BURST: GENERATOR-FIRST SPIKE-COUNT GENERATION WITH ADAPTIVE DISPERSION FOR CROSS-DAY NEUROPROSTHESIS CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intracortical speech decoders degrade across recording days, requiring repeated collection of labeled target-day trials for recalibration. We ask whether generated phoneme-conditioned target-day spike counts can substitute for part of this collection. From an observation-process perspective, we find that attempted-speech counts exhibit rate-dependent sub-Poisson variability and a Fano-rate relationship that changes across sessions beyond estimation noise. We introduce **B**irth–Death **U**nder-Dispersed **R**ates for **S**pike-**T**rain Synthesis (**BURST**), a generator-first count bridge whose state-dependent jump rates define the conditional path and provide explicit dispersion control. Finite-state propagators evaluate analytic reverse-rate targets for simulation-free bridge matching. A historical Fano-rate profile sets the base schedule; a few-shot day code jointly adapts cross-channel tuning and a bounded profile displacement, and an alignment-free phoneme conditioner supports variable-length sequences from sequence-level labels. Across two transfer horizons on an intracortical speech dataset, equal mixtures of real and generated sentences consistently outperform real-only calibration at the same recording budget and approach calibration with twice as many real sentences. Generated counts reproduce the empirical high-rate sub-Poisson regime. BURST therefore provides a data-efficient route to cross-day neuroprosthesis calibration.

1 INTRODUCTION

Implanted speech neuroprostheses now decode attempted speech into text and synthesized voice at increasingly practical rates and vocabularies, with intracortical systems using multichannel spike counts as decoder input (Moses et al., 2021; Metzger et al., 2023; Willett et al., 2023; Card et al., 2024; Wairagkar et al., 2025). Performance nevertheless degrades across days as neural state and recording conditions change, requiring historical decoders to be recalibrated with freshly recorded labeled trials and imposing a recurring recording burden during long-term use (Chestek et al., 2011; Perge et al., 2013; Willett et al., 2023). This burden motivates a direct question: **can target-day spike counts generated for new phoneme sequences substitute for additional recordings?**

We answer this question at the level of the electrode-recorded counts delivered to the decoder. The array records extracellular voltage deflections; filtering and threshold detection extract candidate events, and temporal binning converts them into non-negative integer counts (Gold et al., 2006; Willett et al., 2023). These measurement steps determine the statistical structure observed by the decoder and therefore the distribution the generator must reproduce. In attempted-speech recordings, the Fano factor decreases with firing rate and falls below the Poisson reference when activity is dense. Both neural refractoriness and an effective exclusion interval in threshold detection can produce this short-lag suppression, and their individual contributions remain unresolved. The fitted Fano-rate relationship also varies across sessions beyond within-session estimation noise. A generator that misses this relationship can expose the calibration decoder to a variability structure that differs from real target-day counts, even when mean activity is matched. Faithful target-day synthesis must therefore preserve integer support, reproduce rate-dependent sub-Poisson variability, and adapt its strength to the target day.

To meet these distributional requirements, we introduce Birth–Death Under-Dispersed Rates for Spike-Train Synthesis (BURST), a generator-first count bridge. BURST defines each coordinate

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.

through a birth rate λ_τ and a count-dependent death rate $\kappa_\tau x^{\nu_\tau}$, which together determine the bridge dynamics and conditional path. The frozen-rate stationary law is Conway–Maxwell–Poisson, making ν_τ an explicit dispersion control. A historical Fano-rate profile calibrates the pooled schedule, and a bounded scalar from the target-day code adjusts its strength. Finite-state propagators evaluate the conditional probabilities, yielding analytic reverse rates and a simulation-free bridge-matching loss. Calibration additionally requires each generated trial to express a requested variable-length phoneme sequence in the neural state of a target day represented by only a few recordings. BURST learns the sequence-to-bin correspondence from sequence-level labels with an alignment-free phoneme conditioner, while a permutation-invariant set encoder maps the few-shot trials to a channel-resolved day code. This code adjusts cross-channel tuning through low-rank mixing and affine modulation and supplies the bounded dispersion displacement. Training-only semantic and neural-state critics supervise both conditions and are removed at deployment. In our experiments, we augment the real calibration set with generated counts conditioned on phoneme sequences from a disjoint set of target-day training trials.

On the T12 intracortical speech dataset (Willett et al., 2023), an equal mixture of real and BURST-generated sentences consistently improves cross-day calibration over real-only calibration at the same recording budget and approaches the performance obtained with twice as many real sentences under both transfer horizons. Ablations across calibration budgets show that this substitution behavior remains stable as the amount of target-day data changes, while the generated counts reproduce the empirical high-rate sub-Poisson variability, supporting both substitution value and distributional fidelity. Our contributions are summarized as follows:

(1) To the best of our knowledge, we introduce the first framework that generates target-session intracortical counts jointly conditioned on an unaligned phoneme sequence and a few-shot day state for cross-day speech calibration.

(2) We propose a generator-first birth–death count bridge with analytically derived reverse-rate targets evaluated by finite-state propagators, explicit Fano-calibrated rate-dependent dispersion, and joint day adaptation of cross-channel tuning and dispersion.

(3) Across two transfer horizons, BURST consistently improves over real-only calibration at the same recording budget, approaches calibration with twice as many real sentences, and reproduces the measured high-rate count dispersion.

2 Related Work

2.1 Cross-Day Speech Decoding

Non-invasive recordings have recovered speech representations and the semantic content of continuous language (Défossez et al., 2023; Tang et al., 2023). High-throughput attempted-speech neuroprostheses, however, have relied on invasive cortical recordings for more localized, higher-bandwidth measurements. ECoG has supported phoneme recognition and speech synthesis, whereas intracortical arrays capture action-potential-related events as counts and have enabled large-vocabulary decoding with rapid calibration (Mugler et al., 2014; Anumanchipalli et al., 2019; Metzger et al., 2023; Willett et al., 2023; Card et al., 2024). Intracortical counts vary with neural state and recording conditions, degrading decoder performance across sessions and motivating recalibration (Chestek et al., 2011; Perge et al., 2013). Session-invariant representations and test-time adaptation reduce this shift at the decoder while leaving the paired target-day training set unchanged (Zhang et al., 2026; Feghhi et al., 2025). Generator adaptation has been explored for motor BCIs with continuous behavioral conditions (Wen et al., 2023). Speech introduces the additional challenge that a phoneme sequence specifies symbol order without a bin-level neural alignment. To our knowledge, prior work has not generated phoneme-conditioned target-day intracortical counts for cross-day speech calibration. BURST addresses this gap by using a requested sequence and a few target-day trials to generate additional labeled calibration pairs without further target-day recordings.

2.2 Neural Count Generation

Action potentials generate extracellular voltage deflections that filtering, threshold detection, and binning convert into the non-negative counts consumed by a speech decoder (Gold et al., 2006). A calibration generator must reproduce the resulting count distribution as well as the latent neural dynamics. SpikeGAN and LFADS use adversarial or sequential latent models, while LDNS combines

latent diffusion with a Poisson observation (Molano-Mazón et al., 2018; Pandarinath et al., 2018; Kapoor et al., 2024). Their formulations prioritize population dynamics; recording-specific count statistics remain fixed or implicit, and LDNS ties conditional variance to the decoded mean. Count Bridges develops stochastic count bridges with closed-form conditionals for transcriptomics (Fishman et al., 2026), while Count-FM trains birth–death rates from endpoint-conditioned binomial paths (Wei & Pearson, 2026). In these formulations, recording-specific dispersion remains implicit in the endpoint distribution or path schedule, with no explicit calibration on historical recordings or adjustment for a target session. Recorded counts inherit neural and acquisition variability, so their dispersion can be non-Poisson and session dependent (Chestek et al., 2011; Perge et al., 2013; Stevenson, 2016; Wei & Stevenson, 2023). BURST explicitly controls this cross-session dispersion: a pooled Fano-rate profile calibrated from historical recordings defines the base bridge, and the few-shot day code jointly adjusts cross-channel tuning and a bounded displacement of that profile. This construction adapts the target-day distribution at both the neural-state and dispersion levels.

3 Preliminaries

3.1 Signal and Task

The application context is phoneme decoding from intracortical spiking activity. The signal is a firing-count stream $\mathbf{x} \in \mathbb{Z}_{\geq 0}^{T \times C}$, where $x_{t,c}$ is the number of spiking events detected on channel c within time bin t. The decoding target is a phoneme sequence $\mathbf{y} = (y_1, \ldots, y_L)$, where each phoneme y_l belongs to a finite vocabulary $\mathcal{V}$ and the sequence length is L. The decoder that maps $\mathbf{x}$ to $\mathbf{y}$ drifts across days under electrode displacement and neural-state change, so a target day still requires fresh labeled trials for decoder calibration, and this daily recording burden is what we address. Our task is generative: record only a small set $S_{d^\star}$ on a target day $d^\star$ and expand it into a full training set by generating $\hat{\mathbf{x}} \sim p_\theta(\mathbf{x} \mid \mathbf{y}, S_{d^\star})$ for arbitrary phoneme sequences $\mathbf{y}$, where the historical trials supply the generator and $S_{d^\star}$ supplies the neural state of the target day. The empirical distribution of the historical trials is the data distribution q that the generator learns.

3.2 Rate-dependent Dispersion of Spike Signals

Spike counts present a different modeling target from the continuous time-frequency structure of field potentials. They are point-process observations: small non-negative integers dominated by zeros. Their dispersion about the firing rate, summarized by the Fano factor $\mathrm{Var}(x)/\mathbb{E}[x]$, is a defining statistic that is blurred by continuous relaxation and cannot be recovered through rounding. Physiology constrains this statistic in a specific direction: refractoriness and the dead time of threshold detection suppress count variability precisely when activity is dense, pushing the counts below Poisson variability.

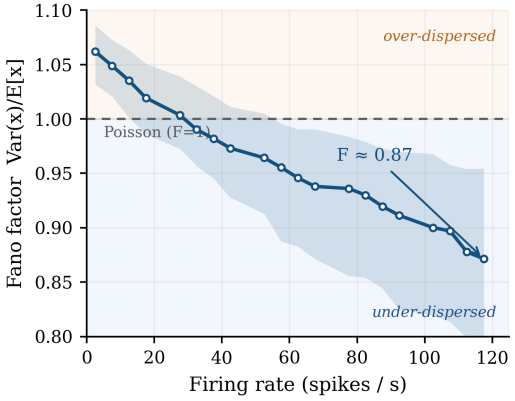


Figure 1: Per-channel Fano factor $\mathrm{Var}(x)/\mathbb{E}[x]$ of T12 firing counts versus firing rate.

The T12 recordings show this regime. Measured with the rate-stratification estimator of Appendix C, the per-channel median Fano factor decreases from approximately one at low firing rates to about 0.87 at high rates (Figure 1). This trend is consistent with neural refractoriness and threshold-detection dead time. The standard generators considered here lack this rate-dependent constraint: an unbounded Poisson law has unit Fano, and the evaluated negative-binomial and Gaussian-latent generators miss the measured sub-Poisson curve. We therefore model the counts directly and make dispersion an explicit function of rate in the process of Sec. 4.1.

4 Birth–Death Under-Dispersed Rates for Spike-Train Synthesis

BURST is a count-bridge flow-matching generator with four parts: a birth–death bridge over count states whose dispersion is prescribable by design (Sec. 4.1), an analytic reverse-rate identity eval-

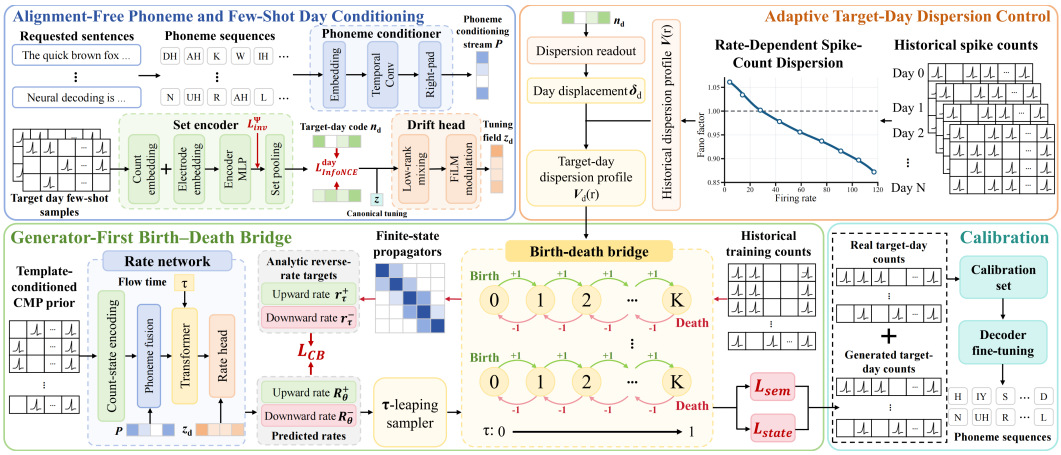


Figure 2: Overview of BURST for data-efficient cross-day neuroprosthesis calibration. A generator-first birth–death bridge models integer spike counts with rate-dependent sub-Poisson variability, while finite-state propagators provide analytic reverse-rate targets for simulation-free bridge matching. Alignment-free phoneme conditioning specifies the requested content, and a few-shot day code jointly adapts cross-channel tuning and a historical dispersion profile. Generated target-day counts augment real recordings for decoder calibration, reducing the need for additional labeled trials. Red arrows "←" denote training-only paths and losses.

uated using finite-state numerical propagation for bridge matching (Sec. 4.2), alignment-free conditioning on the phoneme sequence and on the target day (Sec. 4.3), and a training objective with auxiliary critics that are discarded at deployment (Sec. 4.4).

4.1 A GENERATOR-FIRST BIRTH–DEATH BRIDGE WITH PRESCRIBABLE DISPERSION

We model each entry $x_{t,c}$ by a time-inhomogeneous birth–death chain on the non-negative integers, specified directly through its jump rates: at state x and flow time τ, the count rises by one at rate $b_\tau(x)$ and falls by one at rate $d_\tau(x)$, as follows:

$$b_\tau(x) = \lambda_\tau, \qquad d_\tau(x) = \kappa_\tau\, x^{\nu_\tau}, \tag{1}$$

where $\lambda_\tau > 0$ is the arrival rate, $\kappa_\tau > 0$ the removal scale, and $\nu_\tau > 0$ a dispersion exponent that sets how fast removal accelerates with the count. The dispersion these rates induce is read off the stationary distribution p_∞ at frozen rates, a diagnostic object that the time-inhomogeneous bridge itself never reaches. Detailed balance, $p_\infty(x{+}1)/p_\infty(x) = b_\tau(x)/d_\tau(x{+}1)$, gives

$$p_\infty(x) \;\propto\; \frac{(\lambda_\tau/\kappa_\tau)^x}{(x!)^{\nu_\tau}}, \tag{2}$$

the Conway–Maxwell–Poisson law, in which the $x!$ of the Poisson law is raised to the power ν_τ. The single exponent ν_τ therefore controls dispersion, made precise by the following.

