PARALLEL CAUSAL DENOISING: FROM EXACTNESS TO BATCH SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Parallel generation reduces waiting in sequence models, but denoisers restricted to current and earlier positions can distort relationships when dependent positions are generated together. For Gaussian data and a fixed optimal denoiser, we establish a necessary and sufficient condition for exact generation when noise levels decrease continuously. We also quantify the final generation error, revealing why correlation alone can select the wrong grouping: it ignores how much variation remains after using earlier context. We use this analysis to develop a batching method with a near-optimality guarantee. For a specified Gaussian family with weak correlations only between adjacent positions, its final distributional error (KL divergence) is at most 1% above the optimum at the same batch count. We extend selection to trained linear denoisers using their actual sampling updates at the chosen computational budget. Gaussian experiments support the selection guarantee and show a median 5.07-fold speedup in evaluating and selecting complete groupings over a baseline using numerical integration.

1 INTRODUCTION

Fast sequence generation matters in applications such as streaming video, where new frames must arrive promptly and remain consistent with preceding frames. Self Forcing illustrates this practical demand (Huang et al., 2025). Generating several positions together offers more parallelism, but their relationships must still be preserved. We study causal denoisers: models that gradually remove noise while each position uses only its own noisy value and earlier positions. Completing positions one at a time provides finished context but requires waiting; updating them together means relying on earlier values that are still changing. When can such a model generate several positions together without changing the target distribution? If some error is allowed, how should those positions be grouped?

Recent methods offer flexible ways to organize parallel generation. Diffusion Forcing allows positions to have different noise levels, while Block Diffusion generates positions in parallel within sequentially produced blocks (Chen et al., 2024; Arriola et al., 2025). DEMASK uses estimated dependencies to select parallel positions and bounds the error of an individual update relative to the model's joint prediction under its assumptions (Ringel et al., 2026). For a fixed continuous causal denoiser, these advances leave unresolved how a grouping determines the final generated distribution. In particular, a dependence score alone does not establish which groupings are exact or how their final errors compare at the same batch count. Establishing this connection is difficult for two reasons: errors introduced during denoising may be canceled by later updates, and errors in earlier batches change the context used to generate later ones. Both effects must be accounted for before local grouping decisions can be justified by the final output.

Our insight is to choose groups according to the dependence and uncertainty left after using earlier context. Completing a shared influence can make the remaining randomness at later positions independent, permitting exact parallel generation. When dependence remains, the amount of remaining uncertainty also matters: under the same added noise, a smaller signal becomes informative later. Correlation discards this scale information and can therefore favor the wrong grouping. Under our Gaussian assumptions, we follow these remaining fluctuations through the generation process to determine when the output is exact and to calculate the error contributed by each batch. These contributions add up to the final distributional error even when earlier batches were generated imperfectly.

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.

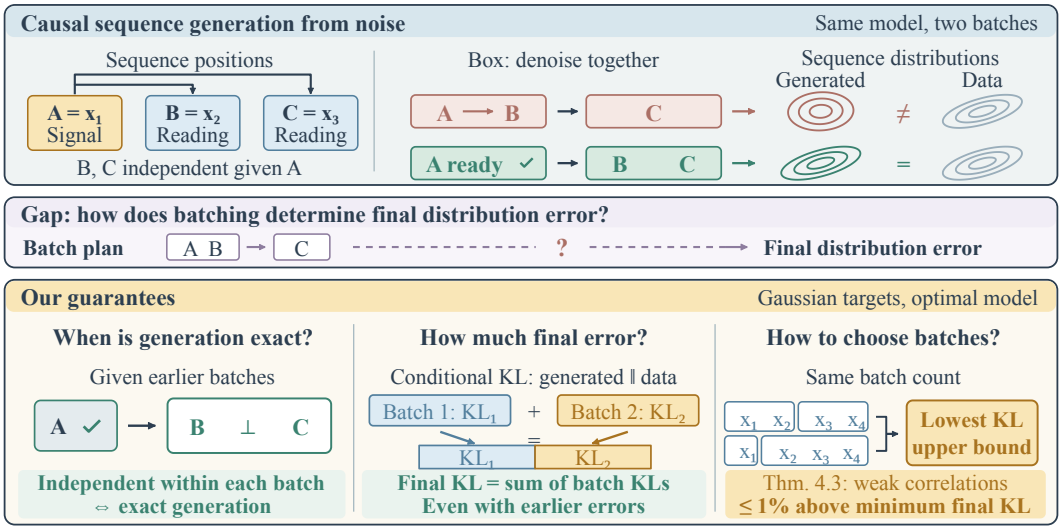


Figure 1: **How batching determines final distribution error.** The two batch plans use the same fixed optimal causal denoiser but yield different joint distributions (schematic ellipses). For Gaussian targets, our theory characterizes exact batching and decomposes final KL into batch costs despite earlier errors (Theorems 3.1 and 4.1). Minimizing a computable KL upper bound achieves error within 1% of the optimum at the same batch count for the weak-adjacent-correlation family of Theorem 4.3.

They therefore provide a direct basis for comparing groupings and selecting one at a fixed batch count.

Our contributions are:

Exactness conditions for parallel generation. We give a necessary and sufficient condition for a fixed optimal causal denoiser to reproduce the final Gaussian distribution under the stated schedules. This identifies which parallel updates preserve the target distribution and which necessarily introduce error.

A quantitative analysis of parallel generation error. We decompose the final distributional error into contributions from individual batches, even when earlier outputs are imperfect. The analysis reveals why correlation alone can select the wrong grouping and provides a computable basis for comparing complete groupings.

A batching method with a near-optimality guarantee. We use bounds on generation error to select consecutive groups, with error at most 1% above the optimum at the same batch count for the specified Gaussian family. We also develop selection using the actual updates of trained linear denoisers and evaluate a nonlinear adaptation on real signals.

2 Background and Motivation

Denoising provides flexible generation schedules. Diffusion models generate samples by progressively recovering structure from noise (Ho et al., 2020). This process can also use deterministic updates, as in DDIM (Song et al., 2021a). For sequences, assigning different noise levels to different positions allows later positions to start changing before earlier ones are complete (Chen et al., 2024). We consider a causal denoiser, where each position can use its own noisy value and earlier positions, but cannot read later positions. This restriction remains in place when positions are updated together. A batch is a consecutive group of positions denoised together, with earlier batches completed first.

Completed context can make parallel generation exact. Figure 1 illustrates three sequence positions: a shared signal A and two readings B and C. In this Gaussian example, A is the only shared influence on the two readings. Once A is known, B and C still vary, but their remaining fluctuations are independent. Completing A first therefore allows B and C to be generated together

exactly. Generating A and B together, followed by C, produces a different joint distribution. Both choices use two batches. Their difference is whether the shared influence has been settled before the dependent quantities are generated together.

The amount of remaining uncertainty also matters. Dependence-based decoding uses relationships between positions to guide parallel generation (Ringel et al., 2026). In our setting, the size of the remaining fluctuations provides additional information. Earlier context may explain most of the variation at one position and little at another. The variance left after using this context is the conditional variance. Standardizing the original data does not make these conditional variances equal. With the same added noise, a smaller remaining signal is harder to distinguish, so positions become informative at different stages of denoising. Correlation normalizes away this difference, even though it affects the error caused by generating positions together.

The remaining question. We still need to connect a grouping decision to the distribution it ultimately generates. Continuous diffusion and flow matching connect suitable update rules to the evolution of a generated distribution (Song et al., 2021b; Lipman et al., 2023). Establishing this connection for the causal rule studied here requires following its complete generation process: an error observed midway could, in principle, be corrected later. Section 3 determines when the final distribution is exactly correct under the stated Gaussian assumptions. Section 4 measures the remaining difference between generated and target distributions using KL divergence, and turns this error analysis into a method for choosing batches.

3 When parallel generation is exact

We now make the distinction in Section 2 precise. The training path describes noisy real data; the sampler repeatedly applies the best causal prediction to its own state. Our goal is to determine when the two processes have the same final distribution. Gaussian targets allow this question to be answered through their means and covariances.

Training path and generated law. Let $X \sim P = \mathcal{N}(\mu, \Sigma)$ on $\mathbb{R}^H$, where Σ is a positive definite correlation matrix and μ is deterministic. For independent $G \sim \mathcal{N}(0, I)$, define

$$Y_i(t) = \alpha_i(t) X_i + b_i(t) G_i, \qquad \alpha_i = \sin(\pi q_i/2), \quad b_i = \cos(\pi q_i/2). \tag{1}$$

Each $q_i : [0,1] \to [0,1]$ is continuous, nondecreasing, starts at zero and ends at one, and is piecewise continuously differentiable with finite one-sided derivatives. Pauses are allowed. Time t indexes denoising, whereas i indexes sequence position. The population squared-loss predictor and its sampler are

$$v_i^{\mathrm{c}}(t, y_{1:i}) = \mathbb{E}[\dot{Y}_i(t) \mid Y_{1:i}(t) = y_{1:i}], \qquad \dot{\widehat{Y}}_i(t) = v_i^{\mathrm{c}}(t, \widehat{Y}_{1:i}(t)), \quad \widehat{Y}(0) = \xi \sim \mathcal{N}(0, I). \tag{2}$$

The fresh initial noise ξ is independent of the variables defining the regression. Dots denote time derivatives, and all differential identities hold almost everywhere. Gaussian regression makes the field affine with integrable coefficients, giving a unique absolutely continuous sampler. Its mean equals that of the training path at every time. We distinguish the training covariance $R(t) = \mathrm{Cov}(Y(t))$ from the generated covariance $S(t) = \mathrm{Cov}(\widehat{Y}(t))$.

The following theorem gives a necessary and sufficient condition. It also establishes that, in this model, correct generation at the final time requires agreement throughout the path.

Theorem 3.1 (Fixed-target exactness)**.** For every target and schedule above, the following are equivalent:

1. $\mathrm{Law}(\widehat{Y}(1)) = P$;
2. $\mathrm{Law}(\widehat{Y}(t)) = \mathrm{Law}(Y(t))$ for every t;
3. for every $i < j$,
$$\dot{\alpha}_i(t)\alpha_j(t)\,\mathrm{Cov}(X_i, X_j \mid Y_{<i}(t)) = 0 \quad \text{almost everywhere.} \tag{3}$$

The covariance in (3) measures the dependence that the current prefix has not resolved. Earlier signal growth is harmless when the later coordinate is still pure noise or this conditional covariance vanishes. Otherwise it produces a terminal error. The condition also detects a reversed generation order: an earlier position cannot generally start after a dependent later position has already become clean.

Why the mismatch cannot disappear. Write the centered sampler as $F(t)\xi$, where F is lower triangular with positive diagonal, and let $T(t) = \mathrm{chol}(R(t))$. Set

$$h_i(t) = \mathrm{Var}(X_i \mid Y_{<i}(t)), \qquad d_i(t) = 1 - \alpha_i(t)^2 + \alpha_i(t)^2 h_i(t).$$

Increasing the earlier signal levels decreases h_i. Direct Gaussian regression gives the identity

$$\frac{d}{dt}\log\frac{F_{ii}(t)}{T_{ii}(t)} = -\frac{\alpha_i(t)^2\dot{h}_i(t)}{2d_i(t)} =: \Gamma_i(t) \geq 0, \qquad \log\frac{\det S(t)}{\det R(t)} = 2\sum_i \int_0^t \Gamma_i(s)\, ds. \tag{4}$$

Thus structural mismatch accumulates in the volume ratio. Terminal equality forces every integral to vanish. Gaussian conditioning identities then turn these zero information rates into (3); an induction over the prefix handles the differing conditioning sets. Conversely, (3) makes the covariance evolution of the training path equal to that of the sampler. The full argument, including regularity at clean boundaries, appears in Appendix A.

The next result quantifies the error when the desired parallelism violates the criterion.

4 From generation error to batch selection

When exact parallel generation is unavailable, batch selection should control the final distributional error at a fixed batch count. We study consecutive partitions $\pi = (B_1, \ldots, B_D)$ of $1{:}H$ into nonempty intervals. Completing these in order maintains a clean prefix, as in block generation (Arriola et al., 2025; Liang et al., 2026). We keep the causal predictor fixed within each batch: active coordinates share a monotone signal profile, earlier batches are clean, and later batches remain pure noise.

Conditioning isolates what a batch must generate. For a Gaussian target, the conditional law of batch B after prefix J is

$$X_B = h_B(X_J) + Z_B, \quad h_B(x_J) = \mu_B + \Sigma_{B,J}\Sigma_{J,J}^{-1}(x_J - \mu_J), \quad Z_B \perp X_J, \tag{5}$$

where $Z_B \sim \mathcal{N}(0, K_B)$ and

$$K_B = \Sigma_{B,B} - \Sigma_{B,J}\Sigma_{J,J}^{-1}\Sigma_{J,B} = T_{B,B}T_{B,B}^{\mathsf{T}}, \qquad T = \mathrm{chol}(\Sigma). \tag{6}$$

For the first batch, $h_B = \mu_B$ and the Schur subtraction is absent. Let $F(K)$ be the terminal linear map of the synchronous causal flow for the local target $\mathcal{N}(0, K)$, starting from standard Gaussian noise (Appendix B.1, (30)). This is the same regression and sampler as before, applied to the residual target.

Let Q_π denote the final distribution generated by partition π. The next theorem shows why a wrong generated prefix does not prevent us from assigning a separate cost to each batch: the prefix changes its conditional mean, while its remaining random variation follows a local flow independent of that prefix value.

Theorem 4.1 (Gaussian batch costs)**.** For every positive definite Gaussian covariance and consecutive synchronous partition, the original population sampler satisfies

$$\widehat{X}_B = h_B(\widehat{X}_J) + F(K_B)\xi_B, \qquad \mathrm{KL}(Q_\pi \| P) = \sum_{B\in\pi} c(K_B), \tag{7}$$

where the ξ_B are independent standard Gaussian noises and

$$c(K) = \tfrac{1}{2}\big(\|U\|_{\mathrm{F}}^2 - |B|\big) - \log\det U, \qquad U = \mathrm{chol}(K)^{-1}F(K). \tag{8}$$

The global generated factor is $T\,\mathrm{blockdiag}(U_B)$, and $c(K_B) = 0$ exactly when K_B is diagonal.

Proof idea. Fix any clean prefix value x_J and subtract $\alpha h_B(x_J)$ from the noisy batch. Gaussian conditional regression becomes the synchronous local regression for $\mathcal{N}(0, K_B)$. The generated prefix remains fixed during that batch, so the same subtraction gives its local ODE even if x_J came from an incorrect generated law. The generated and target conditionals share the translation $h_B(x_J)$; their KL is the local cost, independent of the prefix value. The KL chain rule gives the sum. Expressing these independent residuals in the common coordinates determined by T gives the global factor (Appendix B). □

KL is directed from generated law to target throughout. Conditional variances in K_B are part of the cost: standardizing a residual target changes the flow. This fact will determine the cheap approximation in Section 4.1.

Additive costs make the partition selectable. Writing $c(a, b)$ for the interval $a+1{:}b$, the recurrence

$$V(d, b) = \min_{d-1 \le a < b} \{V(d-1, a) + c(a, b)\}, \qquad V(0, 0) = 0, \quad V(0, b) = +\infty \ (b > 0) \tag{9}$$

finds the minimum actual endpoint KL at depth D by backtracking from $V(D, H)$. For at most D batches, minimize over $d \le D$. Its $O(DH^2)$ operations follow cost preprocessing. In the Gaussian case, prefix consistency obtains all intervals with a common start from one synchronous suffix ODE. We next approximate these costs without solving the local flows.

