on the target-relevant part of the window. A prediction would additionally require that the discarded content is target-relevant and recoverable by a wide-window raw readout, conditions not supplied by the proposition. The MC-RTT zero crossing under approximate matching (-0.026 at $K = 4$) is consistent with lossy filtering.

(iii) Scope of the finite-support premise in BCIM. The BCIM backbone applies non-causal temporal self-attention over the whole token sequence, with no local or causal attention mask during feature extraction, so Z_i depends in principle on every token of the trial. Its explicit temporal-position signal enters through the learned absolute time embedding pos_i added before the attention blocks (Appendix B). Taking L equal to the trial length satisfies the premise. Once the raw readout has access to the full trial, an unrestricted readout can reproduce the frozen pipeline, so the frozen representation has no positive advantage at the population optimum.

With ten tokens per trial, $K = 4$ already spans nine of them at central positions, and the measured $\Delta_{\mathrm{FM-raw}}(4)$ is -0.026 to $+0.040$ across datasets and settings in the index-aligned sweeps. This observation is consistent with the bound but does not test it, because the readout is zero-padded at trial edges and a fitted MLP need not attain the population optima considered by the proposition. This use of trial-length support does not justify inferring a small effective support from these measurements. The smallest L compatible with an observed $\Delta_{\mathrm{FM-raw}}(K)$ is determined jointly by the readout family, the target and the fitted readouts.

G.2 Finite-sample window measurements

Proposition 4 concerns population optima, whereas every measured value is a test-set R^2 of a fitted readout. Writing $\Upsilon(K) := R^{2,\mathrm{pop}}(K) - R^2(K)$ for the *signed* discrepancy between the population optimum for a condition and the test-set value of its fitted readout, with Υ_{raw} and Υ_{FM} for the two conditions, the measured difference is $\Delta_{\mathrm{FM-raw}}(K) = \Delta^{\mathrm{pop}}_{\mathrm{FM-raw}}(K) + \Upsilon_{\mathrm{raw}}(K) - \Upsilon_{\mathrm{FM}}(K)$, and expressing the bound through measured rather than population quantities gives

$$\Delta_{\mathrm{FM-raw}}(K) \;\le\; \big[R^2_{\mathrm{raw}}(K+L) - R^2_{\mathrm{raw}}(K)\big] \;+\; \big[\Upsilon_{\mathrm{raw}}(K+L) - \Upsilon_{\mathrm{FM}}(K)\big].$$

The quantity $\Upsilon(K)$ can be negative, because the reported R^2 is evaluated on a finite test set and can exceed its population counterpart. These measurements leave the sign of the discrepancy term free and its value unidentified, and the term does not quantify estimation error saved by denoising.

The discrepancy term absorbs at least the approximation error of the fixed MLP family, optimization and early-stopping error, the effect of weight decay and dropout, the d_{tok} compression, the asymmetry between a 5-dimensional and a 256-dimensional per-token input, and test-set sampling error. Several of these do not vanish as training data grow. Establishing that the discrepancy term shrinks with sample size would require assumptions about uniform convergence, optimization convergence and the expressive power of the readout family beyond the present analysis.

Finite support under index alignment. This check uses the fixed Gaussian conditions in Table 12. The implementation truncates each Gaussian kernel at 4σ bins. At the fixed widths $\sigma = 4$ on MC-RTT and $\sigma = 2$ on MC-Maze, the radii are 16 and 8 bins, corresponding to $L = 4$ and $L = 2$ tokens at five bins per token. These calculations concern the index-aligned sweeps; the primary-protocol kernels are described with Table 16.

Within the evaluated range of K this leaves only $K = 0$ checkable on MC-RTT and $K \le 2$ on MC-Maze. Where checkable, the bound is respected on MC-RTT with slack 0.211 at fixed width and 0.233 under approximate matching, and on MC-Maze at $K = 0$ with slack 0.019 at fixed width and 0.033 under approximate matching. On MC-Maze it is exceeded at $K = 1$ and $K = 2$ at fixed width, by 0.0035 and 0.0104, and at $K = 2$ under approximate matching, by 0.0003; $K = 1$ under matching is respected, with slack 0.017. All four MC-Maze magnitudes at $K = 1$ and $K = 2$ are at or below the replicate dispersion of the underlying conditions. This weak check has little discriminating power and does not calibrate the residual term.