Lemma 1 (Dispersion control). *Fix $\lambda_\tau, \kappa_\tau > 0$ and write $\mu = \lambda_\tau/\kappa_\tau$. For the Fano factor* $\mathrm{Fano} = \mathrm{Var}(x)/\mathbb{E}[x]$ *of equation 2: (i) $\nu_\tau > 1$ implies* $\mathrm{Fano} < 1$, *$\nu_\tau = 1$ recovers the Poisson law with* $\mathrm{Fano} = 1$, *and $\nu_\tau < 1$ implies* $\mathrm{Fano} > 1$; *(ii) as $\mu^{1/\nu_\tau} \to \infty$ at fixed ν_τ, $\mathbb{E}[x] = \mu^{1/\nu_\tau} + O(1)$ and* $\mathrm{Var}(x) = \mu^{1/\nu_\tau}/\nu_\tau + O(1)$, *hence* $\mathrm{Fano} \to 1/\nu_\tau$. *The proof is in Appendix B.*

Clause (i) fixes the side of the Poisson reference, while clause (ii) calibrates the dispersion magnitude. We invert the pooled Fano-rate curve of Sec. 3.2 to obtain a base profile $\nu(r)$ and apply a bounded target-day adjustment,

$$\nu_d(r) = 1 + \big(\nu(r) - 1\big)\exp(\delta_d), \qquad \delta_d = \mathrm{clip}\big(g_\nu(\bar{\mathbf{n}}_d), -\delta_{\max}, \delta_{\max}\big). \tag{3}$$

Here $\bar{\mathbf{n}}_d = C^{-1}\sum_c \mathbf{n}_d[c]$ is the channel-pooled day code of Sec. 4.3, and g_ν is a linear readout trained through the sample-level terms of equation 9. The adjustment preserves $\nu_d(r) \geq 1$ and recovers the pooled profile at $\delta_d = 0$; Appendix D motivates the day dependence, and Appendix E specifies the schedule.

4.2 Reverse-Rate Identity and Simulation-Free Bridge Matching

We index the generative path from a source distribution at $\tau = 0$ to the empirical count distribution at $\tau = 1$, with x_1 denoting a real training endpoint. Its conditional generative rates $r_\tau^+(y \mid x_1)$ and $r_\tau^-(y \mid x_1)$ are obtained by time-reversing the endpoint-conditioned reference chain and give the rates at which a sample at state y raises or lowers its count by one. Write $p_\tau(\cdot \mid x_1)$ for the conditional path, the law at flow time τ of the forward chain pinned to reach the datum x_1 at $\tau = 1$. For the birth–death chain, these rates have analytic expressions in terms of the forward rates and conditional probabilities.

Lemma 2 (Reverse rates)**.** *For the forward rates in equation 1 with conditional path $p_\tau(\cdot \mid x_1)$, the generative reverse rates of the time-reversed birth–death chain (Anderson, 1991) are*

$$r_\tau^-(y \mid x_1) = b_\tau(y-1)\,\frac{p_\tau(y-1 \mid x_1)}{p_\tau(y \mid x_1)}, \qquad r_\tau^+(y \mid x_1) = d_\tau(y+1)\,\frac{p_\tau(y+1 \mid x_1)}{p_\tau(y \mid x_1)}, \tag{4}$$

which are finite and non-negative wherever $p_\tau(y \mid x_1) > 0$. The proof is in Appendix B.

The Gaussian conditional paths used in continuous flow matching cancel the path density and provide the velocity directly. For the count process, the reverse rates in equation 4 retain the path $p_\tau(\cdot \mid x_1)$ explicitly. This path remains tractable on the truncated lattice $\mathcal{X} = \{0, \dots, K\}$, with the per-channel cap K set well above the observed counts: the birth–death generator Q_τ is tridiagonal, and the conditional path $p_\tau(\cdot \mid x_1) = [\overrightarrow{\exp} \int_\tau^1 Q_s\, ds\,]_{x_1 \to \cdot}$ is evaluated by finite-state numerical propagation (Appendix E). Substitution into equation 4 gives bridge-matching targets without simulating complete sample trajectories. We train a network $R_\theta^\pm(x, \tau \mid \mathbf{P}, \mathbf{z}_d)$ to predict per-channel generative rates, conditioned on a phoneme stream $\mathbf{P}$ and a day-tuning field $\mathbf{z}_d$ (both defined in Sec. 4.3). With these conditions held fixed, the marginal rates are posterior averages of the endpoint-conditioned bridge rates, by the discrete flow-matching identity (Campbell et al., 2022; Gat et al., 2024) (Appendix A), enabling direct regression under a Bregman divergence suited to non-negative quantities,

$$\mathcal{L}_{\mathrm{CB}} = \mathbb{E}\left[\sum_{t,c}\sum_{\sigma \in \{+,-\}} D\big(R_{\theta,t,c}^\sigma(\mathbf{x}_\tau, \tau \mid \mathbf{P}, \mathbf{z}_d) \,\|\, r_{\tau,t,c}^\sigma\big)\right]. \tag{5}$$

Here $D(\hat{r}\|r) = \hat{r} - r\log\hat{r}$, and $r_{\tau,t,c}^\sigma$ is the endpoint-conditioned target from equation 4. The expectation is over training episodes, $\tau \sim \mathcal{U}(0,1)$, and $\mathbf{x}_\tau \sim p_\tau(\cdot \mid \mathbf{x}_1)$ (Appendix H). The divergence stays finite at $x = 0$, where $r_\tau^-(0) = 0$, which avoids the instability of a Poisson log-likelihood in the sparse-count regime.

4.3 Conditioning on the Label Sequence and the Target-Day Neural State

The rate network $R_\theta^\pm(x, \tau \mid \mathbf{P}, \mathbf{z}_d)$ is conditioned on two streams: a phoneme stream $\mathbf{P} = \Psi(\mathbf{y}) \in \mathbb{R}^{T \times d_p}$ that carries the label sequence, and a day-tuning field $\mathbf{z}_d \in \mathbb{R}^{C \times d_z}$ that carries the target-day neural state. The phoneme conditioner Ψ maps the unaligned phoneme sequence $\mathbf{y}$ to a stream on the bin axis and pads it to the signal length T, so the rate network sees phoneme content and generated counts in one sequence model and learns the phoneme-to-bin correspondence in context, which matches the alignment-free nature of the data and needs no separate duration module or onset predictor. The number of requested phonemes determines the generated sequence length via equation 39. The day conditioner constructs $\mathbf{z}_d$ from the few-shot set $S_{d^\star}$ in two stages: a set encoder Φ that averages count-only features over the prompt trials, respecting electrode geometry through learned positional encodings, produces a per-channel day code $\mathbf{n}_d = \Phi(S_{d^\star})$; a drift head then maps a learned canonical tuning $\mathbf{z}$ to the target day by combining feature-wise modulation with a low-rank cross-channel adapter,

$$\mathbf{z}_d = \underbrace{R_d\,\mathbf{z}}_{\text{cross-channel}} \odot \boldsymbol{\gamma}(\mathbf{n}_d) + \boldsymbol{\beta}(\mathbf{n}_d), \qquad R_d = I + U_d V_d^\top, \tag{6}$$

where $\boldsymbol{\gamma}, \boldsymbol{\beta}$ are per-channel scales and shifts and $U_d, V_d \in \mathbb{R}^{C \times r}$ give cross-channel mixing at a rank $r \ll C$. This combination absorbs per-channel gain change, cross-channel leakage from electrode micro-displacement, and, through the additive shift, can synthesize tuning for electrodes that are silent in the historical data but active on the target day, a capability specific to generation. Padding and duration strategies, the physical interpretation of the drift terms, and the set-encoder architecture appear in Appendix G.2.

4.4 TRAINING OBJECTIVE AND SAMPLING

The bridge loss $\mathcal{L}_{\mathrm{CB}}$ of equation 5 fits the phoneme- and day-conditioned network rates to endpoint-conditioned bridge targets at sampled intermediate states. Four auxiliary terms further enforce semantic and day-state fidelity and discourage prompt-content leakage into the day code. Below, $\hat{\mathbf{x}} \sim p_\theta(\cdot \mid \mathbf{y}, S_d)$ denotes a generated trial and $\tilde{\mathbf{x}}$ its straight-through relaxation, which equals $\hat{\mathbf{x}}$ in the forward pass and carries gradients to θ in the backward pass (Appendix F).

We first introduce a semantic loss that keeps a generated trial decodable as its requested label sequence, paired with a neural-state loss that ties the trial to the statistics of the target day,

$$\mathcal{L}_{\mathrm{sem}} = \mathbb{E}\Big[w_\omega \,\mathrm{CTC}\big(D_\omega(\tilde{\mathbf{x}}),\, \mathbf{y}\big)\Big], \qquad \mathcal{L}_{\mathrm{state}} = \mathbb{E}\Big[\|\tilde{\mathbf{s}} - \mathbf{c}_d\|_2^2 + \eta_{\mathrm{day}}\,\mathrm{CE}\big(C_\psi(\tilde{\mathbf{s}}),\, d\big)\Big], \quad (7)$$

where D_ω is a decoding critic trained under the connectionist-temporal-classification (CTC) loss, H_ψ is a day-state encoder with day classifier C_ψ, $\tilde{\mathbf{s}} = H_\psi(\tilde{\mathbf{x}})$ is the state embedding of the sample, $\mathbf{c}_d$ is the mean state embedding of the few-shot set, and w_ω is a stopped-gradient confidence weight that suppresses samples the critic cannot yet explain.

We further impose a day-contrastive loss so that the day code identifies the day: $\mathcal{L}^{\mathrm{day}}_{\mathrm{InfoNCE}}$ is a standard InfoNCE (van den Oord et al., 2018) that pulls the codes of two disjoint prompt subsets of the same day together against the other days in the batch. The last term guards against content leakage: since $S_{d^\star}$ carries both the target-day neural state and the particular calibration sentences, a code retaining the latter would leak calibration content into samples requested for new sentences, so we add an invariance loss that strips label content from the code. With A_ϕ a CTC probe that reads the features $\mathbf{f}_k$ computed by the set encoder for prompt trial k,

$$\mathcal{L}^{\Phi}_{\mathrm{inv}} = -\min\big(\mathrm{CTC}\big(A_\phi(\mathbf{f}_k),\, \mathbf{y}_k\big),\, m_{\mathrm{inv}}\big)\,, \quad (8)$$

clipped at m_{inv} so that the encoder does not spend capacity on an already-defeated probe.

The full objective combines the four terms with the bridge loss,

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{CB}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{state}}\mathcal{L}_{\mathrm{state}} + \lambda_{\mathrm{day}}\mathcal{L}^{\mathrm{day}}_{\mathrm{InfoNCE}} + \lambda_{\mathrm{inv}}\mathcal{L}^{\Phi}_{\mathrm{inv}}, \quad (9)$$

with non-negative weights λ_*. Training alternates between fitting D_ω, H_ψ with C_ψ, and A_ϕ on real labeled trials and updating θ under equation 9 with these estimates frozen. This schedule prevents the critics from adapting to generator artifacts. The critics are discarded after training; deployment retains the set encoder, the drift and dispersion heads, the phoneme conditioner, and the rate network, so the auxiliary terms add no inference cost.

At deployment, we initialize from the template-conditioned finite CMP prior π_{init} as an approximate source (Appendix H.4) and integrate the rate field under both conditions to generate counts for requested label sequences, including previously unrecorded sentences. The template is retrieved from the historical training trials by phoneme-sequence similarity and supplies only the channel-wise rate buckets for initialization; it requires no additional target-day recordings. Classifier-free guidance (Ho & Salimans, 2021) acts on the jump rates while preserving a valid continuous-time Markov chain (Nisonoff et al., 2024). Training and guidance details are given in Appendices F and H.

5 EXPERIMENTS

We evaluate BURST by the criterion that motivates it: whether counts generated for a target day can substitute for freshly recorded calibration data when a speech decoder is adapted to that day. We report the cross-day substitution result on the T12 dataset at two transfer horizons, validate the fidelity of the generated counts, and ablate the two design choices the substitution rests on; the objective, conditioning, guidance, and sampler ablations are reported in Appendix K.

5.1 EXPERIMENTAL SETUP

Dataset. All experiments use the T12 intracortical speech dataset (Willett et al., 2023): 12,100 attempted-speech sentences recorded over 25 sessions spanning approximately four months from a participant with amyotrophic lateral sclerosis, with neural activity acquired using Utah microelectrode arrays in the ventral motor cortex. We use the count-only representation, retaining all 128 threshold-crossing count channels per 20 millisecond bin and discarding spike-band power, so that the decoder reads exactly the integer process that BURST generates; the effect of this choice on the decoder is quantified in Appendix J.

Table 1: Cross-day substitution PER under the medium-horizon protocol.

Method	Day 10	Day 11	Day 12	Day 13	Day 14	Avg.
No fine-tune (no transfer, upper bd.)	$59.25_{\pm 0.00}$	$59.93_{\pm 0.00}$	$63.60_{\pm 0.00}$	$64.41_{\pm 0.00}$	$66.93_{\pm 0.00}$	$62.82_{\pm 2.93}$
Poisson (NR+NG)	$48.38_{\pm 0.83}$	$47.18_{\pm 0.49}$	$52.20_{\pm 0.90}$	$51.14_{\pm 0.45}$	$49.00_{\pm 1.00}$	$49.58_{\pm 2.00}$
Negative binomial (NR+NG)	$49.08_{\pm 0.68}$	$47.85_{\pm 0.61}$	$52.84_{\pm 0.85}$	$51.75_{\pm 0.48}$	$49.77_{\pm 1.22}$	$50.26_{\pm 1.98}$
Gaussian latent (NR+NG)	$49.77_{\pm 0.87}$	$47.88_{\pm 0.64}$	$53.66_{\pm 0.81}$	$52.12_{\pm 0.48}$	$50.06_{\pm 1.14}$	$50.70_{\pm 2.17}$
Count-FM (NR+NG)	$47.36_{\pm 0.81}$	$46.13_{\pm 0.64}$	$50.56_{\pm 0.92}$	$49.87_{\pm 0.51}$	$47.17_{\pm 1.20}$	$48.22_{\pm 1.90}$
BURST (NR+NG)	$44.57_{\pm 0.74}$	$43.00_{\pm 0.57}$	$48.37_{\pm 0.85}$	$46.89_{\pm 0.45}$	$45.12_{\pm 1.14}$	$\mathbf{45.59}_{\pm 2.03}$
NR-alone (N real)	$46.90_{\pm 0.67}$	$45.50_{\pm 0.59}$	$50.63_{\pm 0.88}$	$49.34_{\pm 0.45}$	$47.51_{\pm 1.24}$	$47.98_{\pm 1.99}$
2NR ($2N$ real, anchor)	$43.78_{\pm 0.80}$	$42.04_{\pm 0.78}$	$47.54_{\pm 0.90}$	$45.46_{\pm 0.37}$	$43.94_{\pm 1.22}$	$44.55_{\pm 2.04}$
All-real FT (all sentences, lower bd.)	$41.86_{\pm 0.67}$	$40.48_{\pm 0.64}$	$45.52_{\pm 0.81}$	$44.41_{\pm 0.45}$	$42.42_{\pm 1.16}$	$42.94_{\pm 1.98}$
Δ_{sub} (BURST−2NR)	$+0.79_{\pm 0.24}$	$+0.96_{\pm 0.34}$	$+0.83_{\pm 0.25}$	$+1.43_{\pm 0.33}$	$+1.18_{\pm 0.16}$	$+1.04_{\pm 0.35}$

Table 2: Cross-day substitution PER under the long-horizon protocol.