4.1 Conditional scales matter for batch selection

The exact Gaussian cost requires a local flow solve. To approximate it without that solve, a conditional-correlation score must retain the conditional variances: they determine how much signal each coordinate reveals during a synchronous batch. For its Schur complement K, let $s_i = \sqrt{K_{ii}}$ and $\rho_{ij} = K_{ij}/(s_i s_j)$. Define

$$p(K) = \frac{1}{2} \sum_{i<j} w_{ij}^2 \rho_{ij}^2, \qquad w_{ij} = \frac{s_j}{s_i + s_j}. \tag{10}$$

For fixed batch width, positive diagonal $D_s = \mathrm{diag}(s_i)$, and symmetric C with zero diagonal, the original synchronous sampler satisfies $c(D_s(I + \varepsilon C)D_s) = p(D_s(I + \varepsilon C)D_s) + O(|\varepsilon|^3)$ as $\varepsilon \to 0$. Equal conditional scales give $p(K) = \frac{1}{8} \sum_{i<j} \rho_{ij}^2$. Appendix C proves the expansion and a finite bound for pairs.

The direction of the weight follows from effective signal-to-noise progress. Removing each scalar marginal transport changes the common signal amplitude u into

$$\widetilde{\alpha}_i(u) = \frac{u s_i}{\sqrt{1 - u^2 + u^2 s_i^2}}, \qquad \int_0^1 \widetilde{\alpha}_i \, d\widetilde{\alpha}_j = \frac{s_i}{s_i + s_j}. \tag{11}$$

The integral is the retained first-order transport from earlier i to later j; its missing fraction is w_{ij}. A larger earlier conditional scale provides reliable context sooner and reduces this error. Thus equal signal amplitudes can conceal unequal information progress.

The two-coordinate case gives a finite check on this approximation. For arbitrary positive conditional scales and $|\rho_{12}| \le 1/4$, the exact cost satisfies

$$p(K) \le c(K) \le \frac{7}{5} p(K). \tag{12}$$

The factor is uniform over the variance ratio; the sharper bound in Appendix C tends to one as correlation vanishes. For wider batches, the stated expansion is asymptotic at fixed width and fixed positive scales. Thus retaining conditional scales does not require those scales to be approximately equal.

A different boundary even after standardization. Let $L \ge 1$, $0 < \varepsilon < 1/2$, $\tau = 1/L$, and $a = \sqrt{1 - \tau^2}$. Independent standard normals G_1, G_2, G_3 define a positive definite Gaussian target with unit marginal variances by

$$\begin{aligned} X_1 &= G_1, \qquad X_2 = \varepsilon G_1 + \sqrt{1 - \varepsilon^2} G_2, \\ X_3 &= a G_1 + \tau\big(2\varepsilon G_2 + \sqrt{1 - 4\varepsilon^2} G_3\big). \end{aligned} \tag{13}$$

Its two consecutive two-batch partitions are $\pi_A = (\{1\}, \{2, 3\})$ and $\pi_B = (\{1, 2\}, \{3\})$. Their unweighted conditional-correlation scores are $4\varepsilon^2$ and ε^2, respectively.

Proposition 4.2 (Conditional scales change the optimal boundary)**.** For every $0 < \varepsilon \le 1/4$ at $L = 8$, unweighted conditional correlation uniquely selects π_B, whereas the weighted score and true endpoint KL uniquely select π_A. Moreover,

$$\frac{\mathrm{KL}(Q_{\pi_B} \| P)}{\mathrm{KL}(Q_{\pi_A} \| P)} \ge \frac{8}{5}. \tag{14}$$

For $0 < \varepsilon \le 1/8$, the bound improves to $20/7$.

Conditioning on X_1 leaves standard deviations $(\sqrt{1-\varepsilon^2}, 1/L)$ in the second batch of π_A. These unequal scales make its larger correlation less costly. At fixed $0 < \varepsilon < 1/2$, the same family has an unbounded multiplicative KL ratio as $L \to \infty$. Proposition C.3 gives the proof and the score-tie condition.

Selecting a partition with an estimated covariance. Given a positive definite training-covariance estimate and a batch count, compute each candidate's conditional covariance, evaluate (10), and backtrack (9) to obtain the boundaries. This avoids local flow solves; integrating those flows instead supplies the Gaussian-cost rule (Appendix B). Both select for the Gaussian field defined by the supplied covariance. We next replace point scores by lower and upper costs to control the selected partition's error.

4.2 Batch selection with a near-optimality guarantee

To bound selection error, we need costs for every candidate after its completed prefix. Even an equal-scale nearest-neighbor target loses equal conditional scales: conditioning changes the first remaining variance. Retaining this boundary term in higher-order expansions gives lower and upper costs for every candidate batch.

Minimizing the upper costs selects a partition, while minimizing the lower costs bounds the best attainable error. Their difference bounds how much the selected partition can lose, without solving its exact flow.

Let f_i be coordinate vectors and consider

$$\Sigma = s^2 \left[I + \sum_{i=1}^{H-1} e_i (f_i f_{i+1}^{\mathsf{T}} + f_{i+1} f_i^{\mathsf{T}}) \right], \qquad |e_i| \le \rho, \quad 0 < \rho \le \tfrac{1}{5}, \quad s > 0. \tag{15}$$

The edges can be nonuniform and signed. Define $d_1 = 1$ and $d_j = 1 - e_{j-1}^2/d_{j-1}$. For $B = (a, b]$, write its actual Schur complement as $K_B = s^2(I + M_B)$. The matrix M_B has first diagonal entry $d_{a+1} - 1$, other diagonal entries zero, and adjacent entries $e_{a+1}, \ldots, e_{b-1}$. Further prefix conditioning stays in this family; the changed first variance is retained exactly.

Removing the common scalar transport turns each causal-prefix flow into a finite algebraic expansion U_r of the relative endpoint factor. The expansion retains the changed boundary variance and reuses one suffix calculation for all intervals with the same start (Appendix D). A computable tail bound δ_r on $\|U(K_B) - U_r\|_2$ gives lower and upper costs $\underline{c}_r(B), \overline{c}_r(B)$ by bounding the off-diagonal norm and each diagonal term in (8). The lower cost also uses $c(K_B) \ge (19/160) \sum_{i=a+1}^{b-1} e_i^2$; singletons have exact cost zero. Appendix D gives the explicit tail and interval formulas.

Theorem 4.3 (Reliable complete selection)**.** For (15), these finite intervals enclose all conditional batch costs. Let V_L, V_U be the minimum lower and upper cost sums over consecutive D-batch partitions, and let $\widehat{\pi}$ minimize the upper sum. With $C(\pi) = \mathrm{KL}(Q_\pi \| P)$ and $C_* = \min_\pi C(\pi)$,

$$C(\widehat{\pi}) - C_* \le R_r := V_U - V_L \le 2(H - D)(18\rho\delta_r + 5\delta_r^2). \tag{16}$$

If $\rho/2 \le |e_i| \le \rho$ and $0 < \rho \le 1/5$, then $r = 12$ gives $R_{12} < V_L/100$ for every H and $D < H$. Thus the returned partition satisfies $C(\widehat{\pi}) \le 1.01 C_*$. At $D = H$, the sole all-singleton partition is exact and $R_r = 0$.

Proof idea. The conditional boundary preserves a tridiagonal matrix whose absolute causal prefix powers admit the scalar majorant in (49). Its true adjacent transport errors have a fixed sign, giving the internal-edge cost lower bound. Every D-batch partition retains exactly $H - D$ internal edges, while its nonsingleton batches contain at most $2(H - D)$ coordinates. Both the interval width and the lower cost therefore scale with $H - D$; their ratio is bounded independently of width and depth. Appendix D proves the constants and the 12th-order bound. □

Refinement continues until $R_r \le \eta V_L$ for the requested tolerance η. A selected upper cost below every alternative's lower cost also certifies uniqueness. Refinement eventually separates positive gaps; exact ties retain a regret bound. Appendix E gives the general-scale expansion.

4.3 SELECTING BATCHES FOR A TRAINED DENOISER

Training and a finite sampling budget can change the grouping costs. We therefore evaluate the denoiser's actual updates.

Fix a denoiser affine in the state, causal in both states and noise levels, and a common fixed integration grid for each batch. Composing its updates gives

$$\widehat{X}_B = d_B + G_B \widehat{X}_J + N_B \xi_B, \qquad \widehat{K}_B = N_B N_B^{\mathsf{T}}, \tag{17}$$

with independent standard Gaussian ξ_B and nonsingular N_B. Against the Gaussian reference in (5), define the conditional-mean error $e_B(x) = d_B + G_B x - h_B(x)$ and the covariance cost $\kappa_B = \mathrm{KL}(\mathcal{N}(0, \widehat{K}_B) \| \mathcal{N}(0, K_B))$.

Proposition 4.4 (Finite-step learned costs)**.** For every consecutive partition, its actual output error satisfies

$$C(\pi) := \mathrm{KL}(Q_\pi \| P) = \sum_{B \in \pi} \kappa_B + P_\pi, \qquad P_\pi = \tfrac{1}{2} \sum_{B \in \pi} \mathbb{E}_{Q_{\pi,J}} \left[e_B(\widehat{X}_J)^{\mathsf{T}} K_B^{-1} e_B(\widehat{X}_J) \right] \geq 0. \tag{18}$$

If V_κ is the minimum sum of κ_B over D-batch partitions and π_κ attains it, then $V_\kappa \leq C_* := \min_{|\pi|=D} C(\pi)$ and $C(\pi_\kappa) - C_* \leq P_{\pi_\kappa}$.

The generated prefix depends on earlier groupings, so its conditional-mean contribution generally prevents scalar additivity. Appendix F.3 gives the proof and prefix-moment recurrence.

To select batches, cache the stage maps, minimize the covariance costs κ_B using (9), then propagate the chosen prefix moments to obtain the full cost and gap bound. Intervals with the same start share a suffix computation. For the population field with exact continuous integration, $e_B = 0$ and $\kappa_B = c(K_B)$, recovering Theorem 4.1. Estimated covariances define the supplied Gaussian reference; experiments separately measure true-target error.

5 EXPERIMENTS

Section 4 makes three claims that experiments can test: conditional scales, not correlation alone, determine which grouping is best; the bounds of Theorem 4.3 select near-optimal groupings at low cost; and for a trained sampler with a finite budget, the relevant costs are those of the sampler itself. We test them first on Gaussian targets, where the output KL is exact, and then with a nonlinear causal Transformer on real motion signals. Correlation selection scores squared conditional correlations; weighted and Gaussian-cost selection use (10) and (8); all use the dynamic program (9). Protocols appear in Appendices G to J.

5.1 TESTING THE ROLE OF CONDITIONAL SCALES

Standardizing the data equalizes marginal variances but not conditional ones (Section 2). We test whether this remaining difference changes which grouping generates better samples. Four weakly coupled copies of (13) give twelve positions with unit marginal variances, and L controls their conditional scales. Per target, three independently trained linear denoisers run with six batches and 192 field evaluations (NFE); all rules select from the same training covariance.

Table 1: **Conditional scales change generation error.** Output KL (lower is better), mean ± sample SD over three repetitions at six batches and NFE 192. The paired difference is weighted minus correlation; negative favors weighting. Scale-L targets use $\varepsilon = .30$; stronger correlation uses $(\varepsilon, L) = (.45, 4)$. Reversal preserves the scale-4 spectrum.

Target	Correlation	Weighted	Paired difference
Scale 2	0.0424 ± 0.0017	0.0424 ± 0.0017	0
Scale 4	0.0436 ± 0.0016	0.0331 ± 0.0005	-0.0105 ± 0.0013
Scale 8	0.0453 ± 0.0018	0.0136 ± 0.0012	-0.0317 ± 0.0008
Stronger correlation	0.1065 ± 0.0031	0.1128 ± 0.0026	$+0.0063 \pm 0.0029$
Reversed order	0.1726 ± 0.0158	0.0958 ± 0.0017	-0.0768 ± 0.0141

The gain from weighting grows with the scale contrast. At $L = 2$, all three rules choose identical cuts. At $L = 4$ and 8, weighting changes every repetition's cuts and lowers KL by 21 to 25% and 68 to 71% relative to correlation (Table 1). This follows Proposition 4.2: correlation ignores the small conditional scale of each copy's third position ($1/L$ in (13)) and so overstates the cost of generating it together with the second.

The global spectrum alone does not explain the effect. Reversing the $L = 4$ target preserves its spectrum, yet changes the paired KL difference from $-.0105$ to $-.0768$, because reversal changes the conditional structure and the candidate partitions. With stronger correlation, correlation selection is slightly better, since (10) is only a leading-order score. Conditional scales thus change the best grouping after standardization, but point scores need an error check (Section 5.2).

5.2 Evaluating reliable batch selection

We test whether selection with the bounds of Theorem 4.3 stays within its 1% guarantee and costs less than computing exact costs. On 24 targets from its weakly correlated family, sequence lengths $16, 32, 64$ and three batch counts per target give 72 decisions; refinement stops once the bound on relative excess KL is at most 1%. Numerically integrating the exact flow gives reference costs, which define the optimum, and the timing baseline. Both methods reuse one suffix computation per start, and timing covers all preprocessing, including refinement.

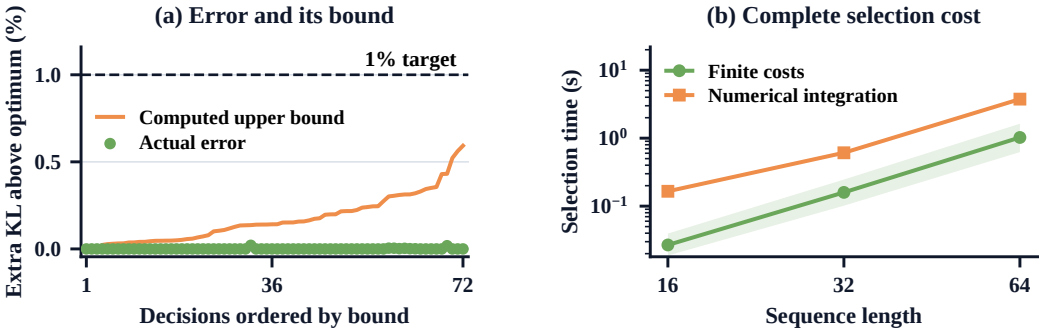


Figure 2: **Reliable, inexpensive batch selection.** Left panel, extra KL and computed upper bounds for all 72 decisions, relative to the optimum at the same batch count. Right panel, complete preprocessing for three batch counts, excluding sampling. Lines and bands show medians and interquartile ranges across eight targets per length, each timed three times and summarized by its median.

The bounds keep every decision near-optimal and cost less to compute. All 72 decisions meet the criterion: the largest computed bound on excess KL is .59%, and the largest measured excess is .021% (Figure 2, left). Point scores lack this check: correlation and weighted selection each reach 6.77% excess on the same grid. Complete preprocessing is also faster on all 24 targets, with a median speedup of 5.07× and a range of 1.42 to 8.73× (right).

5.3 Adapting selection to the sampling budget

A trained sampler with a finite budget departs from the population flow, so its best grouping may differ. We test whether costs computed from the sampler's own updates (Proposition 4.4) track this difference. We reuse the twelve denoisers of Section 5.1, excluding reversal, with six batches at NFE 96 and 192, giving 24 model–budget cases. Learned costs use the actual midpoint updates and an independent validation covariance; Gaussian and weighted costs retain their training-covariance inputs. Enumerating all 462 partitions gives the optimum for each case.