Method	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20	Day 21	Avg.
No fine-tune (no transfer, upper bd.)	$47.01_{\pm 0.00}$	$54.69_{\pm 0.00}$	$52.59_{\pm 0.00}$	$54.10_{\pm 0.00}$	$52.11_{\pm 0.00}$	$54.39_{\pm 0.00}$	$56.18_{\pm 0.00}$	$53.01_{\pm 2.79}$
Poisson (NR+NG)	$35.97_{\pm 0.44}$	$36.98_{\pm 0.68}$	$37.49_{\pm 0.38}$	$39.98_{\pm 1.01}$	$38.00_{\pm 0.69}$	$39.08_{\pm 0.37}$	$35.43_{\pm 0.79}$	$37.56_{\pm 1.64}$
Negative binomial (NR+NG)	$36.62_{\pm 0.43}$	$37.62_{\pm 0.65}$	$38.48_{\pm 0.42}$	$40.64_{\pm 1.07}$	$38.60_{\pm 0.54}$	$39.83_{\pm 0.45}$	$35.91_{\pm 0.78}$	$38.24_{\pm 1.69}$
Gaussian latent (NR+NG)	$37.16_{\pm 0.56}$	$38.36_{\pm 0.79}$	$38.88_{\pm 0.49}$	$41.17_{\pm 0.97}$	$39.14_{\pm 0.57}$	$40.45_{\pm 0.35}$	$36.37_{\pm 0.79}$	$38.79_{\pm 1.71}$
Count-FM (NR+NG)	$34.58_{\pm 0.52}$	$35.62_{\pm 0.76}$	$36.75_{\pm 0.47}$	$38.94_{\pm 1.03}$	$37.02_{\pm 0.63}$	$37.96_{\pm 0.39}$	$34.43_{\pm 0.82}$	$36.47_{\pm 1.71}$
BURST (NR+NG)	$\mathbf{32.20}_{\pm 0.48}$	$\mathbf{33.07}_{\pm 0.72}$	$\mathbf{33.75}_{\pm 0.44}$	$\mathbf{35.96}_{\pm 0.99}$	$\mathbf{34.13}_{\pm 0.59}$	$\mathbf{35.26}_{\pm 0.36}$	$\mathbf{31.58}_{\pm 0.78}$	$\mathbf{33.71}_{\pm 1.59}$
NR-alone (N real)	$34.59_{\pm 0.54}$	$35.55_{\pm 0.84}$	$36.10_{\pm 0.46}$	$38.52_{\pm 1.08}$	$36.56_{\pm 0.72}$	$37.71_{\pm 0.34}$	$33.98_{\pm 0.84}$	$36.14_{\pm 1.66}$
2NR ($2N$ real, anchor)	$31.41_{\pm 0.26}$	$32.03_{\pm 0.69}$	$32.89_{\pm 0.36}$	$34.95_{\pm 0.95}$	$33.25_{\pm 0.64}$	$34.24_{\pm 0.59}$	$30.81_{\pm 0.60}$	$32.80_{\pm 1.51}$
All-real FT (all sentences, lower bd.)	$29.66_{\pm 0.55}$	$30.47_{\pm 0.69}$	$31.32_{\pm 0.51}$	$33.30_{\pm 0.94}$	$31.57_{\pm 0.55}$	$32.65_{\pm 0.40}$	$29.15_{\pm 0.80}$	$31.16_{\pm 1.54}$
Δ_{sub} (BURST−2NR)	$+0.79_{\pm 0.30}$	$+1.04_{\pm 0.08}$	$+0.86_{\pm 0.36}$	$+1.01_{\pm 0.25}$	$+0.88_{\pm 0.40}$	$+1.02_{\pm 0.37}$	$+0.77_{\pm 0.24}$	$+0.91_{\pm 0.30}$

Cross-day transfer protocols. We evaluate BURST under two protocols that differ in the length of the historical window and in its distance from the target deployment day. The medium-horizon protocol takes days 1 through 7 as the historical window and days 10 through 14 as targets; the long-horizon protocol takes days 1 through 12 and targets days 15 through 21. Recalibration therefore occurs three to seven days after the window and three to nine days, respectively. Both protocols share the same data discipline: the historical decoder and the generator, including its day conditioner, see all sentences of the historical window and nothing from the target day; the calibration sentences of a target day are drawn from the training split of that day; and every reported score is computed on the disjoint test split. The metric throughout is the phoneme error rate (PER), with lower values better.

Implementation details. We use a compact three-block Conformer–CTC decoder for all substitution arms and a six-layer Transformer rate network with hidden width 256 for BURST. Counts are modeled on the finite lattice $\{0, \dots, 16\}$, and the historical Fano-rate profile is discretized into eight rate buckets with the day displacement bounded by $\delta_{\max} = 0.35$. We train the generator for 17,000 AdamW updates with a peak learning rate of 2×10^{-4} and sample generated counts using 50-step tau-leaping with guidance weights $(w_p, w_d, w_{\text{int}}) = (1.5, 1.5, 0)$. All calibrated arms share the same two-stage decoder fine-tuning schedule of 20,000 updates with batch size 16 and identical input augmentation. We report the mean and standard deviation of target-day test PER over five random seeds at each target day and transfer horizon. Full network architectures, auxiliary-loss weights, dispersion calibration, and remaining training, sampling, and adaptation settings are provided in Appendices G, E, and H.

Baselines. The real-data references are NO FT, with no calibration; NR-ALONE, fine-tuned on N real sentences; 2NR, fine-tuned on $2N$ real sentences; and ALL-REAL FT, fine-tuned on all available target-day data. Since no directly comparable baseline exists, we construct matched count-generation baselines inspired by alternative generative formulations. Count-FM (Wei & Pearson, 2026) uses its binomial endpoint conditional path together with BURST's phoneme conditioner, day encoder Φ, guidance, and fine-tuning schedule. The other baselines use a unit-Fano Poisson formulation (Santos et al., 2023; Campbell et al., 2022; Gat et al., 2024), an overdispersed negative-binomial formulation (Lopez et al., 2018), or a rounded Gaussian-latent formulation (Wen et al., 2023; Kapoor et al., 2024). All five generator arms use the same NR+NG protocol: N real sentences plus N generated sentences, with the same target-day recording budget and fine-tuning schedule. For each target day, the 2NR reference uses a fixed set of $2N$ recorded trials, while NR-ALONE and the real half of NR+NG use the same N-trial subset, which also contains the shared few-shot set $S_{d^\star}$. The generated half uses only the phoneme sequences of the remaining N sentence prompts in the 2NR set; their neural recordings are not accessed. Thus, NR-ALONE and NR+NG both use N target-day neural recordings, whereas 2NR uses $2N$. We report the seed-paired substitution gap

Table 3: Ablation on conditional fidelity. Ins/Del/Sub decompose the edits.

Condition	PER	Δ_{gen}	Ins	Del	Sub
Real (reference)	$47.93_{\pm 0.28}$	–	5.34	6.42	36.17
Generated, correct	$50.10_{\pm 0.31}$	$+2.17_{\pm 0.26}$	5.71	6.88	37.51
shuffled phonemes	$89.37_{\pm 0.63}$	$+41.44_{\pm 0.58}$	19.76	11.24	58.37
mismatched day	$63.70_{\pm 0.45}$	$+15.77_{\pm 0.41}$	8.13	12.38	43.19

$\Delta_{\text{sub}} = \text{PER}(\text{method}) - \text{PER}(2\text{NR})$, where zero denotes matching the twice-budget real-data reference.

5.2 Main Results

Tables 1 and 2 report the target-day PER for cross-day substitution under a newly recorded-data budget of $N{=}100$, evaluated at both transfer horizons. We highlight three findings. **First, BURST improves real-only calibration at the same recording budget and approaches the $2N$-real reference.** The BURST mixture is within 1.04 PER points of the $2N$-real reference at the medium horizon and within 0.91 points at the long horizon. Augmenting the N real samples with N BURST-generated samples reduces PER by 2.39 and 2.43 points relative to fine-tuning on the N real samples alone, respectively. Equivalently, the generated data recover approximately 70% and 73% of the improvement obtained by collecting the second N real sentences at the medium and long horizons, respectively.

Second, the substitution result depends on the count-generation formulation. Count-FM provides the closest path-level comparison: with the conditioning and adaptation protocol matched, its average PER is higher than NR-alone by 0.24 and 0.33 points at the two horizons, while BURST reduces PER relative to Count-FM by 2.63 and 2.76 points, respectively. The Poisson, negative-binomial, and Gaussian-latent generators underperform NR-alone by 1.42 to 2.72 PER points. Under the shared substitution protocol, BURST performs better than all evaluated alternative generation formulations.

Third, BURST's advantage is consistent across target days. BURST outperforms every same-recording-budget comparator on every target day at both horizons, with a gap of 0.77 to 1.43 PER points to the $2N$-real reference. The relative ordering of Count-FM and NR-alone varies across days. At the long horizon, BURST retains a similar gap to the $2N$-real reference even though absolute PER is more than ten points lower than at the medium horizon, indicating that overall target-day difficulty alone cannot account for the remaining difference.

5.3 Ablation studies

To establish that the substitution rests on the design choices we claim it does, we examine three questions: (i) whether the generated counts reproduce the targeted dispersion and preserve phoneme- and day-specific conditioning, (ii) how much rate-dependent dispersion contributes to substitution, and (iii) whether the substitution gap changes with the calibration budget. We then provide a compact ablation of the day-conditioning architecture. Additional ablations of the training objective, guidance, and tau-leaping discretization are reported in Appendices K.1, K.2, and K.3, respectively. Unless stated otherwise, all experiments use the medium-horizon protocol with $N{=}100$ (Sec. 5.1).

Do generated counts preserve the targeted dispersion and conditioning? At high firing rates, the generated counts have a per-channel Fano factor of 0.867 ± 0.003, close to the empirical value of 0.872 (Figure A1 in Appendix C and Table 4). The Poisson baseline ($\nu \equiv 1$) produces near-unit Fano factors and misses this sub-Poisson regime (Table 4). Because the statistic is measured after guided tau-leaping, the agreement reflects the complete calibrated sampler, including the sampling procedure.

Table 4: Dispersion control (days 10–14). Fano is the high-rate per-channel median; Fano err. is the L_1 distance to the real curve (App. C).

Schedule	Fano	Fano err.	Δ_{sub}
$\nu(r)$ (**BURST**)	0.867	0.012	$+\mathbf{1.04}_{\pm 0.35}$
$\bar{\nu}$ constant	0.903	0.058	$+1.31_{\pm 0.38}$
$\delta_d{=}0$	0.881	0.039	$+1.67_{\pm 0.37}$
$\nu{\equiv}1$	0.994	0.121	$+5.03_{\pm 0.40}$

Evaluating conditioning with the substitution decoder could confound fidelity with adaptation to generator artifacts. We therefore use an independently initialized Conformer–CTC evaluator trained

Table 5: Day-conditioning ablation under the medium-horizon protocol at N=100.

Conditioning variant	Indep. PER	Δ_{sub}
Full drift (BURST; rank-8 adapter and affine)	$50.10_{\pm 0.31}$	$+1.04_{\pm 0.35}$
No adapter ($R_d = I$; affine retained)	$53.08_{\pm 0.39}$	$+2.83_{\pm 0.52}$
Full-rank transform (affine retained)	$50.27_{\pm 0.32}$	$+1.19_{\pm 0.37}$
No affine ($\boldsymbol{\gamma} = 1, \boldsymbol{\beta} = 0$; adapter retained)	$51.18_{\pm 0.35}$	$+1.70_{\pm 0.42}$

only on real target-day sentences and never exposed to generated samples or generator optimization. Under correct conditioning, generated sequences yield 50.10 PER, 2.17 points above the real target-day reference, with errors distributed across all three edit types. Permuting the phoneme order while preserving sequence length and inventory increases PER to 89.37; replacing the target-day neural-state prompt with that of historical day 7 increases it to 63.70 and approximately doubles the deletion rate. These controls indicate that the generated counts preserve both phoneme order and information specific to the requested day.

Does rate-dependent dispersion contribute? Table 4 compares the day-adjusted initialization profile $\nu_d(r)$ with a constant $\bar{\nu}_d$ calibrated to the aggregate high-rate Fano target and the Poisson initialization control ($\nu \equiv 1$), holding the trained rate field and all other learned components fixed. The integrated error relative to the empirical Fano–rate curve increases from 0.012 for $\nu_d(r)$ to 0.058 for $\bar{\nu}_d$ and 0.121 for Poisson. The corresponding substitution gaps are $+1.04$, $+1.31$, and $+5.03$ PER points, with the last obtained from $49.58 - 44.55$. Disabling the per-day adjustment δ_d while retaining the pooled rate-dependent schedule increases the substitution gap by 0.63 PER points, confirming that target-day dispersion adaptation contributes beyond the base profile. Relative to BURST, the constant-exponent and Poisson controls increase the mean gap by 0.27 and 3.99 points, respectively. Because these controls change both the prior mean and its dispersion, the PER differences quantify their joint effect.

Does the day-conditioning architecture contribute? The drift head uses low-rank mixing to model correlated changes across electrodes and affine modulation to account for channel-specific changes. Table 5 evaluates the two components separately. Removing the cross-channel adapter increases independent-decoder PER by 2.98 points and the substitution gap by 1.79 points. Removing affine modulation gives smaller increases of 1.08 and 0.66 points, respectively. Replacing the rank-8 adapter with a full-rank transform changes performance only slightly, suggesting that the low-rank adapter is sufficient to model cross-channel drift in this setting.

Does the substitution gap shrink with the calibration budget? We vary the calibration budget over $N \in \{50, 100, 200\}$ and compare BURST with the corresponding $2N$-real reference (Table 6). The substitution gap remains close to one PER point, $+0.94$, $+0.86$, and $+1.11$, respectively, with no monotonic trend. Within the evaluated range, increasing the calibration budget improves absolute PER while leaving the gap to $2N$ real sentences largely unchanged. This pattern argues against small-sample calibration as the sole source of the residual gap.