The best grouping depends on the sampling budget. On the stronger-correlation target at NFE 96, mean excess KL is 81.52% for Gaussian costs, 2.52% for weighting and .18% for learned costs; at NFE 192, Gaussian costs narrowly lead (Figure 3). Gaussian costs choose the same partition at both budgets, so this reversal comes from the finite-step sampler. Learned costs, which include the sampler, are closest to the optimum on average: across the 24 cases they average .58% excess, with maximum 4.12%, against 3.43% for weighting and 12.15% for Gaussian costs.

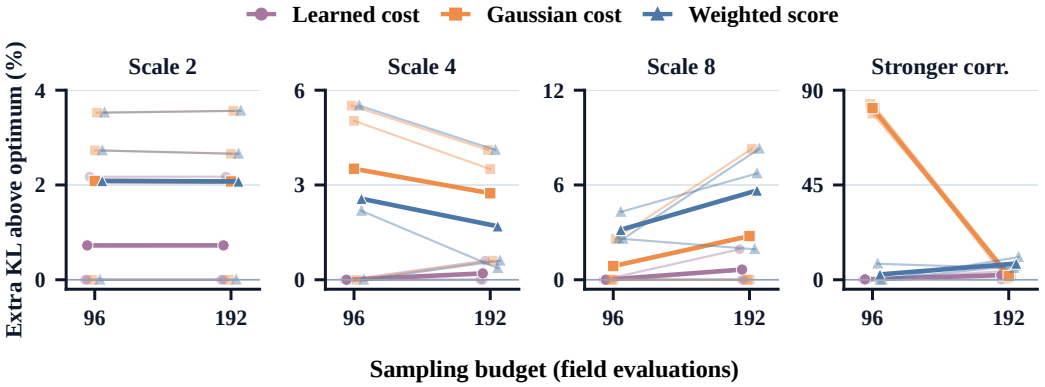


Figure 3: **Extra target KL above each budget's optimum.** Thin lines track each of three repetitions across budgets; thick lines show means. All 24 cases are shown; panel scales differ.

5.4 Evaluating selection on real signals

For a nonlinear model on real signals, exact costs are unavailable, so we test whether costs from sampling updates can replace expensive validation search. Three independently trained causal Transformers generate 32 HAPT acceleration points (2.56 seconds) from the preceding 32, using eight batches and NFE 32, with separate training, validation and test subjects (Reyes-Ortiz et al., 2016).

Response-based selection measures, at 36 training contexts, how each batch's finite updates respond to input changes, composes these affine responses into a Gaussian approximation of each partition's output, and scores 32 candidate partitions by expected energy (Gneiting & Raftery, 2007) against local training references, each as a whole because energy is nonadditive. Full validation search instead samples the same candidates on validation prefixes and keeps the lowest energy.

Table 2: **Selection quality on real signals.** Mean energy reduction of response-based selection over each comparator, measured with actual samples on 158 held-out segments from nine test subjects; positive values favor response-based selection.

Comparator	Energy reduction (%)
Full validation search	+.223
Similar-time validation search	+1.820
Same responses, Gaussian KL	+8.430
Equal batches	+9.148

Response-based selection reaches comparable quality in about a quarter of the time. Including reference fitting, responses and scoring, it takes 3.701s versus 15.479s for full validation search on the same hardware, excluding model loading and generation, with .223% lower mean energy (Table 2). A validation search given the same time is 1.82% worse, and equal batches are 9.15% worse. Scoring the same responses by Gaussian KL instead of energy is 8.43% worse, so the score matters as well.

6 Conclusion

We characterize exact parallel generation for a fixed optimal causal denoiser on Gaussian targets and decompose final KL error into batch costs even when earlier batches are generated imperfectly. Conditional scales explain why correlation alone can select the wrong grouping. Computable bounds yield near-optimal selection for the specified Gaussian family, while costs from actual sampling updates extend selection to trained affine denoisers. These results make final distributional error a basis for choosing which positions to generate together.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and improving readability. The authors reviewed the AI-assisted revisions and take responsibility for the final content of this work.

REFERENCES

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.09573.

Wenrui Bao, Zhiben Chen, Dan Xu, and Yuzhang Shang. Learning to parallel: Accelerating diffusion large language models via learnable parallel decoding. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bFJ8Sdr224.

Yiyi Cai, Yuhan Wu, Kunhang Li, You Zhou, Bo Zheng, and Haiyang Liu. FloodDiffusion: Tailored diffusion forcing for streaming motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.03520.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2407.01392.

Tianqi Du, Lizhe Fang, Weijie Yang, Chenheng Zhang, Zeming Wei, Yifei Wang, and Yisen Wang. Autoregressive models rival diffusion models at ANY-ORDER generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.13228.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Ligong Han, Hao Wang, Han Gao, Kai Xu, and Akash Srivastava. S2D2: Fast decoding for diffusion LLMs via training-free self-speculation. *arXiv preprint arXiv:2603.25702*, 2026. URL https://arxiv.org/abs/2603.25702.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Lanxiang Hu, Siqi Kou, Yichao Fu, Samyam Rajbhandari, Tajana Rosing, Yuxiong He, Zhijie Deng, and Hao Zhang. Fast and accurate causal parallel decoding using Jacobi Forcing. *arXiv preprint arXiv:2512.14681*, 2025. URL https://arxiv.org/abs/2512.14681.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. URL https://arxiv.org/abs/2506.08009.

Mahdi Karami and Ali Ghodsi. Auto-regressive masked diffusion models. *arXiv preprint arXiv:2601.16971*, 2026. URL https://arxiv.org/abs/2601.16971.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19274–19286. PMLR, 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

Jiajun Liang, Yucheng Liao, Yukang Cao, Jiazhe Wei, Ken Li, Wende Tan, Jiankun Zhang, ZY Cui, Jingkang Yang, Liucheng Guo, Shiqi Yang, B. Yang, Caifeng Shan, Ziwei Liu, and Chenyang Si. AURORA-LM: Autoencoding unified representation for continuous-latent diffusion language modeling. *arXiv preprint arXiv:2608.02602*, 2026. URL https://arxiv.org/abs/2608.02602.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xiaocheng Lu, Huabin Liu, Song Guo, and Jianguo Li. Reducing pretraining-generation mismatch in diffusion language models. *arXiv preprint arXiv:2608.09424*, 2026. URL https://arxiv.org/abs/2608.09424.

Peter Pao-Huang, Xiaojie Qiu, and Stefano Ermon. Generative modeling with flux matching. *arXiv preprint arXiv:2605.07319*, 2026. URL https://arxiv.org/abs/2605.07319.

Jorge-L. Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. Transition-aware human activity recognition using smartphones. *Neurocomputing*, 171:754–767, 2016. doi: 10.1016/j.neucom.2015.07.085. URL https://doi.org/10.1016/j.neucom.2015.07.085.

Liran Ringel, Ameen Ali, and Yaniv Romano. Dependency-guided parallel decoding in discrete diffusion language models. *arXiv preprint arXiv:2604.02560*, 2026. URL https://arxiv.org/abs/2604.02560.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/MWR-D-14-00269.1.

Andy Shih, Suneel Belkhale, Stefano Ermon, Dorsa Sadigh, and Nima Anari. Parallel sampling of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.16317.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://arxiv.org/abs/2010.02502.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://arxiv.org/abs/2011.13456.

Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *Journal of Machine Learning Research*, 19(66):1–71, 2018. URL https://jmlr.org/papers/v19/17-747.html.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026a. URL https://arxiv.org/abs/2602.02214.

Yichen Zhu, Xiaoming Shi, Peng Zhao, Weiyu Chen, Debing Zhang, and James Kwok. Dynamic chunking for diffusion language models. *arXiv preprint arXiv:2605.15676*, 2026b. URL https://arxiv.org/abs/2605.15676.

Yuchen Zhu, Jing Shi, Chongjian Ge, Hao Tan, Yiran Xu, Wanrong Zhu, Jason Kuen, Koustava Goswami, Rajiv Jain, Yongxin Chen, Molei Tao, and Jiuxiang Gu. FLARE: Diffusion for hybrid language model. *arXiv preprint arXiv:2606.01774*, 2026c. URL https://arxiv.org/abs/2606.01774.

Tamim Zoabi, Ameen Ali, Liran Ringel, and Lior Wolf. Mean-field parallel decoding for discrete diffusion language models. *arXiv preprint arXiv:2606.15805*, 2026. URL https://arxiv.org/abs/2606.15805.

A WHY A GAUSSIAN PARALLEL-GENERATION ERROR CANNOT CANCEL

The proof compares the generated triangular factor with the Cholesky factor of the training path. Each diagonal ratio accumulates a nonnegative conditional-information rate. The essential final step is to convert its vanishing into the structural condition, whose conditioning set differs from the one in that rate.

A.1 REGRESSION, REGULARITY, AND THE TWO COVARIANCE PATHS

Let $X \sim \mathcal{N}(\mu, \Sigma)$ on $\mathbb{R}^H$, where Σ is positive definite with unit diagonal and μ is deterministic. Choose $0 < \underline{r} \le 1 \le \overline{r}$ with $\underline{r} I \preceq \Sigma \preceq \overline{r} I$. For independent $G \sim \mathcal{N}(0, I)$, set

$$Y_i(t) = \alpha_i(t) X_i + b_i(t) G_i, \qquad \alpha_i = \sin(\pi q_i/2), \quad b_i = \cos(\pi q_i/2). \tag{19}$$

Each $q_i : [0,1] \to [0,1]$ is continuous, nondecreasing, starts at zero, ends at one, and is continuously differentiable on a finite partition with finite one-sided derivatives. Pauses are allowed. All time derivatives below hold almost everywhere, and all ODE solutions are absolutely continuous. In particular, $\dot\alpha_i \ge 0$, $\int_0^1 \dot\alpha_i(t)\, dt = 1$, and $\alpha_i \dot\alpha_i + b_i \dot b_i = 0$. The population field and its sampler are

$$v_i^{\mathrm{c}}(t, y_{1:i}) = \mathbb{E}[\dot Y_i(t) \mid Y_{1:i}(t) = y_{1:i}], \qquad \dot{\widehat{Y}}_i = v_i^{\mathrm{c}}(t, \widehat{Y}_{1:i}), \quad \widehat{Y}(0) = \xi \sim \mathcal{N}(0, I), \tag{20}$$

where ξ is a fresh draw, distinct from the variables defining the regression. Write $< i = \{1, \dots, i-1\}$, and use empty conditioning for $i = 1$. Matrix subscripts select rows and columns; chol denotes the lower-triangular Cholesky factor with positive diagonal.

Put $A_\alpha = \mathrm{diag}(\alpha_i)$, $m = A_\alpha \mu$, and

$$R(t) = \mathrm{Cov}(Y(t)) = \mathrm{diag}(b_i^2) + A_\alpha \Sigma A_\alpha. \tag{21}$$

The quadratic form of R and $\alpha_i^2 + b_i^2 = 1$ give $\underline{r} I \preceq R \preceq \overline{r} I$. This also bounds every prefix covariance. Gaussian regression gives a lower-triangular matrix L with

$$\begin{aligned} g_i &= \mathrm{Cov}(\dot Y_i, Y_{1:i}) = (\dot\alpha_i \alpha_1 \Sigma_{i1}, \dots, \dot\alpha_i \alpha_{i-1} \Sigma_{i,i-1}, 0), \\ L_{i,1:i} &= g_i R^{-1}_{1:i,1:i}, \qquad v_i^{\mathrm{c}}(t, y_{1:i}) = \dot m_i + L_{i,1:i}(y_{1:i} - m_{1:i}). \end{aligned} \tag{22}$$

Indeed, the regression residual of jointly Gaussian variables is uncorrelated with, hence independent of, the conditioning variables. The last entry of g_i is zero because $\mathrm{Var}(Y_i) = 1$. The formula defines a version of the field at every prefix value. Moreover, $\|L_{i,1:i}\|_2 \le \underline{r}^{-1} \|\Sigma_{i,<i}\|_2 \dot\alpha_i$, so all coefficients are integrable. Scalar integrating factors, successively in coordinate order, give the unique solution

$$\widehat{Y} = m + F\xi, \quad \dot F = LF, \quad F(0) = I, \qquad S := \mathrm{Cov}(\widehat{Y}) = FF^{\mathsf{T}}, \quad \dot S = LS + SL^{\mathsf{T}}. \tag{23}$$

Here F is lower triangular and $F_{ii}(t) = \exp(\int_0^t L_{ii}) > 0$. In particular, the generated mean is $m(t)$ throughout the path.

Define

$$E = \dot R - LR - RL^{\mathsf{T}}, \qquad T = \mathrm{chol}(R), \qquad U = T^{-1} F. \tag{24}$$

Cholesky factorization is smooth on the positive definite cone. The spectral bounds make T, T^{-1} absolutely continuous with integrable derivatives. Since U is lower triangular with positive diagonal,

$$S = R \iff UU^{\mathsf{T}} = I \iff U = I. \tag{25}$$

We also use the Schur bound: if $Q \succeq \underline{r} I$, then $Q_{22} - Q_{21} Q_{11}^{-1} Q_{12} \succeq \underline{r} I$, because its quadratic form at z is $\min_w (w, z)^{\mathsf{T}} Q (w, z) \ge \underline{r} \|z\|_2^2$.

A.2 CONDITIONAL-INFORMATION IDENTITIES

All following conditional covariances are Gaussian and do not depend on the values conditioned on. Define

$$\begin{aligned} h_i &= \mathrm{Var}(X_i \mid Y_{<i}), & d_i &= \mathrm{Var}(Y_i \mid Y_{<i}) = 1 + \alpha_i^2 (h_i - 1), \\ \kappa_{ij} &= \mathrm{Cov}(X_i, X_j \mid Y_{<i}) \quad (i < j), & A_{ir} &= \{1, \dots, i-1\} \setminus \{r\} \quad (r < i), \\ \chi_{ir} &= \mathrm{Cov}(X_i, X_r \mid Y_{A_{ir}}), & w_{ir} &= \mathrm{Var}(Y_r \mid Y_{A_{ir}}). \end{aligned} \tag{26}$$

The notation A_{ir} in this subsection denotes a conditioning set.

Lemma A.1 (Residual and information derivative)**.** The residual has zero diagonal, and for $i < j$,

$$E_{ij} = E_{ji} = \frac{\dot{\alpha}_i \alpha_j}{d_i} \kappa_{ij}, \qquad L_{ii} = \frac{\alpha_i \dot{\alpha}_i (h_i - 1)}{d_i}, \qquad \frac{\partial h_i}{\partial \alpha_r} = -\frac{2\alpha_r \chi_{ir}^2}{w_{ir}^2} \quad (r < i). \tag{27}$$

The conditional variances d_i and w_{ir} are at least $\underline{r}$. These conditional covariance functions and signal derivatives extend continuously to the closed signal cube $[0,1]^H$.

Proof. For $i < j$, row j of the regression gives $(LR)_{ji} = \dot{\alpha}_j \alpha_i \Sigma_{ji}$; also $(LR)_{ii} = 0 = \dot{R}_{ii}$. Thus $E_{ij} = \mathrm{Cov}(\dot{Y}_i, Y_j \mid Y_{1:i})$. Conditioning first on $Y_{<i}$ gives

$$\mathrm{Cov}(\dot{Y}_i, Y_i \mid Y_{<i}) = \alpha_i \dot{\alpha}_i (h_i - 1), \quad \mathrm{Cov}(\dot{Y}_i, X_j \mid Y_{<i}) = \dot{\alpha}_i \kappa_{ij}, \quad \mathrm{Cov}(Y_i, X_j \mid Y_{<i}) = \alpha_i \kappa_{ij}.$$

The first formula divided by d_i is L_{ii}. Conditioning once more on Y_i gives

$$\mathrm{Cov}(\dot{Y}_i, X_j \mid Y_{1:i}) = \dot{\alpha}_i \kappa_{ij} - \frac{\alpha_i \dot{\alpha}_i (h_i - 1)}{d_i} \alpha_i \kappa_{ij} = \frac{\dot{\alpha}_i}{d_i} \kappa_{ij}.$$

Multiplication by α_j proves the residual formula.

For the derivative, condition first on $Y_{A_{ir}}$ and put $v_r = \mathrm{Var}(X_r \mid Y_{A_{ir}})$. Adding Y_r reduces the conditional variance of X_i by

$$\frac{\alpha_r^2 \chi_{ir}^2}{1 + \alpha_r^2 (v_r - 1)}, \qquad w_{ir} = 1 + \alpha_r^2 (v_r - 1).$$

Both v_r and χ_{ir} are independent of the variable α_r in this partial derivative. Differentiating proves the third identity. Every joint covariance used here is a principal submatrix of a corruption covariance, assigning signal one to any coordinate written as X_i or X_r; no coordinate appears twice. The spectral and Schur bounds give $d_i, w_{ir} \geq \underline{r}$. The joint covariance entries are polynomial functions of the signals, with uniformly invertible conditioning matrices. This proves the stated regularity also at clean and noisy boundaries. □

A.3 Volume growth forces the endpoint criterion

Theorem A.2 (Fixed-target terminal exactness)**.** Under the assumptions above, the following are equivalent: (i) $\mathrm{Law}(\widehat{Y}(1)) = \mathcal{N}(\mu, \Sigma)$; (ii) $\mathrm{Law}(\widehat{Y}(t)) = \mathcal{N}(m(t), R(t))$ for every t; (iii) for every $i < j$, $\dot{\alpha}_i \alpha_j \kappa_{ij} = 0$ almost everywhere. More precisely, setting

$$\Gamma_i = -\frac{\alpha_i^2 \dot{h}_i}{2 d_i} = \frac{\alpha_i^2}{d_i} \sum_{r<i} \frac{\alpha_r \dot{\alpha}_r \chi_{ir}^2}{w_{ir}^2} \geq 0 \tag{28}$$

gives, for every t,

$$U_{ii}(t) = \exp\left(\int_0^t \Gamma_i(s)\, ds \right), \qquad \log \frac{\det S(t)}{\det R(t)} = 2 \sum_i \int_0^t \Gamma_i(s)\, ds. \tag{29}$$

If E is nonzero on a set of positive measure, the terminal volume ratio is strictly greater than one.

Proof. Since $T_{ii} = \sqrt{d_i}$ and $F'_{ii} = L_{ii} F_{ii}$,

$$\frac{d}{dt} \log U_{ii} = L_{ii} - \frac{\dot{d}_i}{2 d_i} = -\frac{\alpha_i^2 \dot{h}_i}{2 d_i} = \Gamma_i.$$

The conditional variance h_i depends only on the earlier signals. The chain rule and Lemma A.1 give the sum in (28). Its terms are nonnegative, bounded by constants times the integrable signal derivatives. Integrating from $U_{ii}(0) = 1$, then taking determinants, proves (29).

If the endpoint is exact, $U(1) = I$ by (25), so every $\Gamma_i = 0$ almost everywhere. Fix a time outside the common null set of these identities. We may also assume that $\dot{\alpha}_r > 0$ implies $\alpha_r > 0$: an absolutely continuous nonnegative function has derivative zero almost everywhere on its zero set. For such an r, every $i > r$ with $\alpha_i > 0$ has $\chi_{ir} = 0$, since the rth summand in Γ_i vanishes. We must show $\kappa_{ri} = 0$, which uses a smaller conditioning set.

Induct on $i > r$ in coordinate order, considering those with $\alpha_i > 0$. Let $B = \{r+1, \dots, i-1\}$. Conditioning further on Y_B after $Y_{<r}$ gives

$$\chi_{ir} = \kappa_{ri} - \operatorname{Cov}(X_r, Y_B \mid Y_{<r}) \operatorname{Var}(Y_B \mid Y_{<r})^{-1} \operatorname{Cov}(Y_B, X_i \mid Y_{<r}).$$

The subtracted term vanishes for empty B. Otherwise, each entry of its left factor is $\alpha_j \kappa_{rj}$, $j \in B$. This is zero either because $\alpha_j = 0$, or by the induction hypothesis when $\alpha_j > 0$. Hence $\kappa_{ri} = \chi_{ir} = 0$ at this step. This proves the induction. Pairs with $\dot\alpha_r = 0$ or $\alpha_i = 0$ already satisfy (iii), so (i) implies (iii).

By (27), (iii) is equivalent to $E = 0$ almost everywhere. Differentiation then gives

$$\frac{d}{dt}\big(F^{-1} R F^{-\mathsf{T}}\big) = F^{-1} E F^{-\mathsf{T}} = 0.$$

Its initial value is I, so $R = FF^{\mathsf{T}} = S$, proving (ii). Conversely, (ii) implies $E = 0$ through the covariance equations, and implies (i) at $t = 1$. Finally, a terminal volume ratio of one forces every Γ_i to vanish; the same induction forces $E = 0$. The contrapositive proves strictness when the residual persists. □

The universal-order consequence needs only an isolated correlated pair: for $\Sigma = I + \rho(e_i e_j^{\mathsf{T}} + e_j e_i^{\mathsf{T}})$, $i < j$, $0 < |\rho| < 1$, one has $\kappa_{ij} = \rho$. Thus a schedule is exact for every correlation target precisely when $\int_0^1 \alpha_j \dot\alpha_i = 0$ for all $i < j$. This is equivalent to $\alpha_j(t) > 0 \Rightarrow \alpha_i(t) = 1$: a violation at t_0 gives an integral at least $\alpha_j(t_0)(1 - \alpha_i(t_0)) > 0$, while under the pointwise condition the integrand vanishes almost everywhere.

B Actual Gaussian batch errors and partition selection

This section allows any positive definite Σ, since conditioning a correlation target need not leave unit residual variances. The interpolation, population field, and schedule regularity are as in (19)–(20). Gaussian regression remains affine and well posed: use $\underline{r} = \min\{1, \lambda_{\min}(\Sigma)\}$ in the covariance lower bound, and replace the final entry of g_i by $\alpha_i \dot\alpha_i (\Sigma_{ii} - 1)$. All entries remain integrable. We always direct KL from the generated law to the target.

B.1 The synchronous batch cost

For a positive definite $k \times k$ covariance K, define

$$\begin{aligned} R_K(u) &= I + u^2(K - I), \qquad & [L_K(u)]_{i,1:i} &= u(K-I)_{i,1:i}[R_K(u)_{1:i,1:i}]^{-1}, \\ F_K'(u) &= L_K(u) F_K(u), & F_K(0) &= I, \qquad 0 \le u \le 1, \end{aligned} \tag{30}$$

where L_K is zero above the diagonal. Since $R_K \succeq \min\{1, \lambda_{\min}(K)\} I$, the coefficient is continuous, and F_K is uniquely defined, lower triangular, with positive diagonal. Set

$$\begin{aligned} F(K) = F_K(1), \quad T(K) = \operatorname{chol}(K), \quad U(K) = T(K)^{-1} F(K), \\ c(K) = \frac{1}{2}\big(\|U(K)\|_{\mathrm{F}}^2 - k\big) - \log\det U(K). \end{aligned} \tag{31}$$

The Gaussian density formula identifies $c(K)$ as $\mathrm{KL}(\mathcal{N}(0, F(K)F(K)^{\mathsf{T}}) \| \mathcal{N}(0, K))$. In triangular coordinates it is

$$c(K) = \frac{1}{2} \sum_{i<j} U(K)_{ji}^2 + \sum_i \left\{ \frac{U(K)_{ii}^2 - 1}{2} - \log U(K)_{ii} \right\}. \tag{32}$$

This retains the residual variances in K; replacing K by its correlation matrix changes the local flow.

A consecutive partition is $\pi = (B_1, \dots, B_D)$ with $B_d = \{n_{d-1} + 1, \dots, n_d\}$ and $0 = n_0 < n_1 < \cdots < n_D = H$. Its schedule completes the batches in order. Active coordinates share a signal $u(t)$ going from zero to one, earlier coordinates are clean, and later coordinates remain pure noise. For an interval $B = \{a+1, \dots, b\}$ after $J = \{1, \dots, a\}$, write

$$h_B(z) = \mu_B + \Sigma_{B,J} \Sigma_{J,J}^{-1} (z - \mu_J), \qquad K_B = K_{a,b} = \Sigma_{B,B} - \Sigma_{B,J} \Sigma_{J,J}^{-1} \Sigma_{J,B}. \tag{33}$$

For empty J, take $h_B = \mu_B$ and $K_B = \Sigma_{B,B}$.

B.2 THE CONDITIONAL IDENTITY REMAINS VALID AT AN INCORRECT PREFIX

Theorem B.1 (Conditional innovations and endpoint KL)**.** For every such target and partition, the original population sampler has endpoint law Q_π satisfying

$$\widehat{X}_B = h_B(\widehat{X}_J) + F(K_B)\xi_B, \qquad \mathrm{KL}(Q_\pi \| P) = \sum_{B\in\pi} c(K_B), \tag{34}$$

with independent standard Gaussian noises ξ_B. If $T = \mathrm{chol}(\Sigma)$, its global factor is

$$\widehat{X} - \mu = F_\pi \xi, \qquad F_\pi = T\,\mathrm{blockdiag}_{B\in\pi}\, U(K_B). \tag{35}$$

Furthermore, $c(K) = 0$ if and only if K is diagonal.

Proof. Fix an active batch and a clean prefix value z. Gaussian conditioning writes $X_B = h_B(X_J) + Z_B$ with $Z_B \sim \mathcal{N}(0, K_B)$ independent of X_J. At $X_J = z$, the batch interpolation is $uh_B(z) + uZ_B + \sqrt{1-u^2}G_B$. Its centered covariance is $R_{K_B}(u)$, and the covariance of the physical-time velocity in coordinate i with its local prefix is $u\dot{u}(K_B - I)_{i,1:i}$. Regression therefore gives

$$v_i^{\mathrm{c}} = \dot{u}\, h_{B,i}(z) + \dot{u}[L_{K_B}(u)]_{i,1:i}(y_{B,1:i} - uh_{B,1:i}(z)). \tag{36}$$

The diagonal term includes the derivative of the corruption noise. The formula holds at every z, not only under the target prefix law.

Before activation the batch has zero velocity and remains ξ_B. The generated prefix depends only on earlier initial noises, so it is independent of ξ_B. Its completed coordinates have zero velocity and stay fixed during the batch. Hence subtracting $uh_B(\widehat{X}_J)$ from the generated state gives $\dot{Z} = \dot{u}L_{K_B}(u)Z$ starting at ξ_B. The unique solution is $F_{K_B}(u(t))\xi_B$, including pauses. This proves the endpoint identity even if the prefix law is wrong. At every prefix value, the target and generated conditional laws have the same location $h_B(z)$, and their conditional KL is $c(K_B)$ independently of z. Factoring both joint densities into their batch conditionals and integrating the log ratio under Q_π proves the KL sum.

For the factor formula, the leading Cholesky blocks give

$$T_{B,B}T_{B,B}^{\mathsf{T}} = K_B, \qquad \Sigma_{B,J}\Sigma_{J,J}^{-1} = T_{B,J}T_{J,J}^{-1}.$$

Induction on batches in the endpoint identity now gives $F_{\pi,B,J} = T_{B,J}\,\mathrm{blockdiag}_{C\subseteq J}\, U(K_C)$ and $F_{\pi,B,B} = F(K_B) = T_{B,B}U(K_B)$, proving (35).

It remains to characterize zero cost for arbitrary variances. In the local path put $a_i(u) = \mathrm{Var}(X_i \mid Y_{<i}(u))$, $d_i(u) = 1 + u^2(a_i(u) - 1)$, $p_i = K_{i,<i}$, and $C_i(u) = I + u^2(K_{<i,<i} - I)$. Gaussian conditioning gives

$$a_i = K_{ii} - u^2 p_i C_i^{-1} p_i^{\mathsf{T}}, \qquad a_i' = -2u\|C_i^{-1}p_i^{\mathsf{T}}\|_2^2, \qquad \frac{d}{du}\log\frac{[F_K(u)]_{ii}}{\sqrt{d_i(u)}} = -\frac{u^2 a_i'}{2d_i} \geq 0.$$

The derivative of $u^2 C_i^{-1}$ is $2uC_i^{-2}$, which proves the middle identity; for the empty prefix $a_1' = 0$. If K is not diagonal, some $p_i \neq 0$, so the last derivative is strictly positive on $(0,1)$. Then $U(K)_{ii} > 1$, and its diagonal summand in (32) is positive. If K is diagonal, the scalar equations give $F_K(u) = \mathrm{diag}\,\sqrt{1 + u^2(K_{ii} - 1)}$ and $U(K) = I$. This proves both directions, including zero singleton cost. □

B.3 THE INTERVAL DYNAMIC PROGRAM

Proposition B.2 (KL-optimal consecutive partition)**.** For $1 \leq D \leq H$, let $c(a,b) = c(K_{a,b})$. The recurrence

$$V(d,b) = \min_{d-1\leq a<b}\{V(d-1,a) + c(a,b)\}, \qquad V(0,0) = 0, \quad V(0,b) = +\infty\ (b > 0), \tag{37}$$

for $1 \leq d \leq b \leq H$ gives the minimum actual endpoint KL among consecutive D-batch partitions as $V(D,H)$. Backtracking gives a minimizing partition. The recurrence uses $O(DH^2)$ operations after cost preprocessing; at most D batches are handled by minimizing $V(d,H)$ over $d \leq D$.

Proof. Every partition ending at b has a unique last cut a. Its last cost conditions on the entire prefix $1:a$ but does not depend on how that prefix was partitioned. The KL sum and induction on d therefore give (37). There are $O(DH)$ states, each with at most H predecessors. □

For a fixed start a, $K_{a,b}$ is a leading principal block of $K_{a,H}$. The same consistency holds for F_K, by the leading ODEs, and for $T(K)$ by the Cholesky recursion. One suffix ODE per start therefore supplies all terminal interval costs, using the cumulative row sums in (32). The same dynamic program can instead minimize any supplied additive interval score; the next section gives one that requires no local ODE solve.

C Conditional scales and a different optimal boundary

C.1 Leading KL at arbitrary positive conditional scales

For each positive definite K, put

$$s_i = \sqrt{K_{ii}}, \quad \rho_{ij} = K_{ij}/(s_i s_j), \quad w_{ij} = \frac{s_j}{s_i + s_j}, \qquad p(K) = \frac{1}{2}\sum_{i<j} w_{ij}^2 \rho_{ij}^2. \tag{38}$$

This score is the leading term of the same local flow defined in (30). For the calculations only, change clock to $x = u^2$ and denote $\Phi_K(x) = F_K(\sqrt{x})$. Then

$$\Phi_K' = \tfrac{1}{2} M_K \Phi_K, \quad \Phi_K(0) = I, \qquad [M_K(x)]_{j,1:j} = (K-I)_{j,1:j}[I + x(K-I)_{1:j,1:j}]^{-1}, \tag{39}$$

with zero entries above the diagonal. The coefficient extends continuously to both endpoints, and $\Phi_K(1) = F(K)$.