Table 6: Substitution across calibration budgets, pooled over days 10–12. Δ_{sub} is positive for every seed.

N	2NR	BURST	Δ_{sub}
50	47.02	47.96	$+0.94_{\pm 0.35}$
100	44.45	45.32	$+0.86_{\pm 0.28}$
200	43.83	44.94	$+1.11_{\pm 0.32}$

6 Conclusion

We presented BURST, a phoneme-conditioned spike-count generator for data-efficient cross-day neuroprosthesis calibration. Its central innovation is a generator-first birth–death bridge that makes rate- and target-day-dependent dispersion an explicit part of the generative process, while an alignment-free phoneme conditioner and a few-shot day encoder jointly specify linguistic content and target-day state. Across two transfer horizons, an equal mixture of real and generated sentences consistently outperformed real-only calibration at the same recording budget and approached calibration with twice as many real sentences, while reproducing empirical high-rate sub-Poisson variability. BURST retains a small target-day recording set and reduces the repeated calibration burden in long-term neuroprostheses.

AI USE STATEMENT

Generative AI tools were used solely to improve the clarity and grammar of author-written text. They were not used for research ideation, method development, code generation, experiments, data analysis, figure creation, or scientific conclusions. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This study uses only publicly available, de-identified intracortical speech data released by prior work. We conducted no new participant recruitment, intervention, or neural data collection. The original study documents ethical approval and informed consent, and we used the data solely for research under its stated access conditions.

REPRODUCIBILITY STATEMENT

All data used in this study are publicly available. The main text and appendix document data processing, the model and objectives, calibration, architectures, training and sampling, evaluation protocols, and proofs. Complete code and configuration files will be released publicly upon acceptance.

REFERENCES

William J Anderson. *Continuous-Time Markov Chains: An Applications-Oriented Approach*. Springer, 1991.

Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019. doi: 10.1038/s41586-019-1119-1.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, 2022.

Nicholas S. Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R. Deo, Aparna Srinivasan, Eun Young Choi, Matthew F. Glasser, Leigh R. Hochberg, Jaimie M. Henderson, Kiarash Shahlaie, Sergey D. Stavisky, and David M. Brandman. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024. doi: 10.1056/NEJMoa2314132.

Cynthia A. Chestek, Vikash Gilja, Paul Nuyujukian, Justin D. Foster, Joline M. Fan, Matthew T. Kaufman, Mark M. Churchland, Zuley Rivera-Alvidrez, John P. Cunningham, Stephen I. Ryu, and Krishna V. Shenoy. Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. *Journal of Neural Engineering*, 8(4):045005, 2011. doi: 10.1088/1741-2560/8/4/045005.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 2023. doi: 10.1038/s42256-023-00714-5.

Ebrahim Feghhi, Shreyas Kaasyap, Nima Hadidi, and Jonathan C. Kao. Time-masked transformers with lightweight test-time adaptation for neural speech decoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 103247–103272, 2025. doi: 10.52202/085713-3112. URL https://arxiv.org/abs/2507.02800.

Nic Fishman, Gokul Gowri, Tanush Kumar, Jiaqi Lu, Valentin de Bortoli, Jonathan S. Gootenberg, and Omar Abudayyeh. Count bridges enable modeling and deconvolving transcriptomic data. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d2562289abd8bd3ee67b24be367794f1-Paper-Conference.pdf.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T Q Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems*, 2024.

Robert E Gaunt, Satish Iyengar, Adri B Olde Daalhuis, and Burcin Simsek. An asymptotic expansion for the normalizing constant of the conway–maxwell–poisson distribution. *Annals of the Institute of Statistical Mathematics*, 71(1):163–180, 2019.

Carl Gold, Darrell A. Henze, Christof Koch, and György Buzsáki. On the origin of the extracellular action potential waveform: A modeling study. *Journal of Neurophysiology*, 95(5):3113–3128, 2006. doi: 10.1152/jn.00979.2005. URL https://doi.org/10.1152/jn.00979.2005.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pp. 5036–5040, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jaivardhan Kapoor, Auguste Schulz, Julius Vetter, Felix Pei, Richard Gao, and Jakob H. Macke. Latent diffusion for neural spiking data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 118119–118154, 2024. doi: 10.52202/079017-3751.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018.

Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023. doi: 10.1038/s41586-023-06443-4.

Manuel Molano-Mazón, Arno Onken, Eugenio Piasini, and Stefano Panzeri. Synthesizing realistic neural population activity patterns using generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1VVsebAZ.

David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, Pengfei F. Sun, Josh Chartier, Maximilian E. Dougherty, Patricia M. Liu, Gary M. Abrams, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021. doi: 10.1056/NEJMoa2027540.

Emily M. Mugler, James L. Patton, Robert D. Flint, Zachary A. Wright, Stephan U. Schuele, Joshua Rosenow, Jerry J. Shih, Dean J. Krusienski, and Marc W. Slutzky. Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of Neural Engineering*, 11(3):035015, 2014. doi: 10.1088/1741-2560/11/3/035015.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. *arXiv preprint arXiv:2406.01572*, 2024.

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(10):805–815, 2018. doi: 10.1038/s41592-018-0109-9.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

János A. Perge, Mark L. Homer, Wasim Q. Malik, Sydney Cash, Emad Eskandar, Gerhard Friehs, John P. Donoghue, and Leigh R. Hochberg. Intra-day signal instabilities affect decoding performance in an intracortical neural interface system. *Journal of Neural Engineering*, 10(3):036004, 2013. doi: 10.1088/1741-2560/10/3/036004.

Javier E Santos, Zachary R Fox, Nicholas Lubbers, and Yen Ting Lin. Blackout diffusion: Generative diffusion models in discrete-state spaces. In *International Conference on Machine Learning*, 2023.

Galit Shmueli, Thomas P. Minka, Joseph B. Kadane, Sharad Borle, and Peter Boatwright. A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142, 2005. doi: 10.1111/j.1467-9876.2005.00474.x.

Ian H. Stevenson. Flexible models for spike count data with both over- and under-dispersion. *Journal of Computational Neuroscience*, 41(1):29–43, 2016. doi: 10.1007/s10827-016-0603-y.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023. doi: 10.1038/s41593-023-01304-9.

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

Maitreyee Wairagkar, Nicholas S. Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M. Miller, Leigh R. Hochberg, David M. Brandman, and Sergey D. Stavisky. An instantaneous voice-synthesis neuroprosthesis. *Nature*, 644(8075):145–152, 2025. doi: 10.1038/s41586-025-09127-3.

Ganchao Wei and John Pearson. Flow matching for count data. *arXiv preprint arXiv:2605.07746*, 2026. doi: 10.48550/arXiv.2605.07746. URL https://arxiv.org/abs/2605.07746.

Ganchao Wei and Ian H. Stevenson. Dynamic modeling of spike count data with Conway–Maxwell–Poisson variability. *Neural Computation*, 35(7):1187–1208, 2023. doi: 10.1162/neco_a_01593.

Shixian Wen, Allen Yin, Tommaso Furlanello, et al. Rapid adaptation of brain–computer interfaces to new neuronal ensembles or participants via generative modelling. *Nature Biomedical Engineering*, 7:546–558, 2023.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620 (7976):1031–1036, 2023. doi: 10.1038/s41586-023-06377-x.

Zhanqi Zhang, Shun Li, Bernardo L Sabatini, Mikio Aoi, and Gal Mishne. Align: Adversarial learning for generalizable speech neuroprosthesis. *arXiv preprint arXiv:2603.18299*, 2026.

A Background: Probability Paths for Discrete Flow Matching

We recall only the discrete flow-matching identity needed to justify equation 5 (Campbell et al., 2022; Gat et al., 2024). Fix a signal length T, let C denote the number of count channels modeled by the generator, and write $D = T\,C$. A complete signal is a state $\mathbf{x} \in \mathcal{X}^D$, where $\mathcal{X} = \{0, 1, \ldots, K\}$. For variable-length trials, the argument applies conditionally on T. The label sequence, target-day neural state, and stopped day adjustment are held fixed and omitted from the notation.

For a datum $\mathbf{x}_1 \sim q$, let $p_\tau(\cdot \mid \mathbf{x}_1)$ be a conditional probability path. The path factorizes over coordinates, while its marginal mixes over the data distribution:

$$p_\tau(\mathbf{x}) = \sum_{\mathbf{x}_1} p_\tau(\mathbf{x} \mid \mathbf{x}_1)\, q(\mathbf{x}_1), \qquad p_\tau(\mathbf{x} \mid \mathbf{x}_1) = \prod_{i=1}^{D} p_{\tau,i}(x^i \mid \mathbf{x}_1). \tag{10}$$

The factors may depend on endpoint-derived side information, including the rate bucket that remains fixed along the path. At the data endpoint, $p_1(\cdot \mid \mathbf{x}_1) = \delta_{\mathbf{x}_1}$ and hence $p_1 = q$. At the other endpoint, the induced source is $p_{\mathrm{src}}(\mathbf{x}) = \sum_{\mathbf{x}_1} p_0(\mathbf{x} \mid \mathbf{x}_1) q(\mathbf{x}_1)$. A common source p additionally requires $p_0(\cdot \mid \mathbf{x}_1) = p$ for every $\mathbf{x}_1$; the frozen-rate stationary law of Section 4.1 supplies no such condition. An exact sampler for the stated marginal path must therefore be initialized from p_{src}. Practical generation uses the approximate prior π_{init} defined in Appendix H.4.

Let $u_\tau(\mathbf{x}, \mathbf{z} \mid \mathbf{x}_1)$ denote a conditional generator entry, with destination $\mathbf{x}$ and origin $\mathbf{z}$. The following standard identity converts these tractable conditional entries into the generator of the marginal path.

Proposition 1 (Marginalization of conditional rates)**.** *For $\tau \in (0, 1)$ and $p_\tau(\mathbf{z}) > 0$, define*

$$u_\tau(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{x}_1} u_\tau(\mathbf{x}, \mathbf{z} \mid \mathbf{x}_1)\, p_{1|\tau}(\mathbf{x}_1 \mid \mathbf{z}), \qquad p_{1|\tau}(\mathbf{x}_1 \mid \mathbf{z}) = \frac{p_\tau(\mathbf{z} \mid \mathbf{x}_1) q(\mathbf{x}_1)}{p_\tau(\mathbf{z})}. \tag{11}$$

Completing the columns associated with zero-mass origins yields a valid generator of the marginal path p_τ. Moreover, for each off-diagonal entry, the scalar loss $D(\hat{r} \| r) = \hat{r} - r \log \hat{r}$ in equation 5 has $u_\tau(\mathbf{x}, \mathbf{z})$ as its population minimizer after the continuous extension $D(0 \| 0) = 0$.

The proof is given in Appendix B. In BURST, equation 4 supplies the nonzero single-coordinate conditional entries, with rates of jumps outside $\mathcal{X}$ set to zero. Although the conditional paths factorize, the posterior average in equation 11 can retain temporal and cross-channel dependence. The proposition justifies the unguided rate-matching objective; sampling-time guidance modifies the rate field and need not preserve the same marginal path.

B Proofs

B.1 Marginalization of conditional rates

Proof of Proposition 1. Each conditional generator satisfies

$$\frac{\partial p_\tau(\mathbf{x} \mid \mathbf{x}_1)}{\partial \tau} = \sum_{\mathbf{z}} u_\tau(\mathbf{x}, \mathbf{z} \mid \mathbf{x}_1) p_\tau(\mathbf{z} \mid \mathbf{x}_1), \qquad \sum_{\mathbf{x}} u_\tau(\mathbf{x}, \mathbf{z} \mid \mathbf{x}_1) = 0. \tag{12}$$

For $p_\tau(\mathbf{z}) > 0$, equation 11 is a posterior average. It therefore preserves non-negativity of the off-diagonal entries and the zero column-sum condition. If $p_\tau(\mathbf{z}) = 0$, non-negativity of the mixture in equation 10 implies $p_\tau(\mathbf{z} \mid \mathbf{x}_1) = 0$ for q-almost every $\mathbf{x}_1$, so that column does not contribute to the continuity equation and can be completed by any valid generator column. For the remaining columns,

$$\sum_{\mathbf{z}} u_\tau(\mathbf{x}, \mathbf{z}) p_\tau(\mathbf{z}) = \sum_{\mathbf{z}} \sum_{\mathbf{x}_1} u_\tau(\mathbf{x}, \mathbf{z} \mid \mathbf{x}_1) p_\tau(\mathbf{z} \mid \mathbf{x}_1) q(\mathbf{x}_1) \tag{13}$$

$$= \sum_{\mathbf{x}_1} q(\mathbf{x}_1) \frac{\partial p_\tau(\mathbf{x} \mid \mathbf{x}_1)}{\partial \tau} = \frac{\partial p_\tau(\mathbf{x})}{\partial \tau}, \tag{14}$$

where equation 12 and equation 10 give the second and third equalities. Hence, u_τ generates the marginal path.

For a fixed off-diagonal entry, write $r(\mathbf{x}_1) = u_\tau(\mathbf{x}, \mathbf{z} \mid \mathbf{x}_1)$ and $\bar{r} = u_\tau(\mathbf{x}, \mathbf{z})$. Conditional on the current state $\mathbf{z}$, the endpoint follows $p_{1|\tau}(\cdot \mid \mathbf{z})$, and therefore

$$\mathbb{E}[D(\hat{r}\|r) \mid \mathbf{z}] = \hat{r} - \bar{r}\log\hat{r}, \qquad \frac{\partial}{\partial\hat{r}}\mathbb{E}[D(\hat{r}\|r) \mid \mathbf{z}] = 1 - \frac{\bar{r}}{\hat{r}}. \tag{15}$$

The unique minimizer is $\hat{r} = \bar{r}$ when $\bar{r} > 0$. When $\bar{r} = 0$, the infimum is approached as $\hat{r} \downarrow 0$ and is attained under the continuous extension $D(0\|0) = 0$. The generalized Kullback–Leibler divergence differs from this scalar loss only by the target-dependent term $r\log r - r$. Thus, the regression in equation 5 has the marginal rate as its population minimizer. □

B.2 Dispersion control

Proof. Let X follow equation 2 on $\mathbb{Z}_{\geq 0}$, and write $p(x) \propto \mu^x/(x!)^\nu$ with $\mu = \lambda_\tau/\kappa_\tau > 0$ and $\nu = \nu_\tau > 0$. All moments are finite because $(x!)^\nu$ grows faster than any geometric sequence.

For clause (i), define $g(x) = (x+1)p(x+1)/p(x) = \mu(x+1)^{1-\nu}$. Shifting the summation index gives

$$\mathbb{E}[X] = \mathbb{E}[g(X)], \qquad \mathbb{E}[X^2] = \mathbb{E}[(X+1)g(X)], \qquad \mathrm{Var}(X) = \mathbb{E}[X] + \mathrm{Cov}(X, g(X)). \tag{16}$$

When $\nu > 1$, g is strictly decreasing, so $\mathrm{Cov}(X, g(X)) < 0$ and $\mathrm{Fano} < 1$. The inequality is strict because the distribution assigns positive mass to more than one state. When $\nu < 1$, g is strictly increasing, which gives $\mathrm{Cov}(X, g(X)) > 0$ and $\mathrm{Fano} > 1$. When $\nu = 1$, $g \equiv \mu$, and equation 2 is the Poisson distribution with $\mathrm{Fano} = 1$.