Proposition C.1 (Leading cost at arbitrary positive scales)**.** Fix a width k, a positive diagonal matrix $D_s = \mathrm{diag}(s_1, \dots, s_k)$, and a symmetric zero-diagonal matrix C. For $K_\varepsilon = D_s(I + \varepsilon C)D_s$, as $\varepsilon \to 0$ through positive definite covariances,

$$\begin{aligned} F(K_\varepsilon)_{ji} &= \frac{\varepsilon s_i s_j C_{ij}}{s_i + s_j} + O(\varepsilon^2), \qquad & U(K_\varepsilon)_{ji} &= -\varepsilon \frac{s_j}{s_i + s_j} C_{ij} + O(\varepsilon^2) \quad (i<j), \\ U(K_\varepsilon)_{ii} &= 1 + O(\varepsilon^2), & c(K_\varepsilon) &= p(K_\varepsilon) + O(|\varepsilon|^3). \end{aligned} \tag{40}$$

The constants may depend on k, D_s, C.

Proof. Write $A_i(x) = 1 + x(s_i^2 - 1)$ and $J = D_s C D_s$. At $\varepsilon = 0$, $\Phi_{K_0}(x) = \mathrm{diag}(\sqrt{A_i(x)})$. Differentiating the prefix inverse in (39) gives

$$[\partial_\varepsilon M_{K_\varepsilon}(x)|_0]_{ji} = \frac{J_{ji}}{A_j A_i} \quad (i<j), \qquad [\partial_\varepsilon M_{K_\varepsilon}(x)|_0]_{jj} = 0.$$

The first term of the off-diagonal derivative is J_{ji}/A_i; the inverse derivative subtracts $x(s_j^2 - 1)J_{ji}/(A_j A_i)$, leaving the expression displayed. The scalar diagonal integrating factor then gives

$$[\partial_\varepsilon F(K_\varepsilon)|_0]_{ji} = \frac{s_j J_{ji}}{2} \int_0^1 \frac{dx}{A_j^{3/2} A_i^{1/2}} = \frac{J_{ji}}{s_i + s_j}. \tag{41}$$

For unequal scales the integral follows by differentiating $\sqrt{A_i/A_j}$, whose derivative is $(s_i^2 - s_j^2)/(2A_j^{3/2}A_i^{1/2})$; equal scales follow by continuity. The diagonal variation is zero.

Differentiating $T(K_\varepsilon)T(K_\varepsilon)^\mathsf{T} = K_\varepsilon$ gives zero diagonal variation and $[\partial_\varepsilon T(K_\varepsilon)|_0]_{ji} = J_{ji}/s_i$. Since $F(K_0) = T(K_0) = D_s$,

$$[\partial_\varepsilon U(K_\varepsilon)|_0]_{ji} = \frac{1}{s_j}\left(\frac{J_{ji}}{s_i + s_j} - \frac{J_{ji}}{s_i}\right) = -\frac{s_j}{s_i + s_j} C_{ij}.$$

Near zero all prefix inverses are uniformly nonsingular. Their coefficients are smooth in ε, and differentiating the linear ODE successively gives bounded first and second parameter derivatives on

$[0,1]$. The factor expansions therefore have the stated remainders. In (32) the diagonal summand has a double zero at $U_{ii}=1$, so its contribution is $O(\varepsilon^4)$; the strict lower triangle gives

$$c(K_\varepsilon)=\frac{\varepsilon^2}{2}\sum_{i<j}\left(\frac{s_j}{s_i+s_j}\right)^2 C_{ij}^2+O(|\varepsilon|^3)=p(K_\varepsilon)+O(|\varepsilon|^3).$$

□

The direction of the weight can also be read from information progress. Set $Z=D_s^{-1}X$ and remove the exact scalar standard deviation $g_i(u)=\sqrt{1-u^2+u^2s_i^2}$. Then

$$\frac{Y_i(u)}{g_i(u)}=\widetilde{\alpha}_i(u)Z_i+\sqrt{1-\widetilde{\alpha}_i(u)^2}G_i,\qquad \widetilde{\alpha}_i(u)=\frac{us_i}{g_i(u)}.$$

Since $\widetilde{\alpha}_j'(u)=s_j/g_j(u)^3$, the same integral as in (41) yields

$$\int_0^1 \widetilde{\alpha}_i\,d\widetilde{\alpha}_j=s_is_j\int_0^1\frac{u\,du}{g_ig_j^3}=\frac{s_i}{s_i+s_j},\qquad \int_0^1\widetilde{\alpha}_j\,d\widetilde{\alpha}_i=\frac{s_j}{s_i+s_j}. \tag{42}$$

The first fraction is the retained first-order transport after division by the final scale s_j; the second is its missing fraction. The integrals sum to one by integration by parts.

C.2 A FINITE PAIR BOUND UNIFORM IN THE VARIANCE RATIO

Proposition C.2 (Pair costs at finite correlation)**.** For $s,t>0$ and $|\rho|<1$, let $K=\left(\begin{smallmatrix}s^2 & \rho st\\ \rho st & t^2\end{smallmatrix}\right)$, $w=t/(s+t)$, and $p=p(K)=w^2\rho^2/2$. Then

$$p\le c(K)\le p\,\Gamma(w,|\rho|),\qquad \Gamma(w,r)=\exp\left(\frac{2wr^2}{1-r^2}\right)\frac{1+r^2}{(1-r^2)^2}. \tag{43}$$

In particular, $|\rho|\le 1/4$ gives $p\le c(K)\le(7/5)p$. For nonzero $\rho\to 0$, $c(K)/p=1+O(\rho^2)$ uniformly over $s,t>0$; at $\rho=0$ both costs vanish.

Proof. Put $z=\rho st$, $A=1+x(s^2-1)$, $B=1+x(t^2-1)$, and $\Delta=AB-x^2z^2$. The clock-x coefficient in (39) is

$$M_{11}=(s^2-1)/A,\qquad M_{21}=z/\Delta,\qquad M_{22}=((t^2-1)A-xz^2)/\Delta.$$

The path covariance has Cholesky factor

$$T_x=\begin{pmatrix}\sqrt{A} & 0\\ xz/\sqrt{A} & \sqrt{\Delta}/\sqrt{A}\end{pmatrix}.$$

Define $U(x)=T_x^{-1}\Phi_K(x)$, so $U(0)=I$ and $U(1)=U(K)$. Differentiation gives

$$U_{11}=1,\quad U_{22}'=h'U_{22},\quad U_{21}'=-\frac{z}{2A\sqrt{\Delta}}+h'U_{21},\qquad h'=\frac{xz^2}{2A\Delta},\quad h(0)=0. \tag{44}$$

Set $q=xz/\sqrt{AB}$ and $f_0=|z|/(2A^{3/2}B^{1/2})$. Since $xs^2/A\le 1$ and $xt^2/B\le 1$, $|q|\le|\rho|$. The source magnitude is $f_0/\sqrt{1-q^2}$ and

$$h'=\frac{f_0|q|}{1-q^2},\qquad \int_0^1 f_0\,dx=w|\rho|.$$

The last integral follows from (41), exchanging the two diagonal factors. Write $h_*=h(1)$. The integrating-factor formula, the fixed sign of the source, and $h'\ge 0$ give

$$0\le h_*\le\frac{w\rho^2}{1-\rho^2},\qquad w|\rho|\le|U_{21}(1)|\le\frac{e^{h_*}w|\rho|}{\sqrt{1-\rho^2}},\qquad U_{22}(1)=e^{h_*}. \tag{45}$$

Thus

$$c(K) = \tfrac{1}{2}U_{21}(1)^2 + \tfrac{1}{2}(e^{2h_*} - 1) - h_*, \qquad 0 \le \tfrac{1}{2}(e^{2h_*} - 1) - h_* = \int_0^{h_*} (e^{2t} - 1)\, dt \le h_*^2 e^{2h_*}.$$

The lower bound is $c(K) \ge p$. For the upper bound, the two terms are at most

$$p\frac{e^{2h_*}}{1-\rho^2} + pe^{2h_*}\frac{2\rho^2}{(1-\rho^2)^2} = pe^{2h_*}\frac{1+\rho^2}{(1-\rho^2)^2},$$

which proves (43). If $|\rho| \le 1/4$, using $w \le 1$ and $e^x \le (1-x)^{-1}$ for $0 \le x < 1$ gives

$$\Gamma(w, |\rho|) \le e^{2/15}\frac{272}{225} \le \frac{15}{13}\frac{272}{225} = \frac{272}{195} < \frac{7}{5}.$$

The same expression gives $\Gamma(w, r) = 1 + O(r^2)$ uniformly in $0 < w < 1$. When $\rho = 0$, (44) gives $U = I$. □

C.3 Decision separation after global standardization

Proposition C.3 (Complete comparison of two boundaries)**.** For finite $L \ge 1$ and $0 < \varepsilon < 1/2$, set $\tau = 1/L$, $a = \sqrt{1-\tau^2}$, and let independent standard normals G_1, G_2, G_3 define

$$\begin{aligned} X_1 &= G_1, \qquad X_2 = \varepsilon G_1 + \sqrt{1-\varepsilon^2}G_2, \\ X_3 &= aG_1 + \tau\big(2\varepsilon G_2 + \sqrt{1-4\varepsilon^2}G_3\big). \end{aligned} \tag{46}$$

The target $P_{\varepsilon,L}$ is positive definite with unit marginal variances. Write C_A, C_B for the true endpoint KLs of $\pi_A = (\{1\}, \{2,3\})$ and $\pi_B = (\{1,2\}, \{3\})$, and P_A, P_B for their sums of $p(K_B)$. Unweighted conditional correlation uniquely selects π_B, whereas

$$P_A = \frac{2\varepsilon^2}{(L\sqrt{1-\varepsilon^2}+1)^2}, \qquad P_B = \frac{\varepsilon^2}{8}, \qquad \frac{P_B}{P_A} = \frac{(L\sqrt{1-\varepsilon^2}+1)^2}{16}. \tag{47}$$

Thus the weighted minimizer is uniquely π_A when $L\sqrt{1-\varepsilon^2} > 3$, uniquely π_B when it is below 3, and both partitions tie at equality. For $L = 8$ and $0 < \varepsilon \le 1/4$, π_A uniquely minimizes true KL, with $C_B/C_A \ge 8/5$; for $\varepsilon \le 1/8$ this improves to $20/7$.

At each fixed finite L, $\lim_{\varepsilon\downarrow 0} C_B/C_A = (L+1)^2/16$. At fixed $0 < \varepsilon < 1/2$, $C_B > 0$ is independent of L and $C_A = \Theta(L^{-2})$. Hence the unweighted selector's multiplicative loss diverges as $L \to \infty$, while its additive regret is bounded by C_B and tends to C_B. The target condition number diverges in this limit.

Proof. The construction has lower-triangular factor

$$\begin{pmatrix} 1 & 0 & 0 \\ \varepsilon & \sqrt{1-\varepsilon^2} & 0 \\ a & 2\varepsilon\tau & \tau\sqrt{1-4\varepsilon^2} \end{pmatrix}.$$

Its diagonal is positive and every row has squared norm one. For π_A, the nonsingleton Schur complement after X_1 is

$$K_A = \begin{pmatrix} 1-\varepsilon^2 & 2\varepsilon\tau\sqrt{1-\varepsilon^2} \\ 2\varepsilon\tau\sqrt{1-\varepsilon^2} & \tau^2 \end{pmatrix},$$

whose scales are $(\sqrt{1-\varepsilon^2}, \tau)$ and correlation is 2ε. For π_B, it is $K_B = \left(\begin{smallmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{smallmatrix}\right)$. Singleton costs vanish, so Theorem B.1 gives $C_A = c(K_A)$ and $C_B = c(K_B)$. The unweighted scores are $4\varepsilon^2$ and ε^2. Substituting into (38) gives (47) and all weighted strict/tied cases.

For $L = 8$ and $\varepsilon \le 1/4$, $\sqrt{1-\varepsilon^2} \ge 7/8$ gives $P_B/P_A \ge 4$ and $w_A = 1/(8\sqrt{1-\varepsilon^2}+1) \le 1/8$. The correlation of K_A is at most $1/2$, and therefore

$$\Gamma(w_A, 2\varepsilon) \le e^{1/12}\frac{20}{9} \le \frac{80}{33} < \frac{5}{2}.$$

Proposition C.2 yields $C_B/C_A \geq P_B/((5/2)P_A) \geq 8/5$. If $\varepsilon \leq 1/8$, its $7/5$ bound gives $C_B/C_A \geq 20/7$. Both costs are positive, and these are the only two consecutive two-batch partitions; hence the optimum is unique. The pair's uniform relative expansion gives $C_A/P_A \to 1$ and $C_B/P_B \to 1$ as $\varepsilon \downarrow 0$, proving the small correlation ratio.

At fixed positive ε, K_B is constant and $C_B \geq \varepsilon^2/8 > 0$. Meanwhile $P_A = \Theta(L^{-2})$, and the pair bound gives the same order for C_A because $\Gamma(w_A, 2\varepsilon) \leq \Gamma(1, 2\varepsilon) < \infty$. Thus π_A is optimal for sufficiently large L, $C_B/\min\{C_A, C_B\} \to \infty$, and $0 \leq C_B - \min\{C_A, C_B\} \leq C_B$ with limit C_B. Finally $\mathrm{Var}(X_3 - aX_1) = \tau^2$, so $\lambda_{\min}(\Sigma) \leq \tau^2/(1+a^2) \to 0$, whereas $\lambda_{\max}(\Sigma) \geq \mathrm{tr}(\Sigma)/3 = 1$. This proves the condition-number statement. □

D Proof of the complete partition guarantee

We prove Theorem 4.3. The boundary variance created by conditioning stays in the finite transport expansion. A lower bound on the true adjacent transport error then turns the resulting interval widths into a complete-decision bound. All norms without a subscript are operator 2-norms, and all guarantees below use exact arithmetic.

Removing the common scalar transport and changing its clock converts the original local flow into an algebraic expansion. For $M = M_B$, set

$$\begin{aligned}(P_n)_{j,1:j} &= M_{j,1:j}(M_{1:j,1:j})^{n-1}, \qquad J_0 = I,\\ J_n &= \frac{1}{2n}\sum_{\ell=1}^{n}(-1)^{\ell-1}P_\ell J_{n-\ell}, \qquad U_r = \mathrm{chol}(I+M)^{-1}\sum_{n=0}^{r}J_n.\end{aligned} \tag{48}$$

Each row uses its own causal prefix and includes the boundary diagonal. All candidates with the same start are leading blocks of one suffix calculation.

The finite cost intervals. Put $\beta_* = (1-\sqrt{1-4\rho^2})/2$, $h = 2\rho$, and $\ell = \sqrt{\rho^2+\beta_*^2}$. For $\rho > 0$, a scalar recurrence supplies a common tail for the entire table:

$$\zeta_0 = 1, \quad \zeta_n = h\frac{n-1+\ell/(2h)}{n}\zeta_{n-1}, \qquad \|U(K_B) - U_r\|_2 \leq \delta_r = \frac{\zeta_{r+1}}{(1-h)\sqrt{1-2\rho}}. \tag{49}$$

Valid candidate-specific norm bounds can sharpen this envelope.

To obtain KL endpoints, let $k = |B|$, $t_r = \|\mathrm{tril}_< U_r\|_F$, $\epsilon_r = \sqrt{k}\delta_r$, and $\psi(v) = (v^2-1)/2 - \log v$. The actual relative factor has diagonal at least one, where ψ is increasing. Consequently,

$$\begin{aligned}\underline{c}_r^{\mathrm{raw}}(B) &= \tfrac{1}{2}(t_r - \epsilon_r)_+^2 + \sum_j \psi\big(\max\{1, (U_r)_{jj} - \delta_r\}\big),\\ \overline{c}_r(B) &= \tfrac{1}{2}(t_r + \epsilon_r)^2 + \sum_j \psi\big((U_r)_{jj} + \delta_r\big).\end{aligned} \tag{50}$$

The true cost also obeys $c(K_B) \geq (19/160)\sum_{i=a+1}^{b-1} e_i^2$. Use the larger of this analytic lower bound and $\underline{c}_r^{\mathrm{raw}}(B)$ as $\underline{c}_r(B)$. Singletons have exact cost zero and receive $[0,0]$.

Schur closure and the original flow. Let e_i be the target's edges, $B = (a, b]$, and $k = b - a$. The global Cholesky factor divided by s is bidiagonal, with pivots

$$d_1 = 1, \qquad d_j = 1 - e_{j-1}^2/d_{j-1}, \qquad d_* := \frac{1+\sqrt{1-4\rho^2}}{2} \leq d_j \leq 1. \tag{51}$$

The lower bound follows by induction from $d_* = 1 - \rho^2/d_*$. For $\rho \leq 1/5$, $d_* \geq 19/20$ and $1 - d_j \leq (20/19)\rho^2$. Conditioning on $1{:}a$ retains the Cholesky block on B: if $K_B = s^2(I+M)$ and $C = \mathrm{chol}(I+M)$, then

$$C_{jj} = \sqrt{d_{a+j}}, \qquad C_{j,j-1} = e_{a+j-1}/\sqrt{d_{a+j-1}} \quad (j \geq 2). \tag{52}$$

Thus M has first diagonal $-\beta_a = -(1 - d_{a+1})$, other diagonal entries zero, and unchanged internal edges. Further leading conditioning gives the same global pivot sequence. This covers every

candidate, independently of how its completed prefix was partitioned. Also $0 \le \beta_a \le \beta_* := 1 - d_*$ and $\lambda_{\min}(I + M) \ge 1 - 2\rho$: the global covariance divided by s^2 has this lower bound, which every Schur complement inherits.

For the original local equation $F_K' = L_K F_K$, $F_K(0) = I$, the prefix row of L_K is $u(K-I)_{j,1:j}[I_j + u^2(K-I)_{1:j,1:j}]^{-1}$. Set $g^2 = 1 - u^2 + u^2 s^2$, $x = u^2 s^2/g^2$, and $F_K(u) = g(u)V(x)$. Using $I + u^2(K-I) = g^2(I + xM)$ and $s^2 - (s^2-1)x = s^2/g^2$ gives, for each prefix,

$$[L_K - (g'/g)I]_{j,1:j} = \frac{us^2}{g^4} M_{j,1:j}(I_j + xM_{1:j,1:j})^{-1}, \qquad \frac{dx}{du} = \frac{2us^2}{g^4}. \tag{53}$$

Consequently $V' = A(x)V$, $V(0) = I$, with prefix row $A_{j,1:j} = M_{j,1:j}(I_j + xM_{1:j,1:j})^{-1}/2$. These identities extend continuously to zero. The row includes the negative boundary diagonal. At the endpoint $F_K(1) = sV(1)$, so the actual relative factor is $U = C^{-1}V(1)$. This removes one common scalar transport and preserves the conditional variance. By Theorem 4.1, the resulting costs add to the original sampler's endpoint KL, including at incorrectly generated prefix values.

Bounding the truncated expansion. Replace M by zM. Expanding each prefix inverse gives generator coefficient $(-1)^{n-1}x^{n-1}P_n/2$ at order z^n, with P_n as in (48). Substitution of $V_z(x) = \sum_{n\ge 0} z^n x^n J_n$ gives the stated recurrence. Write $W_r = \sum_{n=0}^{r} J_n$, so $U_r = C^{-1}W_r$. Let $N = |M|$ entrywise and $B_0 = \mathrm{tril}_{\le} N$. Allowing paths outside the causal prefix gives $|P_n| \le B_0 N^{n-1}$ entrywise. The columns of B_0 have disjoint supports: its first contains β_a and the first edge; each other nonzero column contains one edge. Therefore the full candidate table admits the common bounds

$$\|N\| \le h := 2\rho < 1, \qquad \|B_0\| \le \ell := \sqrt{\rho^2 + \beta_*^2} \le h, \qquad \alpha = \frac{\ell}{2h} \le \tfrac{1}{2}. \tag{54}$$

The first bound also follows from absolute row sums, using $\beta_* \le \rho$. Norm monotonicity for nonnegative matrices gives $\|P_n\| \le \ell h^{n-1}$; no norm contraction of triangular projection is assumed. The majorant in (49) satisfies

$$\begin{gathered}(1-hz)^{-\alpha} = \sum_{n\ge 0} \zeta_n z^n, \qquad n\zeta_n = \tfrac{1}{2}\sum_{m=1}^{n} \ell h^{m-1}\zeta_{n-m}, \\ \|V(1) - W_r\| \le \sum_{n>r} \zeta_n \le \frac{\zeta_{r+1}}{1-h}.\end{gathered} \tag{55}$$

Induction in (48) bounds $\|J_n\|$ by ζ_n. Uniform convergence for $|z| < 1/h$ permits termwise substitution; uniqueness identifies the series with V_z. Successive positive coefficient ratios are at most h, which bounds the omitted sum as displayed. Since $\|C^{-1}\| \le (1-2\rho)^{-1/2}$, this proves the factor error δ_r in (49). Prefix consistency of the recurrence and Cholesky solve allows one suffix calculation per start for all its terminal intervals.

A lower bound on the true cost. Write the local edges as $w_j = e_{a+j}$. The pivots of $I + xM$ satisfy

$$q_1 = 1 - x\beta_a, \qquad q_j = 1 - x^2 w_{j-1}^2/q_{j-1}, \qquad 0 < q_j \le 1, \quad q_j' \le 0, \quad q_j - xq_j' \ge 1. \tag{56}$$

Positivity follows from the covariance. The derivative bounds follow inductively; for $j \ge 2$, the last expression is $1 + x^2 w_{j-1}^2/q_{j-1} - x^3 w_{j-1}^2 q_{j-1}'/q_{j-1}^2 \ge 1$, and for $j = 1$ it equals one. The last inverse-prefix diagonal and subdiagonal are $1/q_j$ and $-xw_{j-1}/(q_j q_{j-1})$, respectively. Thus

$$\begin{gathered}A_{jj} = (1 - q_j^{-1})/(2x) \le 0, \qquad A_{j,j-1} = w_{j-1}/(2q_j q_{j-1}), \\ \frac{d}{dx}\log\frac{V_{jj}}{\sqrt{q_j}} = \frac{xw_{j-1}^2}{2q_j q_{j-1}} - \frac{x^2 w_{j-1}^2 q_{j-1}'}{2q_j q_{j-1}^2} \ge 0 \quad (j \ge 2).\end{gathered} \tag{57}$$

The log derivative is zero for $j = 1$. Hence $\sqrt{q_j} \le V_{jj} \le 1$ throughout and $U_{jj} \ge 1$ at the endpoint.

For $j \ge 2$ put $z_j = V_{j,j-1}/V_{j-1,j-1} - xw_{j-1}/q_{j-1}$. Only rows j and $j-1$ enter column $j-1$ of the triangular flow; substitution of (57) gives

$$z_j' = (A_{jj} - A_{j-1,j-1})z_j + \frac{w_{j-1}}{q_{j-1}^2}(\tfrac{1}{2} - q_{j-1} + xq_{j-1}'), \qquad z_j(0) = 0. \tag{58}$$

The source has sign opposite to w_{j-1} and magnitude at least $|w_{j-1}|/2$. Its integrating kernel obeys

$$\exp\left(\int_t^1 (A_{jj} - A_{j-1,j-1})\,dx\right) \geq \frac{V_{jj}(1)}{V_{jj}(t)} \geq V_{jj}(1).$$

There is no sign cancellation, so $|z_j(1)| \geq |w_{j-1}|V_{jj}(1)/2$. The bidiagonal solve then yields the bound on the actual factor,

$$\begin{aligned} |U_{j,j-1}| &= \frac{V_{j-1,j-1}(1)}{\sqrt{q_j(1)}}|z_j(1)| \geq \tfrac{1}{2}|w_{j-1}|\sqrt{d_{a+j-1}}, \\ c(K_B) &\geq \tfrac{1}{2}\sum_{j=2}^{k} U_{j,j-1}^2 \geq q(B) := \frac{19}{160}\sum_{i=a+1}^{b-1} e_i^2. \end{aligned} \tag{59}$$

The remaining KL summands are nonnegative. For a singleton, the scalar flow is $\sqrt{1-u^2+u^2s^2d_{a+1}}$; its cost is exactly zero, and its interval is assigned $[0,0]$ without truncation.

Cost enclosures and their width. For $k \geq 2$, retain t_r, ϵ_r, ψ defined before (50) and put $l_j = \max\{1, (U_r)_{jj} - \delta_r\}$, $u_j = (U_r)_{jj} + \delta_r$. The actual cost is $\frac{1}{2}\|\operatorname{tril}_< U\|_F^2 + \sum_j \psi(U_{jj})$. Frobenius error is at most $\sqrt{k}\delta_r$ and each diagonal error is at most δ_r. Since $U_{jj} \geq 1$ and ψ increases on $[1,\infty)$, these facts prove both raw endpoints in (50). The arguments u_j are at least one because they bound U_{jj}. Taking the larger lower endpoint is valid by (59).

For their width, $\beta_* \leq (20/19)\rho^2$ gives $\ell \leq 21\rho/20$. The logarithm of the scalar majorant at one is at most $\ell/[2(1-h)] \leq 7\rho/8$. Using $e^y - 1 \leq y/(1-y)$ yields $\|W_r - I\| \leq 35\rho/33$ for every $r \geq 0$. The bidiagonal structure and $\sqrt{d_j} \geq 19/20$ also give

$$\begin{aligned} \|C - I\| &\leq \frac{400}{741}\rho^2 + \frac{20}{19}\rho < \frac{6}{5}\rho, \\ \|U_r - I\| &\leq \|C^{-1}\|(\|W_r - I\| + \|C - I\|) < \frac{13}{10}\left(\frac{35}{33} + \frac{6}{5}\right)\rho < 3\rho. \end{aligned} \tag{60}$$

Here $\|C^{-1}\| \leq \sqrt{5/3} < 13/10$. The lower-triangle interval width is at most $2t_r\epsilon_r + \epsilon_r^2$, including when $t_r < \epsilon_r$, and hence at most $k(6\rho\delta_r + \delta_r^2)$. Each diagonal interval has length at most $2\delta_r$ and lies below $1 + 3\rho + \delta_r$. Because $\psi'(v) \leq 2(v-1)$ on $[1,\infty)$, each contributes at most $12\rho\delta_r + 4\delta_r^2$. The lower clipping only reduces width. Thus

$$0 \leq \bar{c}_r(B) - \underline{c}_r(B) \leq k\omega_r, \qquad \omega_r = 18\rho\delta_r + 5\delta_r^2. \tag{61}$$

From interval bounds to a complete partition. Let $L_r(\pi), U_r^{\mathrm{cost}}(\pi)$ be the sums of the interval endpoints and V_L, V_U their minima. If $\widehat{\pi}$ minimizes the upper sum and $C_* = \min_\pi C(\pi)$, additivity gives $V_L \leq C_* \leq C(\widehat{\pi}) \leq V_U$. Each D-batch partition has $H - D$ internal edges and $\sum_{B:|B|\geq 2}|B| \leq 2\sum_B(|B|-1) = 2(H-D)$. Applying (61) to a lower-table minimizer π_L therefore bounds the computed certificate:

$$\begin{aligned} R_r = V_U - V_L \leq U_r^{\mathrm{cost}}(\pi_L) - L_r(\pi_L) \leq E_r := 2(H-D)\omega_r, \\ V_L \geq \frac{19}{160}\kappa^2\rho^2(H-D) \qquad (D < H). \end{aligned} \tag{62}$$

The second line uses $|e_i| \geq \kappa\rho$ and the analytic lower clipping. These bounds prove (16). Their ratio gives

$$\frac{R_r}{V_L} \leq \frac{320}{19\kappa^2}\left(18\frac{\delta_r}{\rho} + 5\frac{\delta_r^2}{\rho^2}\right). \tag{63}$$

This controls the complete computed relative certificate, including when D is close to H.

Finally $\alpha \leq 21/80 < 1/3$. If $a_n = [z^n](1-z)^{-1/3}$, coefficientwise monotonicity gives $\zeta_n \leq a_n h^n$. The a_n decrease for $n \geq 1$, and $a_8 = \prod_{j=1}^{8}(j-2/3)/j = 5434/59049 < 1/10$. Using $a_{13} \leq a_8$,

$(1-h)^{-1} \le 5/3$, and $\|C^{-1}\| < 13/10$ in (55) gives $\delta_{12} \le (13/60)(2\rho)^{13}$. Consequently

$$\frac{\delta_{12}}{\rho} \le \frac{13}{30}\left(\frac{2}{5}\right)^{12} = \frac{53248}{7324218750} < \frac{1}{125000}, \qquad \frac{1280}{19}\left(\frac{18}{125000} + \frac{5}{125000^2}\right) < \frac{1}{100}. \tag{64}$$

Thus $\kappa \ge 1/2$ gives $R_{12} < V_L/100$ and $C(\widehat{\pi}) \le 1.01 C_*$. When $D = H$, the sole partition consists of exact singletons, so $C_* = V_L = V_U = R_r = 0$ without relative division.

Refinement, positive gaps, and ties. The coefficient ratios also imply $\delta_r \le \ell h^r/[2(1-h)\sqrt{1-2\rho}]$. For fixed H, D, ρ, the bound E_r therefore tends geometrically to zero, so every positive additive regret tolerance is certified at finite order, without a minimum edge assumption. With $V_L > 0$, stopping at $V_U - V_L \le \eta V_L$ certifies relative regret at most η; at $V_L = 0$, use $V_U - V_L \le \varepsilon_{\mathrm{abs}}$ for a prescribed absolute tolerance $\varepsilon_{\mathrm{abs}} > 0$. For uniqueness, let $L_{\mathrm{alt}}(\widehat{\pi}) = \min_{\pi \ne \widehat{\pi}} L_r(\pi)$, with value $+\infty$ if there is no alternative. The strict test $U_r^{\mathrm{cost}}(\widehat{\pi}) < L_{\mathrm{alt}}(\widehat{\pi})$ certifies the unique optimum. If the true optimum π_* has gap $\Delta > 0$, every partition interval has width at most E_r, and $2E_r < \Delta$ implies

$$U_r^{\mathrm{cost}}(\pi_*) \le C_* + E_r < C_* + \Delta - E_r \le L_r(\pi) \qquad (\pi \ne \pi_*).$$

The upper-table minimizer is then π_* and passes the strict test. An exact tie cannot be certified as unique; it does not obstruct a positive-tolerance regret certificate. The minima use the standard additive interval DP, and exclusion of one partition can be tracked by a flag recording a different cut. The guarantees above apply to the population costs defined by the supplied covariance.