For clause (ii), let $Z(\mu, \nu) = \sum_{x\geq 0}\mu^x/(x!)^\nu$. The identities $\mathbb{E}[X] = \partial\log Z/\partial\log\mu$ and $\mathrm{Var}(X) = \partial^2\log Z/\partial(\log\mu)^2$, together with the asymptotic expansion of the Conway–Maxwell–Poisson normalizer (Shmueli et al., 2005; Gaunt et al., 2019), give

$$\mathbb{E}[X] = \mu^{1/\nu} + O(1), \qquad \mathrm{Var}(X) = \frac{\mu^{1/\nu}}{\nu} + O(1). \tag{17}$$

Therefore, $\mathrm{Fano} = 1/\nu + O(\mu^{-1/\nu}) \to 1/\nu$.

The exact statements above concern the unbounded law on $\mathbb{Z}_{\geq 0}$. On a capped lattice $\{0, \ldots, K\}$ with $K \geq 1$, the same covariance argument gives $\mathrm{Fano} < 1$ for every $\nu \geq 1$, including $\nu = 1$, whereas the dispersion sign for $\nu < 1$ depends on the finite lattice. Clause (ii)'s high-count limit is also inapplicable at fixed K. The untruncated conclusions are recovered as $K \to \infty$ at fixed (μ, ν), so the finite-lattice effect must be assessed separately through the choice of cap. □

B.3 Reverse rates

Proof. Fix x_1, abbreviate $p_\tau(y) = p_\tau(y \mid x_1)$, and work on $\mathcal{X} = \{0, \ldots, K\}$. The capped generator sets the death rate at 0 and the birth rate at K to zero. With the endpoint-derived schedule held fixed, write $Q_\tau(z \to y)$ for its jump rates. In decreasing flow time, the conditional path satisfies

$$-\frac{\partial p_\tau(y)}{\partial\tau} = \sum_{z\neq y}\left[Q_\tau(z \to y)p_\tau(z) - Q_\tau(y \to z)p_\tau(y)\right]. \tag{18}$$

For $\tau \in (0, 1)$ and $p_\tau(y) > 0$, define

$$\rho_\tau(y \to z \mid x_1) = Q_\tau(z \to y)\frac{p_\tau(z)}{p_\tau(y)}, \qquad z \neq y. \tag{19}$$

Substitution into the forward equation for a process run in increasing τ gives

$$\sum_{z\neq y}\left[\rho_\tau(z \to y \mid x_1)p_\tau(z) - \rho_\tau(y \to z \mid x_1)p_\tau(y)\right] = \sum_{z\neq y}\left[Q_\tau(y \to z)p_\tau(y) - Q_\tau(z \to y)p_\tau(z)\right] \tag{20}$$

$$= \frac{\partial p_\tau(y)}{\partial\tau}, \tag{21}$$

where the last equality follows from equation 18. Thus, p_τ satisfies the reverse-time Kolmogorov equation. For every $\varepsilon \in (0,1)$, irreducibility of the capped birth–death chain gives $p_\tau(y) > 0$ and locally bounded reverse rates on $[0, 1-\varepsilon]$. A reverse process initialized from $p_0(\cdot \mid x_1)$ therefore has marginals p_τ on this interval by uniqueness of the finite-dimensional linear system. Since ε is arbitrary, the equality holds for every $\tau < 1$, while $p_\tau \to \delta_{x_1}$ as $\tau \uparrow 1$.

The only nonzero rates into y are the birth from $y-1$ and the death from $y+1$. Hence,

$$\begin{aligned} \rho_\tau(y \to y-1 \mid x_1) &= b_\tau(y-1)\frac{p_\tau(y-1 \mid x_1)}{p_\tau(y \mid x_1)}, \quad 1 \le y \le K, \\ \rho_\tau(y \to y+1 \mid x_1) &= d_\tau(y+1)\frac{p_\tau(y+1 \mid x_1)}{p_\tau(y \mid x_1)}, \quad 0 \le y < K. \end{aligned} \tag{22}$$

These are r_τ^- and r_τ^+ in equation 4, with $r_\tau^-(0 \mid x_1) = 0$ and $r_\tau^+(K \mid x_1) = 0$. This is the standard time reversal of a continuous-time Markov chain (Anderson, 1991). □

C Rate-Dependent Dispersion Estimation

Figure 1 is estimated from the 128 threshold-crossing channels at the native 20-millisecond bin width. Because phoneme labels are not aligned to individual bins, the local rate proxy for bin t and channel c excludes the center observation,

$$\hat{\lambda}_{t,c} = \frac{1}{2r} \sum_{0<|j|\le r} x_{t+j,c}, \qquad r = \lfloor W/2 \rfloor. \tag{23}$$

The reported curve uses $W = 9$ to balance the finite-sample bias at small windows against the smoothing of within-phoneme rate variation at large windows, and groups bins within each channel into bands of width 0.1 counts per bin before taking the median Fano factor across channels. Center exclusion prevents circular bin assignment; temporal dependence and within-band rate variation remain, so the estimate relies on local stationarity and is interpreted as a coarse dispersion target. It decreases from approximately one at low rates to approximately 0.87 near 120 spikes per second.

Both figures of this appendix plot the same quantity, so we state the shared display conventions once. At the native 20-millisecond bin width a band of 0.1 counts per bin corresponds to 5 spikes per second, and every marker is placed at the midpoint of its band. Rates are reported in spikes per second throughout, while the estimator of equation 23 and the banding operate on counts per bin. In Figure 1 the line is the per-channel median within each band and the shaded region is the interquartile range across the 128 channels, pooled over the sessions of the historical window; occupied bands run from 2.5 to 117.5 spikes per second, so the curve terminates below the upper limit of the $[0, 150]$ calibration range of Appendix E, where the two highest buckets are extrapolated rather than measured. The high-rate Fano factor quoted in the text and in Table 4 is the value of this median curve in the highest occupied band.

Figure A1 compares the Fano factor of BURST-generated counts with the empirical T12 curve. Generated counts are drawn from the medium-horizon protocol using the default guidance weights and sampled over 200 prompts per target day. The generated distribution closely tracks the empirical curve across the full firing-rate range, reproducing the measured sub-Poisson behavior at high rates. The two curves are estimated with the identical procedure and are shown on the same 5 spikes per second grid; the legend entry *T12 raw target* denotes the raw, uncorrected estimator of equation 23 applied to the target-day recordings, matching Figure 1. The shaded region around the BURST curve is the spread over the five sampling seeds and is narrow enough to be comparable to the line width above roughly 60 spikes per second; the empirical curve is a single fixed estimate from the recordings and therefore carries no such region.

D Effective Exclusion-Time Model and Cross-Day Variation

The decreasing curve is consistent with a stationary renewal model that includes an effective exclusion interval. Let the detected inter-event interval be $Y = S + \tau_d$, where S is the free interval and τ_d is a deterministic nonparalyzable exclusion time. With $F_0 = \mathrm{Var}(S)/\mathbb{E}[S]^2$ and detected rate

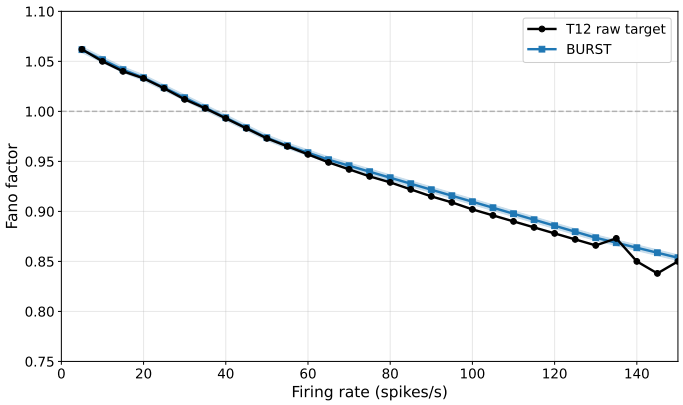


Figure A1: Comparison of Fano factor $\mathrm{Var}(x)/\mathbb{E}[x]$ between T12 empirical recordings and BURST-generated counts across firing rates. Both curves use the estimator of equation 23 with $W{=}9$ on a common 5 spikes per second grid; the shaded region is the spread of the BURST curve over the five sampling seeds. Generated counts reproduce the measured rate-dependent sub-Poisson variability.

$r = 1/\mathbb{E}[Y]$, the renewal limit is

$$\lim_{\Delta\to\infty} \frac{\mathrm{Var}(N_\Delta)}{\mathbb{E}[N_\Delta]} = F_0\big(1 - r\tau_d\big)^2, \qquad 0 \le r\tau_d < 1. \tag{24}$$

This relation provides an effective long-window model. Because neural refractoriness and other forms of short-lag suppression can produce the same trend, the fitted parameter does not separately identify detector dead time. A least-squares fit to the pooled curve gives $\tau_d = 0.74$ milliseconds and $F_0 = 1.05$, with a root-mean-square residual of 0.006 over eight rate buckets.

Session-wise fits show resolvable cross-day variation. After excluding one outlying fit, the between-session standard deviation of τ_d is 0.28 milliseconds, compared with a split-half noise floor of 0.081 milliseconds, while every session remains sub-Poisson in the high-rate band. Direct few-shot estimation remains unreliable: at $N = 100$, bias correction and shrinkage give a mean absolute error of 0.121 milliseconds, compared with 0.092 for the pooled constant. We use the observed cross-day variation to motivate the bounded scalar in equation 3. The scalar δ_d represents an adjustment to dispersion strength and is optimized at the sample level through the finite CMP prior-mean gradient surrogate in equation 30, subject to the sensitivity condition in equation 31; it is not an estimate of τ_d.

E Forward-Schedule Calibration

We partition $[0, 150]$ spikes per second into eight equal-width buckets with midpoint rates r_b. Inverting clause (ii) of Lemma 1 gives the pooled base profile

$$\nu_b = \mathrm{clip}\left(\frac{1}{\mathrm{Fano}(r_b)},\, 1.0,\, 1.7\right), \tag{25}$$

where $\mathrm{Fano}(r_b)$ is the per-channel median curve and the lower bound prevents a low-rate over-dispersed schedule. Let m_b be the pooled mean count in bucket b over the historical training sessions. Setting $\kappa_\tau = 1$ as a time-scale convention, the day-conditioned schedule is

$$\nu_{\tau,d}(b) = 1 + \big(\nu_d(r_b) - 1\big)\tau, \qquad \kappa_\tau(b) = 1, \qquad \lambda_{\tau,d}(b) = m_b^{\,\nu_{\tau,d}(b)}. \tag{26}$$

The pooled values ν_b and m_b are fixed. The linear readout g_ν is initialized at zero, with $\delta_{\max} = 0.35$; when the day condition is dropped, δ_d is set to zero. During training, p_τ and $r_\tau^{\pm}$ use $\mathrm{sg}(\delta_d)$, where sg denotes stop-gradient, keeping the bridge target fixed within an update. The auxiliary losses retain gradients through δ_d via the finite CMP prior mean equation 30. Bridge matching and auxiliary generation assign buckets from the real target trial's channel mean firing rates; deployment uses a retrieved historical template, following equation 40. For each day code, finite-state propagation

Table A1: Nominal forward-schedule exponent per rate bucket. The Fano factor is the per-channel median empirical estimate at each bucket midpoint, and the exponent follows the high-count rule in equation 25. The two lowest buckets are clipped to the Poisson exponent because their empirical Fano factors exceed one.

Bucket midpoint (sp/s)	9.4	28.1	46.9	65.6	84.4	103.1	121.9	140.6
Empirical median Fano	1.044	1.002	0.969	0.941	0.926	0.900	0.864	0.845
Calibrated ν	1.000	1.000	1.032	1.063	1.080	1.111	1.158	1.183

approximates the conditional probabilities on 60 flow-time points using 48 frozen-rate substeps per bucket; reverse-rate targets then follow from equation 4. Both equation 25 and equation 26 use unbounded high-count approximations, so on the lattice $\{0, \dots, 16\}$ they define nominal targets rather than exact moment matches. Table A1 reports the pooled base profile.

F Content Invariance, Auxiliary Critics, and Jump-Rate Guidance

Section 4.4 introduces four auxiliary terms and classifier-free guidance. This section specifies their roles and optimization. The phoneme probe A_ϕ discourages sentence content in the day pathway, the semantic critic D_ω enforces the requested content in generated counts, and the pair (H_ψ, C_ψ) identifies the neural state of a historical training day. These recognition networks are fit only on real historical trials, frozen during generator updates, and discarded after training. Sampling-time guidance uses only the rate network and therefore requires no classifier at deployment.

F.1 Content-invariant day codes and semantic critic

The few-shot set contains both day-level neural state and the content of the recorded calibration sentences. If Φ retains the latter in $\mathbf{n}_d$, the day condition can leak prompt content into samples requested for new phoneme sequences. Content suppression in the day code and semantic fidelity of the generated trial are separate requirements. We therefore apply an adversarial probe to the day-encoder features and a semantic critic to the generated counts.

Let $\mathbf{f}_k \in \mathbb{R}^{T_k \times C \times d_u}$ be the per-bin, per-channel features computed by Φ for prompt trial k before set pooling. A CTC probe A_ϕ reads $\mathbf{f}_k$ through the feature adapter in Table A5 and predicts the prompt sequence $\mathbf{y}_k$. The recognition step fits the probe while stopping gradients to Φ,

$$\mathcal{L}^A_{\text{inv}} = \text{CTC}\big(A_\phi(\text{sg}(\mathbf{f}_k)),\, \mathbf{y}_k\big), \tag{27}$$

where sg denotes stop-gradient. The generator step freezes A_ϕ and updates Φ through the sign-reversed objective equation 8. The clip at m_{inv} stops the adversarial pressure once the probe loss reaches the prescribed margin. Alternating the two updates separates the measurement of residual phoneme information from its removal. Because invariance alone would permit a constant day code, the day-contrastive term below simultaneously requires $\mathbf{n}_d$ to remain informative about the recording day.

The semantic critic D_ω is a separate Conformer CTC decoder trained only on real labeled trials from the historical training window,

$$\mathcal{L}_{\text{realCTC}}(\omega) = \text{CTC}\big(D_\omega(\mathbf{x}),\, \mathbf{y}\big), \tag{28}$$

and never on generated counts, target-day calibration trials, or target test trials. It is also distinct from both the transfer decoder and the independent audit decoder used for evaluation. During a generator update, D_ω is frozen and supplies the semantic term in equation 7. Its confidence weight is

$$w_\omega = \text{sg}\left[\exp\left(-\frac{\text{CTC}(D_\omega(\hat{\mathbf{x}}), \mathbf{y})}{|\mathbf{y}|}\right)\right]. \tag{29}$$

When CTC is viewed as a sequence negative log-likelihood, $w_\omega \in (0, 1]$ is a length-normalized confidence. It downweights unreliable semantic gradients early in training. Stop-gradient prevents the generator from reducing its penalty by changing the weight instead of improving the sample.