E Finite-order reliability of conditional batch costs

We analyze the same synchronous Gaussian population field as in the batch-cost theorem. All matrix 2-norms below are operator norms. Write $K = S(I+E)S$, where $S = \mathrm{diag}(s_i)$ is positive and E is symmetric with zero diagonal, and put $C_K = \mathrm{chol}(I+E)$. Define

$$g_i(u) = \sqrt{1 - u^2 + u^2 s_i^2}, \quad a_i(u) = \frac{u s_i}{g_i(u)}, \quad a_i'(u) = \frac{s_i}{g_i(u)^3}, \quad G_u = \mathrm{diag}(g_i(u)). \tag{65}$$

The normalized flow is $V(u) = G_u^{-1} F_K(u)$, so $U(K) = C_K^{-1} V(1)$ and

$$c(K) = \tfrac{1}{2}\|\operatorname{tril}_< U(K)\|_F^2 + \sum_i \psi(U_{ii}(K)), \qquad \psi(v) = \tfrac{1}{2}(v^2 - 1) - \log v. \tag{66}$$

We use $U_{ii}(K) \ge 1$, including at unequal scales. Indeed, with $q_j(u) = \mathrm{Var}(X_j \mid Y_{<j}(u))$ and $d_j(u) = 1 + u^2(q_j(u) - 1)$, Gaussian conditioning gives

$$q_j' = -2u\|[I + u^2(K_{<j,<j} - I)]^{-1} K_{<j,j}\|^2 \le 0, \qquad \frac{d}{du}\log\frac{[F_K(u)]_{jj}}{\sqrt{d_j(u)}} = -\frac{u^2 q_j'}{2 d_j} \ge 0.$$

The ratio starts at one and ends at $U_{jj}(K)$.

E.1 Finite expansion and a cost interval at arbitrary scales

Replace E by zE and denote the normalized flow by V_z. Let $D_a = \mathrm{diag}(a_i(u))$, $R_u = D_a E D_a$, and define the strictly lower matrix $B_{ji} = a_j' a_i E_{ji}$ for $i < j$. The exact normalized generator satisfies

$$V_z' = A_z V_z, \quad V_z(0) = I, \qquad [A_z]_{j,1:j} = z B_{j,1:j}[I_j + z(R_u)_{1:j,1:j}]^{-1}. \tag{67}$$

To verify this directly, the interpolation covariance is $G_u(I + zR_u)G_u$. In the original local regression, $[L_{K_z}]_{j,1:j} = u(K_z - I)_{j,1:j}[I + u^2(K_z - I)]^{-1}_{1:j,1:j}$. Multiply $G_u^{-1} L_{K_z} G_u - G_u^{-1} G_u'$ on its row prefix by $I_j + z(R_u)_{1:j,1:j}$. Its diagonal is zero; its $i < j$ entry is $z u s_j s_i E_{ji}/(g_j^3 g_i) = z a_j' a_i E_{ji}$. This proves (67), with the required prefix inverse.

Expanding that inverse defines the homogeneous coefficients

$$[A_n]_{j,1:j} = (-1)^{n-1} B_{j,1:j}(R_u)^{n-1}_{1:j,1:j}, \quad V_0(u) = I, \quad V_n(u) = \int_0^u \sum_{\ell=1}^{n} A_\ell(t) V_{n-\ell}(t)\, dt. \tag{68}$$

Set $W_r = \sum_{n=0}^{r} V_n(1)$ and $U_r = C_K^{-1} W_r$. The first coefficient is $(V_1(1))_{ji} = s_i E_{ji}/(s_i + s_j)$. For $r \geq 1$, subtracting the first Cholesky variation yields $(U_r - I)_{ji} = -s_j E_{ji}/(s_i + s_j) + O(E^2)$; the leading cost is $p(K)$.

Choose nonnegative measurable envelopes b_0, h_0, with $\int_0^1 b_0 < \infty$ and $\sup_u h_0(u) < 1$, such that

$$\| \, |B(u)| \, \|_2 \leq b_0(u), \qquad \|D_a |E| D_a\|_2 \leq h_0(u). \tag{69}$$

For example, the computable bounds

$$b_0 = \sqrt{\left(\max_j a'_j \sum_{i<j} a_i |E_{ji}| \right) \left(\max_i a_i \sum_{j>i} a'_j |E_{ji}| \right)}, \qquad h_0 = \max_i a_i \sum_j a_j |E_{ij}|$$

retain the actual conditional scales and satisfy $\sup_u h_0(u) < 1$ when $\max_i \sum_j |E_{ij}| < 1$. Sharper spectral envelopes are allowed; they exist whenever $\| \, |E| \, \|_2 < 1$. Define

$$\begin{gathered} \beta_n = \int_0^1 b_0(u) h_0(u)^{n-1} \, du, \qquad \mathcal{G}(z) = \exp\!\left(\int_0^1 \frac{z b_0(u)}{1 - z h_0(u)} \, du \right) = \sum_{n \geq 0} \zeta_n z^n, \\ \zeta_0 = 1, \quad n\zeta_n = \sum_{\ell=1}^{n} \ell \beta_\ell \zeta_{n-\ell}, \qquad \mathcal{T}_r = \mathcal{G}(1) - \sum_{n=0}^{r} \zeta_n. \end{gathered} \tag{70}$$

Theorem E.1 (Finite-order population-cost interval)**.** Under (69), $\|V(1) - W_r\|_2 \leq \mathcal{T}_r$ and $\|U(K) - U_r\|_2 \leq \delta_r = \mathcal{T}_r / \sqrt{\lambda_{\min}(I + E)}$. Let $t_r = \| \operatorname{tril}_< U_r \|_F$, $\epsilon_r = \sqrt{k}\delta_r$, and $\ell_i = \max\{1, (U_r)_{ii} - \delta_r\}$, $u_i = (U_r)_{ii} + \delta_r$. Then

$$\underbrace{\tfrac{1}{2}(t_r - \epsilon_r)_+^2 + \sum_i \psi(\ell_i)}_{\underline{c}_r(K)} \leq c(K) \leq \underbrace{\tfrac{1}{2}(t_r + \epsilon_r)^2 + \sum_i \psi(u_i)}_{\overline{c}_r(K)}. \tag{71}$$

Proof. Taking absolute values and allowing additional, non-prefix paths gives $|A_n(u)| \leq |B(u)|(D_a|E|D_a)^{n-1}$ entrywise. Operator norm is monotone under entrywise order of nonnegative matrices, since $|Mx| \leq M|x|$. Thus $\|A_n(u)\|_2 \leq b_0 h_0^{n-1}$; no contraction of triangular projection in operator norm is assumed. Let $\zeta_n(u)$ be the coefficients in (70) with upper integration endpoint u. They satisfy $\zeta'_n = \sum_{\ell=1}^{n} b_0 h_0^{\ell-1} \zeta_{n-\ell}$. Induction in (68) gives $\|V_n(u)\|_2 \leq \zeta_n(u)$. The envelope assumptions ensure uniform convergence on a complex disk of radius greater than one. Termwise integration then shows that the series solves (67); uniqueness identifies it with V_z. Summing the omitted terms proves the first bound, and multiplication by $\|C_K^{-1}\|_2 = \lambda_{\min}(I + E)^{-1/2}$ proves the second. The Frobenius error is at most $\sqrt{k}\delta_r$; every diagonal error is at most δ_r. Strict-lower projection contracts Frobenius norm, and ψ increases on $[1, \infty)$. Together with $U_{ii}(K) \geq 1$, these facts prove (71). □

When U_r has positive diagonal, define the corrected cost $c_r = \frac{1}{2}t_r^2 + \sum_i \psi((U_r)_{ii})$. Otherwise the interval remains valid and its midpoint can serve as the estimate. The construction is an exact-arithmetic statement. A certified coefficient error $\|\overline{W}_r - W_r\|_2 \leq \nu_r$ is included by replacing $\mathcal{T}_r$ with $\mathcal{T}_r + \nu_r$; rounding and solve errors also need enclosure.

E.2 An algebraic correction at equal scales

If $s_i = s > 0$, use the clock $x = a_i(u)^2$. Let $L = \operatorname{tril}_<(E)$ and

$$(P_n)_{j,1:j} = L_{j,1:j}(E_{1:j,1:j})^{n-1}, \quad J_0 = I, \qquad J_n = \frac{1}{2n} \sum_{\ell=1}^{n} (-1)^{\ell-1} P_\ell J_{n-\ell}. \tag{72}$$

Indeed, $A_n(x) = (-1)^{n-1} x^{n-1} P_n/2$, so substitution in (68) gives $V_n(x) = x^n J_n$ and $W_r = \sum_{n=0}^{r} J_n$. These finite matrix operations require no time quadrature. Choose $h \geq \| \, |E| \, \|_2$ and $\ell \geq \| \operatorname{tril}_< |E| \|_2$, with $0 < \ell \leq h < 1$. The envelopes $b_0 = \ell/2$, $h_0 = xh$ give

$$\mathcal{G}(z) = (1 - hz)^{-\ell/(2h)}, \qquad \zeta_n = h \frac{n - 1 + \ell/(2h)}{n} \zeta_{n-1}, \qquad \mathcal{T}_r \leq \frac{\zeta_{r+1}}{1 - h}. \tag{73}$$

The last inequality follows from successive coefficient ratios at most h. One may take row/column norm bounds and reduce ℓ to h if necessary, since entrywise monotonicity also bounds the strict lower part by h. For $E = 0$, the cost is exactly zero. Dense coefficient construction costs $O(r^2k^3)$, followed by whitening; band structure limits the traversed paths. Section 5.2 reports complete preprocessing times with suffix reuse for both the finite method and the ODE reference.

F FINITE SAMPLERS AND ENDPOINT ERROR

F.1 ENDPOINT ERROR DECOMPOSITION

The preceding costs use the optimal causal predictor with exact integration. Fitting and numerical integration can reinforce or partly cancel its structural error. For a fitted velocity linear in the state, the finite-step output remains Gaussian.

For a partition π, write the population and actual endpoints as $\mu + TU_\pi\xi$ and $\widehat{\mu}_\pi + \widehat{F}_\pi\xi$, where $T = \mathrm{chol}(\Sigma)$ and $\xi \sim \mathcal{N}(0, I)$. Assume $\widehat{F}_\pi$ is lower triangular with positive diagonal. Applying T^{-1} puts both maps in coordinates where the target has unit covariance. Their difference and the relative changes of their diagonal entries are

$$\begin{aligned} D_\pi = T^{-1}\widehat{F}_\pi - U_\pi, \quad b_\pi = T^{-1}(\widehat{\mu}_\pi - \mu), \quad u_{\pi i} = (U_\pi)_{ii}, \\ \eta_{\pi i} = (D_\pi)_{ii}/u_{\pi i}, \qquad A_\pi = U_\pi - \mathrm{diag}(u_{\pi i}^{-1}). \end{aligned} \tag{74}$$

Although U_π is block diagonal, the endpoint error D_π may couple different batches.

Proposition F.1 (Signed endpoint error)**.** Let C_π and $\widehat{C}_\pi$ be the population and actual generated-to-target KL. Then

$$\widehat{C}_\pi - C_\pi = \langle A_\pi, D_\pi\rangle_F + R_\pi, \qquad R_\pi = \tfrac{1}{2}(\|D_\pi\|_F^2 + \|b_\pi\|^2) + \sum_i [\eta_{\pi i} - \log(1+\eta_{\pi i})] \geq 0. \tag{75}$$

Proof. Suppress the partition subscript in (74) and write $F = TU$ for the population factor. The actual factor is lower triangular with positive diagonal, so $1 + \eta_i = \widehat{F}_{ii}/F_{ii} > 0$. Whitening the target and expanding the Gaussian generated-to-target KL gives

$$\begin{aligned} \widehat{C} - C &= \langle U, D\rangle_F + \tfrac{1}{2}(\|D\|_F^2 + \|b\|^2) - \sum_i \log(1+\eta_i) \\ &= \langle A, D\rangle_F + \tfrac{1}{2}(\|D\|_F^2 + \|b\|^2) + \sum_i h(\eta_i), \qquad h(z) = z - \log(1+z) \geq 0. \end{aligned} \tag{76}$$

The second equality uses $\langle U, D\rangle_F = \langle A, D\rangle_F + \sum_i \eta_i$, which proves the proposition. The whitened mean error contributes only $\|b\|^2/2$, since the initial noise is centered. □

The signed term measures cancellation or reinforcement of structural error, while the remainder is nonnegative. For two partitions w, c, the actual advantage is exactly

$$\widehat{C}_c - \widehat{C}_w = C_c - C_w + \langle A_c, D_c\rangle_F - \langle A_w, D_w\rangle_F + R_c - R_w. \tag{77}$$

In particular, $C_c - C_w > \|A_c\|_F\|D_c\|_F + \|A_w\|_F\|D_w\|_F + R_w - R_c$ suffices to preserve the population ordering. This comparison is conditional on the realized fits and partitions, so it permits their estimation from the same data.

For fixed-step Runge–Kutta integration of a causal affine field, $\widehat{F}$ is lower triangular and can be obtained by propagating the initial basis and subtracting the zero-input trajectory. Finite integration preserves triangularity but does not ensure a positive diagonal. A nonsingular factor with a negative diagonal still has finite Gaussian KL computed with $\log|\det \widehat{F}|$, but the stated $h(\eta_i)$ formula requires positive diagonals. A singular factor gives infinite KL to P.

Fitting and integration. Let $\widehat{F}_{\mathrm{fine}}$ be a refined integration reference for the same fitted field. Define

$$D^{\mathrm{fit}} = T^{-1}(\widehat{F}_{\mathrm{fine}} - F), \qquad D^{\mathrm{num}} = T^{-1}(\widehat{F} - \widehat{F}_{\mathrm{fine}}).$$

Then $D = D^{\mathrm{fit}} + D^{\mathrm{num}}$. The signed term in (76) adds across these two errors. The quadratic term includes $\langle D^{\mathrm{fit}}, D^{\mathrm{num}}\rangle_F$, and the diagonal remainder is evaluated at $\eta_i = (D^{\mathrm{fit}}_{ii} + D^{\mathrm{num}}_{ii})/u_i$. Thus separate nonnegative fitting and solver penalties do not recover the total KL change. The refined solution remains a numerical reference whose accuracy must be checked; the same applies when the population factor F is computed numerically.