Relaxed trajectory and the soft count embedding. The semantic and neural-state losses must differentiate through a discrete sample. Let $\hat{\mathbf{x}}$ be the hard integer sample used for all reported statistics and downstream augmentation. For the auxiliary backward pass, the deterministic trajectory $\mathbf{u}$ starts at the mean of the same finite CMP prior used for hard initialization (Appendix H.4). For $i = (t, c)$, with the training target's channel buckets held fixed, define

$$u_{0,i}(\delta_d) = \sum_{k=0}^{K} k\, \pi_{b_c,d}(k). \tag{30}$$

We differentiate through the normalized CMP probabilities and δ_d, with $\alpha_b = m_b$ fixed; the bridge-matching targets continue to use $\mathrm{sg}(\delta_d)$. Using equation 3, the initialization sensitivity is

$$\frac{\partial u_{0,i}}{\partial \delta_d} = -\big(\nu_d(r_{b_c}) - 1\big)\, \mathrm{Cov}_{X\sim\pi_{b_c,d}}(X, \log(X!))\,. \tag{31}$$

This sensitivity vanishes for buckets with $\nu_d(r_{b_c}) = 1$, and clipping can suppress gradients through the full surrogate. Varying ν_d at fixed α_b changes both the prior mean and variance.

For each coordinate i and step size $\Delta\tau$,

$$u_i \;\leftarrow\; \mathrm{clip}\Big(u_i + \Delta\tau\big[R^{+}_{\theta,i}(\mathbf{u}, \tau \mid \mathbf{P}, \mathbf{z}_d) - R^{-}_{\theta,i}(\mathbf{u}, \tau \mid \mathbf{P}, \mathbf{z}_d)\big],\, 0,\, K\Big), \tag{32}$$

integrated over sixteen steps with step size $\Delta\tau = 1/16$. After integration, we set $\tilde{\mathbf{x}} = \mathrm{sg}(\hat{\mathbf{x}} - \mathbf{u}) + \mathbf{u}$. Thus, $\tilde{\mathbf{x}} = \hat{\mathbf{x}}$ in the forward pass, while both auxiliary losses backpropagate through $\mathbf{u}$. This sixteen-step path is a training surrogate; reported samples use the hard 50-step tau-leaping sampler of Section 5.1. The straight-through gradient is biased, and the nonlinear rate network makes $\mathbf{u}$ a surrogate trajectory rather than the true stochastic-process mean, even when it starts from the exact mean of the initialization prior.

The rate network reads its state through a lookup table over $\{0, \ldots, K\}$. Rounding a coordinate u would break the surrogate gradient, so the relaxed pass linearly interpolates adjacent count embeddings,

$$e(u) = e_{\lfloor u \rfloor} + \big(u - \lfloor u \rfloor\big)\, \big(e_{\min(\lfloor u \rfloor + 1,\, K)} - e_{\lfloor u \rfloor}\big), \tag{33}$$

where e_k is the embedding of count k. This extension agrees with the hard lookup at every integer and is used only by the surrogate path. The bridge targets and integer sampler remain unchanged.

F.2 Neural-state critic and prompt prototype

Phoneme decodability and target-day neural state are separate aspects of conditional fidelity. We train H_ψ and its day classifier C_ψ on real historical trials to evaluate the latter. Their objective combines day classification with supervised contrastive learning,

$$\mathcal{L}^{H}_{\mathrm{day}} = \mathrm{CE}\big(C_\psi(H_\psi(\mathbf{x})), d\big) + \lambda_{\mathrm{sup}} \mathcal{L}_{\mathrm{supcon}}\big(H_\psi(\mathbf{x}), d\big). \tag{34}$$

Positive pairs are different sentences from the same day, and negatives come from other historical days. The critic must therefore identify neural statistics that persist across phoneme content. In a training episode, the first prompt subset defines

$$\mathbf{c}_d = \frac{1}{|S^{(1)}_d|} \sum_{(\mathbf{x},\mathbf{y}) \in S^{(1)}_d} H_\psi(\mathbf{x}). \tag{35}$$

During the generator update, H_ψ and C_ψ are frozen. The state loss in equation 7 pulls the generated embedding toward $\mathbf{c}_d$ and requires classification as day d. This loss is evaluated only in historical-day training episodes. At deployment on an unseen target day, neither the critic nor the prototype is used; the learned set encoder transfers the constraint amortized over the historical days.

F.3 Day-contrastive term

The invariance loss alone could remove all information from $\mathbf{n}_d$. The day-contrastive term prevents this collapse. Let $\mathbf{q}_d = g_\xi(\mathbf{n}_d) \in \mathbb{R}^{d_e}$ and $\mathbf{q}^{+}_d$ be the embeddings of two disjoint prompt subsets from day d. Embeddings $\mathbf{q}_j$ from the other days in the minibatch are negatives:

$$\mathcal{L}^{\mathrm{day}}_{\mathrm{InfoNCE}} = -\log \frac{\exp\big(\cos(\mathbf{q}_d, \mathbf{q}^{+}_d)/\tau_c\big)}{\exp\big(\cos(\mathbf{q}_d, \mathbf{q}^{+}_d)/\tau_c\big) + \sum_{j:\, d_j \neq d} \exp\big(\cos(\mathbf{q}_d, \mathbf{q}_j)/\tau_c\big)}. \tag{36}$$

Here, cos is cosine similarity, τ_c is the contrastive temperature, and g_ξ is a training-only projection head. Gradients update both g_ξ and Φ, while g_ξ is discarded after training.

F.4 Coupled guidance on jump rates

Independent dropout of the phoneme and day conditions trains one rate network in four modes: unconditional, phoneme-only, day-only, and jointly conditioned. This permits classifier-free guidance without any auxiliary classifier at sampling time (Ho & Salimans, 2021). For an exact condition c, the Doob h-transform multiplies an unconditional jump rate by the likelihood ratio between the post-jump and current states (Nisonoff et al., 2024); the conditional-to-unconditional network ratio is the learned counterpart of this factor.

Let $a \in \{+, -\}$ denote a birth or death jump of coordinate i. At state $\mathbf{x}_\tau$ and time τ, let $R^a_{\varnothing,i}$, $R^a_{\mathbf{P},i}$, $R^a_{\mathbf{z}_d,i}$, and $R^a_{\mathbf{P},\mathbf{z}_d,i}$ denote the four rate-network outputs above, with their common arguments omitted. For each admissible jump with positive base rate, define

$$\rho^a_{\mathbf{P},i} = \frac{R^a_{\mathbf{P},i}}{R^a_{\varnothing,i}}, \qquad \rho^a_{\mathbf{z}_d,i} = \frac{R^a_{\mathbf{z}_d,i}}{R^a_{\varnothing,i}}, \qquad \rho^a_{\mathrm{int},i} = \frac{R^a_{\mathbf{P},\mathbf{z}_d,i} / R^a_{\varnothing,i}}{\rho^a_{\mathbf{P},i}\rho^a_{\mathbf{z}_d,i}}. \tag{37}$$

The guided rate is

$$\widetilde{R}^a_{\theta,i} = R^a_{\varnothing,i}\left[\rho^a_{\mathbf{P},i}\right]^{w_p}\left[\rho^a_{\mathbf{z}_d,i}\right]^{w_d}\left[\rho^a_{\mathrm{int},i}\right]^{w_{\mathrm{int}}}, \tag{38}$$

where w_p and w_d control phoneme and day adherence, and w_{int} controls the dependence correction. The setting $(0, 0, 0)$ uses the unconditional network rate, $(1, 1, 0)$ composes the two single-condition ratios without an interaction correction, and $(1, 1, 1)$ exactly recovers the jointly conditioned network rate. When w_p or w_d differs from one, setting $w_{\mathrm{int}} = 1$ adds the dependence correction while retaining the amplification, so the resulting rate differs from the unamplified joint rate. The reported setting $(1.5, 1.5, 0)$ amplifies the two conditions independently. The four network modes are evaluated at the same current state, initialized once from π_{init}. Changing guidance weights changes the rate field while leaving the initialization rule fixed. Because template retrieval and the prior exponent still depend on the requested phonemes and day code, $(0, 0, 0)$ yields an unconditional network rate within a sampler that retains these two sources of conditioning.

Only admissible jumps are reweighted; rates forbidden at 0 or K remain zero. The guided off-diagonal rates are non-negative, and resetting the diagonal to their negative sum yields a valid continuous-time Markov generator. This establishes generator validity; the guided process may follow a different marginal path from the unguided process.

F.5 Alternating training step

The four auxiliary terms enter $\mathcal{L}_G$ in equation 9. A minibatch spans several historical days, with a target trial and two disjoint prompt subsets for each day; Appendix H gives the sampling details. Condition dropout is applied only to the rate-matching pass that trains the four guidance modes. The generated samples used by $\mathcal{L}_{\mathrm{sem}}$ and $\mathcal{L}_{\mathrm{state}}$ retain both conditions, as required by their definitions in equation 7.

Algorithm 1 One alternating training step of BURST

1: sample a minibatch of days; for each d, draw a target $(\mathbf{x}, \mathbf{y})$ and disjoint prompts $S_d^{(1)}, S_d^{(2)} \subseteq S_d \setminus \{(\mathbf{x}, \mathbf{y})\}$
2: encode prompt features $\mathbf{f}_k \leftarrow \Phi_{\text{feat}}(\mathbf{x}_k)$ and day codes $\mathbf{n}_d^{(i)} \leftarrow \Phi(S_d^{(i)})$ for $i \in \{1, 2\}$; form $\mathbf{z}_d$ by equation 6 and δ_d by equation 3
3: **Recognition update:** update the semantic critic D_ω on real CTC loss equation 28; update the state encoder H_ψ, C_ψ on equation 34; update the invariance probe A_ϕ on equation 27
4: set $x_1 \leftarrow \mathbf{x}$ and pad $\mathbf{P} \leftarrow \Psi(\mathbf{y})$ to length T; for the rate-matching pass, independently drop the phoneme and day conditions, setting $\delta_d = 0$ when the day condition is dropped
5: sample $\tau \sim \mathcal{U}(0, 1)$ and $x_\tau \sim p_\tau(\cdot \mid x_1)$ under equation 26 using $\text{sg}(\delta_d)$
6: predict $R_\theta^{\pm}(x_\tau, \tau \mid \mathbf{P}, \mathbf{z}_d)$, evaluate the targets $r_\tau^{\pm}$ from equation 4, and form $\mathcal{L}_{\text{CB}}$
7: restore both conditions, set $T \leftarrow \widehat{T}(\mathbf{y})$ by equation 39, repad $\mathbf{P} \leftarrow \Psi(\mathbf{y})$ to T, and recompute δ_d by equation 3
8: use the real target $\mathbf{x}$ as the reference for equation 40, with T_{tmpl} equal to its observed length; form $\pi_{\text{init}}(\cdot \mid \mathbf{y}, S_d^{(1)}; \mathbf{b})$ by equation 41–equation 42, draw $\mathbf{x}_0$ from this prior, and apply the 50-step hard sampler to obtain $\hat{\mathbf{x}}$
9: initialize $\mathbf{u}_0$ from the same CMP probabilities by equation 30 using unstopped δ_d; integrate the 16-step deterministic surrogate and form $\tilde{\mathbf{x}} = \text{sg}(\hat{\mathbf{x}} - \mathbf{u}) + \mathbf{u}$
10: freeze D_ω, H_ψ, C_ψ, and A_ϕ; form $\mathcal{L}_{\text{sem}}$ and $\mathcal{L}_{\text{state}}$ by equation 7, $\mathcal{L}_{\text{InfoNCE}}^{\text{day}}$ by equation 36, and $\mathcal{L}_{\text{inv}}^{\Phi}$ by equation 8
11: **Generator update:** update the rate network, phoneme conditioner, set encoder, drift and dispersion heads, canonical tuning, and g_ξ on $\mathcal{L}_G$ in equation 9

G NETWORK ARCHITECTURES

This section reports the layer configurations and separates the networks by their role. The deployed generator comprises the phoneme encoder Ψ, the rate network R_θ, the set encoder Φ, the drift head, and the dispersion readout g_ν. Training additionally uses the phoneme probe A_ϕ, the semantic critic D_ω, the day-state encoder and classifier (H_ψ, C_ψ), and the day-contrastive projection head g_ξ; all are discarded before sampling. The transfer decoder and the independent conditional-fidelity decoder are evaluation networks and are not components of BURST. Both the generator and the evaluation decoders use all $C = 128$ threshold-crossing count channels, with each count represented on $\{0, \ldots, 16\}$.

G.1 GENERATOR: RATE NETWORK AND CONDITIONERS

Given the current state x_τ, flow time τ, phoneme stream $\mathbf{P} = \Psi(\mathbf{y})$, and day tuning $\mathbf{z}_d$, the rate network predicts the two reverse jump-rate fields R_τ^+ and R_τ^-. Table A2 specifies the rate network, and Table A3 specifies the modules that produce $\mathbf{P}$, $\mathbf{z}_d$, and δ_d. Learned null phoneme and day embeddings support the four conditioning modes required by the guidance rule in Appendix F.4.

G.2 CONDITIONING DETAILS

Phoneme stream and padding. The phoneme encoder embeds $\mathbf{y} = (y_1, \ldots, y_L)$ and applies three residual temporal convolutional blocks. Each block contains a depthwise convolution with kernel size 7, layer normalization, and a two-layer feed-forward module with expansion factor four. The output is right-padded or truncated to T with a learned padding token, giving $\mathbf{P} \in \mathbb{R}^{T \times d_p}$. Padding only matches the tensor lengths, and the rate network infers the phoneme boundaries.

For rate-matching training, T is the observed length of the real endpoint trial. For auxiliary generation and deployment, the generated count-sequence length is determined by the number of phonemes in the requested sequence:

$$\widehat{T}(\mathbf{y}) = \lfloor |\mathbf{y}|\, \widehat{\rho} \rceil\,, \qquad \widehat{\rho} = \text{median}_{i \in \mathcal{I}_{\text{hist}}} \frac{T_i}{L_i}, \tag{39}$$

Table A2: Rate network R_θ, the BURST generator core. Dimensions: count embedding d_c=64, hidden d_h=256, phoneme d_p=128, tuning d_z=64; lattice cap K=16, channels C=128.

Stage	Module	Configuration
Input	Count state / conditions	$x_\tau \in \{0, \dots, 16\}^{T \times C}$, time τ, stream P, tuning $\mathbf{z}_d$
Coordinate embedding	Count embedding	$\mathrm{Embedding}(17 \to 64)$
	Channel embedding	parameter 128×64
	Projection	$\mathrm{Linear}(64 \to 256)$
Temporal token	Channel pooling	mean over C channels, one token per bin
	Phoneme fusion	$\mathrm{Linear}(256{+}128 \to 256)$
	Temporal position	sinusoidal encoding in $\mathbb{R}^{256}$
	Flow-time MLP	$\mathrm{Linear}(1 \to 256)$, GELU, $\mathrm{Linear}(256 \to 256)$
Backbone	Transformer encoder ($\times 6$)	d=256, heads= 8, feed-forward $256 \to 1024 \to 256$
Rate heads	Bin context $W^{\pm}$	$\mathrm{Linear}(256 \to 64)$
	Coordinate state $C^{\pm}$	$\mathrm{Linear}(64 \to 64)$, fused with $\mathbf{z}_d[c]$
	Per-channel bias	parameter 128
Guidance	Null conditions	learned $\mathrm{null}_P \in \mathbb{R}^{128}$, $\mathrm{null}_z \in \mathbb{R}^{64}$
Output	Reverse rates	$R_\tau^+, R_\tau^- \geq 0$ per coordinate (t, c); invalid boundary jumps are masked

Table A3: Conditioning modules: phoneme encoder Ψ, few-shot set encoder Φ, drift head, and dispersion readout g_ν. Dimensions: phoneme d_p=128, day code d_u=64, tuning d_z=64, drift rank r=8, day embedding d_e=128; conditioner vocabulary 41 (40 phonemes + padding).

Module	Component	Configuration
Phoneme encoder Ψ	Embedding	$\mathrm{Embedding}(41 \to 128)$, right-pad/truncate to T
	ConvNeXt1d ($\times 3$)	depthwise $\mathrm{Conv1d}(128, k{=}7)$, LN, $\mathrm{Linear}(128 \to 512)$, GELU, $\mathrm{Linear}(512 \to 128)$, residual
	Output	phoneme stream $P \in \mathbb{R}^{T \times 128}$
Set encoder Φ	Count and position	$\mathrm{Embedding}(17 \to 64)$ + learned electrode-position parameter 128×64
	Encoder MLP	$\mathrm{Linear}(64 \to 64)$, GELU, $\mathrm{Linear}(64 \to 64)$
	Pooling	mean over bins and trials $\to$ day code $\mathbf{n}_d \in \mathbb{R}^{128 \times 64}$
Drift head	Canonical tuning	parameter $\mathbf{z} \in \mathbb{R}^{128 \times 64}$
	Low-rank mixing	$U_d, V_d = \mathrm{Linear}(64 \to 8)$, $R_d\mathbf{z} = \mathbf{z} + U_d(V_d^\top \mathbf{z})$
	FiLM modulation	$\gamma = 1 + \mathrm{Linear}(64 \to 64)$, $\beta = \mathrm{Linear}(64 \to 64)$; $\mathbf{z}_d = (R_d\mathbf{z}) \odot \gamma + \beta$
Dispersion readout g_ν	Day pooling	mean over channels $\to \bar{\mathbf{n}}_d \in \mathbb{R}^{64}$
	Scalar head	zero-initialized $\mathrm{Linear}(64 \to 1)$, clip to $[-0.35, 0.35]$
Day projection g_ξ	MLP (training only)	$\mathrm{Linear}(64 \to 128)$, GELU, $\mathrm{Linear}(128 \to 128)$, mean over channels

where $\mathcal{I}_{\mathrm{hist}}$ indexes the training trials in the historical window of the corresponding transfer protocol, and T_i and L_i are their count-bin and phoneme lengths. The same rule is used for every requested sentence, including sentences that have historical occurrences. Generated length is determined by the rule above and is independent of the historical template used for rate-bucket assignment.

The initial count state, hard generated sample, and deterministic backward surrogate all use $\widehat{T}(\mathbf{y})$ bins, and the phoneme stream is padded to this length separately from the rate-matching pass. Generation therefore adapts its total length to the requested phoneme count using historical statistics alone. The rule supplies only the total length; the rate network learns individual phoneme durations and bin-level boundaries from sequence-level supervision.

Rate-network integration of phoneme and count state. For each coordinate (t, c), the rate network adds count and channel embeddings and projects the result to the hidden width. Mean pooling over channels gives one count-state vector $\mathbf{h}_t$ per bin. The concatenation of $\mathbf{h}_t$ and $\mathbf{P}_t$ is projected back to the hidden width before the positional encoding and the flow-time embedding are added, so every term in the temporal token has width 256. A six-layer Transformer then processes the temporal sequence. Finally, two coordinate-wise heads combine the resulting bin context with the current coordinate embedding and $\mathbf{z}_d[c]$ to produce $R_\theta^+(t, c)$ and $R_\theta^-(t, c)$.

Set encoder and geometry-aware day code. The set encoder receives only the count signals in $S_d = \{(\mathbf{x}_i, \mathbf{y}_i)\}$; the labels supervise the losses of Section 4.4 but are never inputs to Φ. A shared count embedding is combined with a learned electrode-position embedding indexed by array, row, and column, followed by a coordinate-wise MLP. Mean pooling is applied over bins and prompt trials, but not over channels, yielding $\mathbf{n}_d \in \mathbb{R}^{C \times d_u}$. The result is invariant to prompt order while retaining channel identity and electrode position for the drift head.

Drift head and physical interpretation. Separate linear projections of $\mathbf{n}_d$ produce the scale $\boldsymbol{\gamma}$, shift $\boldsymbol{\beta}$, and low-rank factors $U_d, V_d \in \mathbb{R}^{C \times r}$ in equation 6, with $r = 8$. The adapter $R_d = I + U_d V_d^\top$ models correlated cross-channel change, while $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ provide channel-wise feature modulation (Perez et al., 2018). In particular, the additive term can produce nonzero target-day tuning when the corresponding transformed canonical tuning is zero, which a purely multiplicative reweighting cannot do. The zero-initialized readout g_ν maps $\bar{\mathbf{n}}_d = C^{-1} \sum_c \mathbf{n}_d[c]$ to the bounded dispersion adjustment in equation 3. All day modules are shared across days and evaluated in one forward pass; no target-day parameter is optimized at deployment.

G.3 Downstream Conformer CTC transfer decoder

We use a compact Conformer–CTC decoder as a proof of concept. BURST emits counts in the same format as the recordings, so a phoneme decoder that accepts threshold-crossing counts can take its place. The transfer decoder reads the 128 threshold-crossing channels and applies a fixed Gaussian temporal smoother, a day-specific input layer, and an overlapping-window projection. Three Conformer blocks (Gulati et al., 2020) process the projected sequence, followed by a linear CTC head. All substitution arms use this architecture, and all calibrated arms use the same fine-tuning schedule; the arms differ in the composition of their calibration data.

Table A4: Downstream Conformer CTC decoder for the cross-day substitution protocol. Count-only input over 128 channels; all baselines and BURST variants share this backbone.

Stage	Module	Configuration
Input	Count features	128 threshold-crossing counts, 20 ms bins
Front-end	Gaussian smoothing	temporal, width 20, σ=2.0
	Day input layer	$\mathrm{Linear}(128 \to 128)$, identity-initialized, Softsign
	Unfold	window $= 32$, stride $= 4$
	Input projection	$\mathrm{Linear}(128{\times}32 \to 256)$
Encoder ($\times$3)	FFN (macaron, $\frac{1}{2}$)	hidden$= 256$, expansion$= 4$, dropout$= 0.4$
	Local MHSA	heads$= 8$, window$= 64$, unidirectional
	Conv module	expansion$= 2$, kernel$= 31$, BatchNorm1d
	FFN (macaron, $\frac{1}{2}$)	hidden$= 256$, expansion$= 4$, dropout$= 0.4$
Output	CTC head	$\mathrm{Linear}(256 \to 41)$ (40 phonemes + blank)
Objective	Loss	CTC

G.4 Independent conditional-fidelity decoder

The conditional-fidelity evaluation in Table 3 uses the architecture of Table A4 with an independent initialization. For each target day, this decoder is trained on the real training split, frozen, and then applied to real and generated sequences associated with the disjoint test split. It remains separate from generator optimization, substitution fine-tuning, and the semantic critic trained on real historical trials. Because its layer configuration is unchanged, we do not repeat the table.

G.5 TRAINING-ONLY CRITICS

Table A5 lists the recognition networks defined in Appendix F. The phoneme probe A_ϕ applies a shared feature adapter to all 128 channels of $\mathbf{f}_k$ before the Conformer CTC backbone. The semantic critic D_ω applies the same decoder family directly to count sequences. The day-state encoder H_ψ maps temporal count statistics to a normalized state embedding, and C_ψ predicts the historical recording day. These networks are fitted on real historical trials, frozen during generator updates, and discarded after training. The training-only projection head g_ξ is listed with the conditioning modules in Table A3; it is not a critic.

Table A5: Training-only recognition networks. Conformer configurations follow Table A4; all listed modules are discarded after training.

Network	Input	Configuration
Phoneme probe A_ϕ	Φ features on all 128 channels	shared LN(64) and $\mathrm{Linear}(64 \to 1) \to T \times 128$; Conformer CTC
Semantic critic D_ω	128-channel count sequence	Conformer CTC
Day-state encoder H_ψ	128-channel count sequence	frame MLP $128 \to 256 \to 256$; depthwise $\mathrm{Conv1d}(k{=}7)$ and pointwise convolution; masked mean $\oplus$ standard deviation; MLP $512 \to 256 \to 128$; ℓ_2 normalization
Day classifier C_ψ	H_ψ embedding	$\mathrm{Linear}(128 \to$ number of historical days$)$

H TRAINING AND SAMPLING

This section specifies the historical-day episodes used for training, the order of the recognition and generator updates, and the distinction between differentiable training samples and the deployment sampler. The corresponding losses and one complete update are given in Section 4.4 and Algorithm 1.

H.1 OPTIMIZATION HYPERPARAMETERS

We optimize the generator with AdamW for 17,000 steps with a weight decay of 10^{-2}. The learning rate increases linearly to 2×10^{-4} over 200 warmup steps and then follows a cosine schedule to 10^{-6}. The recognition-network optimizers use twice the generator learning rate. The auxiliary weights are $\lambda_{\mathrm{sem}} = 0.15$, $\lambda_{\mathrm{state}} = 0.05$, $\lambda_{\mathrm{day}} = 0.15$, and $\lambda_{\mathrm{inv}} = 0.02$, with invariance margin $m_{\mathrm{inv}} = 8$. The episodic minibatch construction and condition dropout are specified below.

H.2 HISTORICAL-DAY TRAINING EPISODES

Generator training uses only the historical window. Each sampled historical day is treated as an episodic target, teaching the shared day pathway to condition on a new prompt set. We sample $B = 6$ distinct days uniformly so that days with more trials do not dominate the day-contrastive loss. For each day d, we then sample a target trial $(\mathbf{x}, \mathbf{y})$ and two prompt subsets $S_d^{(1)}$ and $S_d^{(2)}$, each containing $m = 6$ trials. The target and the two subsets are pairwise disjoint: $S_d^{(1)}, S_d^{(2)} \subseteq S_d \setminus \{(\mathbf{x}, \mathbf{y})\}$ and $S_d^{(1)} \cap S_d^{(2)} = \varnothing$. The target trial supplies the bridge endpoint and its requested phoneme sequence. For auxiliary generation, its mean firing rates provide the bucket assignment, while the hard generated sample and backward surrogate both use $\widehat{T}(\mathbf{y})$ bins and are initialized independently of its binwise counts. The first prompt subset supplies the day condition and the neural-state prototype in equation 35; the two prompt subsets form the positive pair in equation 36, while codes from other sampled days serve as negatives. Pairwise disjoint sampling restricts the set encoder and prompt prototype to the prompt trials. At deployment, a historical template supplies the channel means used for bucket assignment. Prompt labels enter the training objectives but are never inputs to Φ.

H.3 ALTERNATING OPTIMIZATION

Each iteration performs a recognition update followed by a generator update. The recognition update fits D_ω with equation 28, (H_ψ, C_ψ) with equation 34, and A_ϕ with equation 27, using real historical trials only. Gradients from the probe are stopped at the set-encoder features during this

update. The recognition networks are then frozen while the generator is optimized under equation 9; consequently, they evaluate generated samples without adapting to generator artifacts. The recognition-network optimizers use twice the learning rate of the generator optimizer, as specified in Appendix H.1.

Condition dropout is applied only to the rate-matching pass. The phoneme and day conditions are dropped independently with probability 0.15, and dropping the day condition also sets $\delta_d = 0$. The conditional path and reverse-rate targets use $\mathrm{sg}(\delta_d)$ and are therefore fixed during backpropagation. For the sample-level losses, equation 30 uses the unstopped finite CMP probabilities of equation 41. Gradients can reach g_ν through this finite-normalizer mean and the deterministic surrogate, subject to equation 31 and clipping. The bridge-target propagators remain outside this backward path. By contrast, the samples evaluated by $\mathcal{L}_{\mathrm{sem}}$ and $\mathcal{L}_{\mathrm{state}}$ retain both conditions. The same generator update also optimizes the day-contrastive and invariance terms, including the training-only projection head g_ξ.

H.4 Training samples and deployment sampling

Bridge matching samples states at selected flow times without simulating complete forward trajectories. During training, only the semantic and neural-state losses require complete generated trials, using hard samples in the forward pass and the sixteen-step surrogate of equation 32 for backpropagation. Auxiliary generation and deployment share the finite CMP parameterization and hard sampler, with buckets derived from the real training target and a retrieved historical template, respectively (Appendix H.4). Reported statistics and downstream augmentation use only hard samples.

At deployment, the few-shot set is encoded once to obtain $\mathbf{z}_{d^\star}$ and $\delta_{d^\star}$; no target-day parameter is optimized. For each requested phoneme sequence, the generation length is set to $\widehat{T}(\mathbf{y})$ by equation 39. The reverse chain on this count-bin axis is initialized from the template-conditioned finite CMP prior π_{init} of Appendix H.4 and integrated by 50-step clipped Poisson tau-leaping as specified in equation 43 with guidance weights $(w_p, w_d, w_{\mathrm{int}}) = (1.5, 1.5, 0)$. Each guidance weight scales a log-rate ratio between a single-condition and the unconditional jump rate. Thus, $w_p = w_d = 1.5$ amplifies each single-condition log-rate ratio by 50%, while $w_{\mathrm{int}} = 0$ omits the learned interaction correction (Appendix F.4). The learned components retained at deployment are Ψ, R_θ, and the day pathway comprising Φ, the canonical tuning, and the drift and dispersion heads. The critics and g_ξ are absent. Boundary masking suppresses inadmissible instantaneous jumps, while clipping after each approximate tau-leap enforces the finite count lattice.

Initialization. For auxiliary generation and deployment, set $T = \widehat{T}(\mathbf{y})$ by equation 39 and initialize from the finite CMP prior below. During auxiliary training, the reference $\mathbf{x}_{\mathrm{tmpl}} = \mathbf{x}$ is the real target trial. At deployment, retrieve $\mathbf{x}_{\mathrm{tmpl}}$ from $\mathcal{R}_d$, the historical training window excluding target-day trials, by minimizing the Levenshtein distance between its phoneme sequence and $\mathbf{y}$ with unit edit costs. Break ties by earliest recording day and then lowest within-day trial index. Let T_{tmpl} denote the reference's observed count-bin length.

The reference's channel mean firing rates determine fixed buckets:

$$\begin{aligned} \bar{r}_c &= \frac{1}{T_{\mathrm{tmpl}}\Delta t_{\mathrm{bin}}} \sum_{t=1}^{T_{\mathrm{tmpl}}} x_{\mathrm{tmpl},t,c}, \qquad \Delta t_{\mathrm{bin}} = 0.02\,\mathrm{s}, \\ b_c &= \min\left\{8,\ 1 + \left\lfloor \frac{\mathrm{clip}(\bar{r}_c, 0, 150)}{18.75} \right\rfloor \right\}. \end{aligned} \tag{40}$$

These eight buckets cover 0–150 spikes per second, with higher rates assigned to the last bucket. For each bucket, define

$$\begin{aligned} \pi_{b,d}(k) &= \frac{\alpha_b^k / (k!)^{\nu_d(r_b)}}{Z_K(\alpha_b, \nu_d(r_b))}, \qquad k \in \{0, \dots, K\}, \\ Z_K(\alpha, \nu) &= \sum_{j=0}^{K} \frac{\alpha^j}{(j!)^\nu}, \qquad \alpha_b = m_b. \end{aligned} \tag{41}$$

Here, m_b is the pooled historical mean count per bin (Appendix E); setting $\alpha_b = m_b$ specifies a CMP parameter, whose resulting finite-distribution mean can differ from m_b. Conditional on the

day code and $\mathbf{b} = (b_1, \ldots, b_C)$, draw coordinates independently:

$$\pi_{\text{init}}(\mathbf{x}_0 \mid \mathbf{y}, S_d; \mathbf{b}) = \prod_{t=1}^{\widehat{T}(\mathbf{y})} \prod_{c=1}^{C} \pi_{b_c,d}(x_{0,t,c}), \qquad \mathbf{x}_0 \in \{0, \ldots, K\}^{\widehat{T}(\mathbf{y}) \times C}. \tag{42}$$

The backward surrogate uses the mean of these same coordinate distributions via equation 30, with the same buckets and generated length. For auxiliary samples, the shorthand $p_\theta(\cdot \mid \mathbf{y}, S_d)$ additionally conditions on the target trial's fixed bucket vector.

The finite CMP prior is normalized directly and may differ from the induced source p_{src} of Appendix A. The exact marginal-path guarantee applies to the induced source, while this initialization is a practical approximation.

Tau-leaping sampler. The hard sampler starts from equation 42 and uses $N_{\text{step}} = 50$ equal flow-time steps, with $\Delta\tau = 1/N_{\text{step}}$ and $\tau_m = m\Delta\tau$ for $m = 0, \ldots, N_{\text{step}} - 1$. At each step, evaluate all guided rates $a^{\pm}_{m,i} = \widetilde{R}^{\pm}_{\theta,i}(\mathbf{x}_m, \tau_m \mid \mathbf{P}, \mathbf{z}_d)$ at the current full state and hold them fixed during the leap (Appendix F.4). For $i = (t, c)$, draw conditionally independent increments across coordinates and jump directions:

$$\begin{aligned} B_{m,i} &\sim \text{Poisson}(\Delta\tau\, a^{+}_{m,i}), \\ D_{m,i} &\sim \text{Poisson}(\Delta\tau\, a^{-}_{m,i}), \\ x_{m+1,i} &= \text{clip}(x_{m,i} + B_{m,i} - D_{m,i},\, 0,\, K). \end{aligned} \tag{43}$$

Return $\mathbf{x}_{N_{\text{step}}}$ at flow time 1. Boundary masking sets $a^{-}_{m,i} = 0$ at $x_{m,i} = 0$ and $a^{+}_{m,i} = 0$ at $x_{m,i} = K$; clipping projects out-of-range proposals to the nearest boundary. This finite-step scheme approximates the continuous-time chain. The sixteen-step auxiliary surrogate in equation 32 uses the fully conditioned rates $R^{\pm}_\theta$ and the soft embedding of equation 33, whereas the hard sampler uses the guided rates $\widetilde{R}^{\pm}_\theta$.

H.5 Decoder adaptation and input augmentation

All calibrated arms use the same two-stage decoder fine-tuning schedule, with batch size 16 and a cosine learning-rate schedule in each stage: 4,000 batches from 5×10^{-5} to 10^{-6}, followed by 16,000 batches from 3×10^{-5} to 10^{-6}. Input augmentation adds white noise with standard deviation 0.8 and a constant per-channel offset with standard deviation 0.2, identically for real and generated sentences. At both transfer horizons, each target-day experiment is repeated over five random seeds; we report the mean and standard deviation of test PER.

H.6 Generation texts and evaluation seeds

For the main substitution experiments at $N = 100$ (Tables 1 and 2), we select a fixed set of $2N$ trials from each target day's training split for the 2NR reference. An N-trial subset is used for NR-ALONE and as the real half of every NR+NG arm, and it also contains the shared few-shot set $S_{d^\star}$. The phoneme sequences of the remaining N sentence prompts request generation, while their neural recordings are withheld from NR+NG. The selection is made once per target day and held fixed across all five evaluation seeds. The target-day test split is reserved for PER evaluation and supplies neither calibration trials nor generation requests.

The five seeds vary the tau-leaping Poisson draws in equation 43 and the minibatch order and dropout masks during decoder fine-tuning. Within each seed, BURST, Count-FM, and the three alternative generators use the same real calibration trials, requested phoneme sequences, and initial random-number-generator state for tau-leaping. Decoder fine-tuning follows the shared two-stage schedule. The real-only references use N real calibration trials for NR-ALONE, $2N$ real training trials for 2NR, and all available target-day training trials for ALL-REAL FT. Method–2NR differences are paired by evaluation-seed index.

I PREVIOUS SINGLE-PASS DISENTANGLEMENT OBJECTIVE

We retain the earlier method variant for ablation and historical comparison. The bridge construction, phoneme conditioner, drift parameterization, and sampler are identical to the main method. The difference is the treatment of phoneme content in the day code and the absence of explicit semantic and neural-state critics on generated samples.

In the previous variant, the per-trial feature sequence $\mathbf{f}_k$ from the set encoder was passed through a Conformer CTC probe h_ϕ with a gradient-reversal layer. The same loss was intended both to remove phoneme information from Φ and to align the probe with the downstream decoder family:

$$\mathcal{L}^{\text{phon}}_{\text{CTC,old}} = \text{CTC}\big(h_\phi(\text{GRL}(\mathbf{f}_k)),\, \mathbf{y}_k\big), \qquad \mathcal{L}_{\text{dis,old}} = \mathcal{L}^{\text{day}}_{\text{InfoNCE}} + \beta_{\text{dis}}\mathcal{L}^{\text{phon}}_{\text{CTC,old}}. \tag{44}$$

The gradient-reversal layer is the identity on the forward pass and negates the gradient on the backward pass (Ganin et al., 2016), so a single backward pass trains h_ϕ to decode phonemes while driving Φ to remove phoneme information from $\mathbf{f}_k$. The resulting generator objective was

$$\mathcal{L}_{\text{old}} = \mathcal{L}_{\text{CB}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis,old}}. \tag{45}$$

This objective explicitly enforces invariance of the day code while leaving phoneme decodability and target-day neural state unconstrained in the generated sequence $\hat{\mathbf{x}}$. The main method addresses all three properties through the alternating invariance, semantic, and neural-state losses of Appendix F.

J SCOPE OF THE COUNT-ONLY EVALUATION

BURST models threshold-crossing counts, and our substitution experiments therefore use a count-only decoder. Adding spike-band power improves real-only calibration, as shown in Table A6.

Table A6: Real-only feature comparison using the same reference Conformer and data splits: historical training on Days 1–7 and evaluation on Days 10–14. Direct denotes no target-day fine-tuning; RandomFT-100 (NR-alone in Table 1) fine-tunes the full decoder on 100 randomly selected target-day training sentences, averaged over five seeds. Entries report mean PER across target days; lower is better. Δ reports count-only minus mixed-feature PER for direct transfer and fine-tuning, respectively.

Input feature	Direct avg.	RandomFT-100 avg.	Δ to mixed
Spike count + spike-band power	60.71	44.25	–
Spike count only	62.82	47.98	+2.11 / +3.73

Our results establish the benefit of generated data within the count-only setting; whether this benefit extends to calibration with both feature types remains to be evaluated.

K ADDITIONAL ABLATIONS

The main text evaluates dispersion, conditional fidelity, calibration budgets, and the day-conditioning architecture. This section examines the training objective, followed by sensitivity analyses of inference-time guidance and tau-leaping discretization. The objective table reports the paired substitution gap $\Delta_{\text{sub}} = \text{PER}(\text{NR+NG}) - \text{PER}(\text{2NR})$ under the medium-horizon protocol at $N{=}100$. Lower is better, and 0 denotes exact substitution for the matched $2N$-real reference. The inference-only tables report the independent-decoder PER and the high-rate Fano factor.

K.1 TRAINING-OBJECTIVE ABLATIONS

Beyond $\mathcal{L}_{\text{CB}}$, the generator uses semantic, episodic day-state, day-contrastive, and day-code invariance terms (Section 4.4). Table A7 removes each term in turn and retrains the generator with the remaining settings fixed. Removing the semantic term has the largest observed effect: independent-decoder PER increases from 50.10 to 52.43, and Δ_{sub} increases from $+1.04$ to $+2.21$, corresponding to changes of 2.33 and 1.17 points. Removing the day-state, day-contrastive, and invariance terms increases Δ_{sub} by 0.25, 0.37, and 0.14 points, respectively. The semantic ablation provides the clearest evidence for the contribution of $\mathcal{L}_{\text{sem}}$ to phoneme decodability. The smaller mean

changes are consistent with an aggregate contribution from the other terms, although the table cannot attribute a distinct mechanism to each one. The previous single-pass objective also performs worse on both PER metrics. That variant changes both the loss composition and the optimization scheme, so the comparison evaluates the combined objective and training procedure.

Table A7: Training-objective ablations on T12 (PER $\times 100$). Four variants remove one auxiliary term and retrain the generator; the final variant replaces the complete objective with equation 45. Independent-decoder PER follows Table 3 and is reported as mean $\pm$ standard deviation over five sampling seeds. Δ_{sub} is the mean $\pm$ standard deviation of the seed-paired substitution gap under the medium-horizon protocol at $N{=}100$.

Objective variant	Indep. PER	Δ_{sub}
Full objective (BURST)	$50.10_{\pm 0.31}$	$+1.04_{\pm 0.35}$
$-\lambda_{\mathrm{sem}}$ (semantic)	$52.43_{\pm 0.36}$	$+2.21_{\pm 0.46}$
$-\lambda_{\mathrm{state}}$ (day state)	$50.38_{\pm 0.30}$	$+1.29_{\pm 0.37}$
$-\lambda_{\mathrm{day}}$ (day contrast)	$50.62_{\pm 0.33}$	$+1.41_{\pm 0.39}$
$-\lambda_{\mathrm{inv}}$ (day invariance)	$50.24_{\pm 0.29}$	$+1.18_{\pm 0.36}$
Previous objective (equation 45)	$51.87_{\pm 0.34}$	$+2.01_{\pm 0.44}$

K.2 Guidance sensitivity

We vary $(w_p, w_d, w_{\mathrm{int}})$ at inference while holding the checkpoint and sampler fixed (Appendix F.4). In addition to independent-decoder PER and high-rate Fano, we report the frozen-critic distance $\|H_\psi(\hat{\mathbf{x}}) - \mathbf{c}_d\|_2^2$ as a diagnostic of day-state fidelity. The critic distance measures day-state fidelity in the training critic's representation, while the independent decoder measures phoneme fidelity. Relative to the unamplified setting $(1, 1, 0)$, the default $(1.5, 1.5, 0)$ reduces independent-decoder PER by 0.33 points and the critic-space day-state error by 0.04. Enabling the interaction correction changes PER by only 0.02 points and leaves the day-state error unchanged at this operating point. Across the conditioned settings, the high-rate Fano factor remains between 0.867 and 0.870, close to the empirical value 0.872. The reported weights are stable across the evaluated diagnostics; global optimality and downstream substitution require separate evaluation.

Table A8: Guidance sensitivity on T12 (inference only; 200 prompts per day). Independent-decoder PER follows Table 3. Day-state error is the critic-space squared distance $\|H_\psi(\hat{\mathbf{x}}) - \mathbf{c}_d\|_2^2$. All entries are mean $\pm$ standard deviation over five sampling seeds.

Guidance $(w_p, w_d, w_{\mathrm{int}})$	Indep. PER	Day-state err.	High-rate Fano
Unconditional $(0, 0, 0)$	$74.00_{\pm 0.62}$	$0.42_{\pm 0.03}$	$0.950_{\pm 0.010}$
Unamplified composition $(1, 1, 0)$	$50.43_{\pm 0.31}$	$0.09_{\pm 0.01}$	$0.870_{\pm 0.004}$
Interaction enabled $(1.5, 1.5, 1)$	$50.08_{\pm 0.32}$	$0.05_{\pm 0.01}$	$0.868_{\pm 0.003}$
Default $(1.5, 1.5, 0)$ (BURST)	$50.10_{\pm 0.31}$	$0.05_{\pm 0.01}$	$0.867_{\pm 0.003}$

K.3 Tau-leaping discretization

All reported samples use 50 tau-leaping steps. To assess sensitivity to this numerical choice, we reuse the trained checkpoint and vary the step count from 50 to 800 with the guidance and conditioning fixed. Across the sweep, independent-decoder PER lies between 50.10 and 50.34, while the high-rate Fano factor lies between 0.864 and 0.870. The locked 50-step result falls within both ranges, indicating limited sensitivity of these two fidelity diagnostics to the integration budget. The downstream substitution gap was not recomputed at every step count, so this sweep is limited to the reported fidelity diagnostics.

Table A9: Sensitivity to tau-leaping discretization on T12 (inference only). Independent-decoder PER follows Table 3. The default row is mean $\pm$ standard deviation over five sampling seeds; the sweep row gives the minimum and maximum over the five tested step counts.

Tau-leaping setting	Indep. PER	High-rate Fano
Default: 50 steps (BURST)	$50.10_{\pm 0.31}$	$0.867_{\pm 0.003}$
Sweep: $\{50, 100, 200, 400, 800\}$ steps	50.10–50.34	0.864–0.870