F.2 Changing the fitted field can change exactness

A two-coordinate example shows why reducing prediction error need not reduce generation error. Let X be centered Gaussian with unit variances and correlation $\rho = 3/5$, and let $m_i(q, y) = \mathbb{E}[\partial_{q_i} Y_i \mid q_{1:i}, Y_{1:i} = y_{1:i}]$ for the training path in (1). For uniform noise levels q independent of the data and corruption noise, conditional-expectation orthogonality gives

$$\frac{\mathcal{R}(\lambda m) - \mathcal{R}(m)}{\mathbb{E}\|m\|^2} = (\lambda - 1)^2, \qquad \mathcal{R}(f) = \mathbb{E}\sum_i |f_i - \partial_{q_i} Y_i|^2.$$

Thus m minimizes prediction risk. On the synchronous schedule, use the clock $x = \sin^2(\pi q/2)$. Gaussian regression gives the sampler coefficient $\lambda A(x)$, where

$$A(x) = \frac{1}{2(1-\rho^2 x^2)} \begin{pmatrix} 0 & 0 \\ \rho & -\rho^2 x \end{pmatrix}, \qquad T_x = \begin{pmatrix} 1 & 0 \\ \rho x & \sqrt{1-\rho^2 x^2} \end{pmatrix}. \tag{78}$$

Direct differentiation of this triangular transport (Spantini et al., 2018) gives $T_x' T_x^{-1} = 2A(x)$ and $T_0 = I$. Hence $\lambda = 2$ generates the target exactly. For $\lambda = 1$, Theorem 3.1 implies an inexact endpoint because the correlated pair is denoised synchronously. Moving from $2m$ to m therefore minimizes prediction risk while increasing endpoint KL. This changes the prediction field; the exactness theorem keeps that field fixed.

F.3 Finite-step batch costs before choosing a partition

Proposition 4.4 uses batch maps computed before a complete partition is selected. Let the fixed field be $f_i(q, x) = r_i(q_{1:i}) + \ell_i(q_{1:i})^{\mathsf{T}} x_{1:i}$. During a batch the completed prefix is frozen, active noise levels follow the prescribed common grid, and later coordinates stay inactive. A fixed explicit Runge–Kutta method composes affine lower-triangular maps, proving (17). Earlier batches depend on earlier initial noises, which are independent of ξ_B. A singular N_B yields a singular output and infinite generated-to-target KL; finite costs below use nonsingular maps.

For a fixed start a, compute the map of the whole suffix $(a, H]$. Its first $b - a$ outputs give the map for $(a, b]$: at every solver substage, rows up to b see the same state and noise-level prefixes, regardless of the later active coordinates. Induction over the substages proves this restriction at the actual finite grid. This reuse assumes that the grid is state independent and the field has no additional batch-index input.

Cost and generated-prefix moments. Write $h_B(x) = h_{B,0} + H_B x$ and $k = |B|$. Gaussian conditional KL gives

$$\kappa_B = \tfrac{1}{2}\left\{ \mathrm{tr}(K_B^{-1}\widehat{K}_B) - k + \log \frac{\det K_B}{\det \widehat{K}_B} \right\}, \qquad \mathrm{KL}(Q_{\pi,B|x} \| P_{B|x}) = \kappa_B + \tfrac{1}{2} e_B(x)^{\mathsf{T}} K_B^{-1} e_B(x). \tag{79}$$

If the incoming generated prefix has mean m_J and covariance S_J, put $a_B = d_B - h_{B,0}$ and $D_B = G_B - H_B$. Its expected mean penalty is

$$\tfrac{1}{2}\left\{ (a_B + D_B m_J)^{\mathsf{T}} K_B^{-1} (a_B + D_B m_J) + \mathrm{tr}(K_B^{-1} D_B S_J D_B^{\mathsf{T}}) \right\}. \tag{80}$$

The finite sampler propagates these moments by

$$m_B = d_B + G_B m_J, \qquad S_{B,J} = G_B S_J, \qquad S_{B,B} = G_B S_J G_B^{\mathsf{T}} + \widehat{K}_B, \tag{81}$$

retaining the existing prefix blocks. The first batch uses the empty prefix and no regression term.

Proof of Proposition 4.4. Apply the KL chain rule to (79), taking each expectation under $Q_{\pi,J}$. This gives (18); (81) follows from independence of the fresh noise. Covariance KL and expected squared conditional-mean error are nonnegative, so dropping the latter gives a lower cost for every partition. Minimizing those lower costs yields $V_\kappa \le C_*$. For the returned minimizer, $C(\pi_\kappa) = V_\kappa + P_{\pi_\kappa}$, proving the gap bound. □

More generally, any selected partition has additive gap at most $C(\pi) - V_\kappa$. When $V_\kappa > 0$, a bound $C(\pi) - V_\kappa \le \eta V_\kappa$ implies $C(\pi) \le (1+\eta)C_*$. When it is zero, the additive bound still applies.

G Conditional scales in Gaussian experiments

G.1 Standardized targets and training-based partitions

The controlled targets have $H = 12$ and four consecutive triples. For group $g \in \{0,1,2,3\}$, let $a = 3g$ and $\sigma_g = (-1)^g$. Construct a lower-triangular matrix B with diagonal blocks

$$B_{a+1:a+3,a+1:a+3} = \begin{pmatrix} 1 & 0 & 0 \\ \sigma_g\varepsilon & \sqrt{1-\varepsilon^2} & 0 \\ \sqrt{1-L^{-2}} & 2\sigma_g\varepsilon/L & \sqrt{1-4\varepsilon^2}/L \end{pmatrix}. \tag{82}$$

For $g > 0$, set $B_{a+1:a+3,1:a} = \nu Z_g/\sqrt{a}$, where the $3 \times a$ matrices Z_g have independent standard Gaussian entries, drawn successively with seed 331. All other entries are zero. Normalize each row: with $D_{ii} = \|B_{i,:}\|_2$, set $T_* = D^{-1}B$ and $X = T_*G$, $G \sim \mathcal{N}(0, I)$. Then $\Sigma = T_*T_*^\mathsf{T}$ is positive definite with unit diagonal for $0 < \varepsilon < 1/2$ and finite $L \ge 1$. The cross-group entries introduce residual dependence while row normalization preserves global standardization.

The learning comparison fixes $\nu = .1$ and four targets: $(\varepsilon, L) = (.30, 2), (.30, 4), (.30, 8), (.45, 4)$. These span weak decision separation, two larger conditional-scale contrasts, and a setting where the leading proxy can fail. The target set was fixed before the learning comparison.

Each target has three repetitions $s \in \{0,1,2\}$, each with 32,768 independent training observations and 8,192 independent validation observations. The mean is known to be zero. The same training second moment $\widehat{\Sigma} = n^{-1}\sum_{r=1}^n x_r x_r^\mathsf{T}$ determines both the learned field and the interval costs. From every interval's Schur complement K, the three scores are $\sum_{i<j}\rho_{ij}^2$ (conditional correlation), $\frac{1}{2}\sum_{i<j}(s_j/(s_i+s_j))^2\rho_{ij}^2$ (weighted), and $c(K)$ (exact Gaussian flow cost), with $s_i = \sqrt{K_{ii}}$ and $\rho_{ij} = K_{ij}/(s_i s_j)$. Interval dynamic programming selects partitions at depths $D \in \{2,3,4,6\}$. The exact score describes the Gaussian population flow for $\widehat{\Sigma}$. Equal batches use cuts $\lfloor jH/D \rfloor$, $j = 1, \dots, D$; sequential and synchronous schedules complete one coordinate and all coordinates per stage, respectively. Identical cut sequences share an evaluation.

G.2 A learned field with exactly evaluable output KL

The learned field is linear in the noisy state, $f_{\theta,i}(y,q) = \sum_{j\le i} L_{\theta,ij}(q_{1:i})y_j$. A shared MLP has two SiLU hidden layers of width 192 and H outputs. For row i, its inputs are the padded prefix $q_{1:i}$, its sine and cosine features at frequencies $1, 2, 4, 8$ times π, and a length-H one-hot position vector. Every future-position input and output coefficient is masked, giving input dimension $10H$.

Gaussian corruption noise can be integrated out of its empirical flow-matching risk. For a fixed noise-level vector q, write $a_i = \sin(\pi q_i/2)$, $b_i = \cos(\pi q_i/2)$, $Y = D_a X + D_b G$, and $V_i = \partial_{q_i} Y_i = (\pi/2)(b_i X_i - a_i G_i)$. With X uniform over the training data and $D_{ab} = \mathrm{diag}(a_i b_i)$,

$$\begin{aligned} R_q &= \mathbb{E}[YY^\mathsf{T}] = D_a\widehat{\Sigma}D_a + D_b^2, \qquad & C_q &= \mathbb{E}[VY^\mathsf{T}] = \tfrac{\pi}{2}(D_b\widehat{\Sigma}D_a - D_{ab}), \\ \ell_i^* &= (C_q)_{i,1:i}(R_q)^{-1}_{1:i,1:i}, & \mathcal{L}(\theta) &= \frac{1}{H}\mathbb{E}_q\sum_i(\ell_{\theta,i} - \ell_i^*)R_q(\ell_{\theta,i} - \ell_i^*)^\mathsf{T}. \end{aligned} \tag{83}$$

Rows in the loss are padded with zeros. Completing the square shows that $H^{-1}\mathbb{E}\|f_\theta(Y,q) - V\|_2^2$ equals $\mathcal{L}(\theta)$ plus a parameter-independent constant. Thus the loss uses the original finite training samples through their second moments, with their corruption noise integrated out.

We fix 8,192 training and 2,048 held-out noise-level vectors. Their mixture consists of independent uniform levels (30%), one shared uniform level (15%), a uniformly chosen stage of a uniformly chosen candidate schedule (30%), and a random consecutive interval (25%). In the last two cases, the active interval shares a uniform level, its prefix is clean, and its suffix is pure noise. Random intervals choose the preceding cut uniformly from $0{:}H-1$ and the end uniformly from the remaining positions. Candidate schedules are the twelve method–depth combinations above plus sequential and synchronous, including repeated cut sequences. The held-out noise loss uses the training second moment; the independent validation data serve the distributional budget below.

AdamW uses batch size 256, weight decay 10^{-6}, gradient clipping at 10, and 20,000 updates. The learning rate is initially 10^{-3}, then 2×10^{-4} after 14,000 updates and 5×10^{-5} after 18,000. Each repetition independently initializes and trains the model. All schedules share its final checkpoint.

Sampling, budgets, and endpoint attribution. For each partition, explicit midpoint uses total NFE $N\in\{48,96,192,384\}$, allocating $N/(2D)$ steps to each active stage. In that stage, q increases linearly from zero to one and only active rows move. If $A(q)$ denotes the masked network coefficient matrix and $h=2D/N$, the actual generated factor obeys

$$F_{r+1}=F_r+hA(q_{r+1/2})\big(F_r+\tfrac{h}{2}A(q_r)F_r\big),\qquad F_0=I. \tag{84}$$

The numerical output law is therefore $\mathcal{N}(0,F_{\mathrm{end}}F_{\mathrm{end}}^{\mathsf{T}})$, evaluated against the true target through the Gaussian KL formula. Network coefficients are queried in float32 and factors propagated in float64. The primary comparison is depth six at NFE 192, with paired weighted-minus-correlation KL differences within each repetition. Reported dispersion is the sample standard deviation across the three independently trained models.

For the auxiliary budget comparison, each policy searches its depths and NFE values using KL to $\mathcal{N}(0,\widehat{\Sigma}_{\mathrm{val}})$, where $\widehat{\Sigma}_{\mathrm{val}}$ is the independent validation second moment. Budgets are .05 and .10. Selection minimizes NFE among passing configurations, breaking ties by validation KL; final acceptance uses the same budget against the true Σ. Sampling NFE counts field evaluations and excludes cost preprocessing. In this comparison, correlation and weighting both choose NFE 96 at $L=4$ and 192 at $L=8$.

The population and empirical-covariance reference flows use DOP853 with relative/absolute tolerances $10^{-10}/10^{-12}$; exact interval costs use $10^{-9}/10^{-11}$. The learned reference uses 256 midpoint steps per active coordinate, with 128-to-256 refinement measured as $\|T^{-1}(F_{256}-F_{128})\|_{\mathrm{F}}$, $T=\mathrm{chol}(\Sigma)$. In the endpoint KL decomposition, the displacement from the population factor to F_{256} contains data and field-approximation error plus reference discretization; the displacement from F_{256} to the tested factor measures the additional integration change. Their quadratic expansion includes the cross term between these two displacements. The 128-to-256 comparison records the sensitivity to the numerical reference. The factor decomposition is applied when its diagonal is positive.

G.3 Conditional scales and reversing the causal order

Table 3: **Conditional scales affect the learned output.** Actual KL at six batches and NFE 192, mean ± SD over three data/model repetitions. The paired difference is weighted minus correlation; negative is better. 4^{R} reverses the $L=4$ target, preserving its global spectrum.

ε	L	Correlation	Weighted	Gaussian cost	Equal	Paired difference
.30	2	.0424±.0017	.0424±.0017	.0424±.0017	.0895±.0021	0
.30	4	.0436±.0016	.0331±.0005	.0334±.0011	.0485±.0015	−.0105±.0013
.30	8	.0453±.0018	.0136±.0012	.0132±.0008	.0362±.0017	−.0317±.0008
.45	4	.1065±.0031	.1128±.0026	.1066±.0030	.1424±.0042	+.0063±.0029
.30	4^{R}	.1726±.0158	.0958±.0017	.0958±.0017	.1478±.0014	−.0768±.0141

The control uses the standardized target $(\varepsilon,L,\nu)=(.30,4,.1)$ from Section G.1, and compares its original order with one prespecified full coordinate reversal. If P is the reversal matrix, the reverse covariance is $P\Sigma P^{\mathsf{T}}$. Each original finite training and validation array is generated using the original Cholesky factor and random stream, then its columns are reversed. Thus each pair has exactly the

same samples up to permutation. The global spectrum, unit marginal variances, and distances between sample points are preserved.

The three forward models from Table 3 are reused. Three reverse models use the same architecture, optimization, 20,000 updates, seeds $0, 1, 2$, and noise-mixture protocol in Section G.2. All covariance-based cuts are recomputed in the new causal order. Within an order and repetition, policies share the final checkpoint. The candidate-schedule component of the training mixture is retained in both orders. Depth six and NFE 192 are primary, with the same $48, 96, 192, 384$ NFE grid as an integration control.

Table 4: **Spectrum alone does not determine selection effects.** Mean population KL of the data-selected depth-six cuts and the paired weighted-minus-correlation differences. Actual differences use the trained models at NFE 192; dispersion is the sample standard deviation over the three repetitions.

Order	Pop. corr.	Pop. weighted	Pop. difference	Actual difference
Forward	.04194	.03275	$-.00918 \pm .00026$	$-.01052 \pm .00132$
Reverse	.17136	.09572	$-.07565 \pm .01386$	$-.07685 \pm .01413$

Weighting improves the primary actual KL in all three repetitions of each order, but by substantially different amounts (Table 4). In the reverse order it selects the Gaussian-cost cuts in all three runs. This fixed-spectrum comparison shows that global conditioning alone is insufficient to determine the effect of the selection rule.

G.4 Population and estimated-cost selection

These comparisons use the population causal regression, with the interval scores and dynamic program defined in Section G.1. Evaluation divides the selected true endpoint KL by the optimum at the same depth when the latter exceeds 10^{-12}.

Dense targets. For $H = 12$, construct $\Sigma = D_L(I + \varepsilon R_s)D_L$. Draw a Gaussian matrix, symmetrize it, set its diagonal to zero, and divide by its operator norm to obtain R_s. Target seeds are $s = 0{:}9$. The diagonal of D_L contains L^{z_i}, where z is a permutation of twelve equally spaced points in $[-1, 1]$, using seed $8301 + s$. The grid is $\varepsilon \in \{.05, .2, .5, .8\}$, $L \in \{1, 2, 4, 8\}$, and depths $D \in \{2, 3, 4, 6\}$. It has 160 covariances and 640 paired target–depth decisions. Here L varies marginal variances; the learning targets in Section G.1 have unit marginal variances. Each seed reuses its correlation direction and scale permutation across strengths, scales, and depths. Weighting selects an optimum in 581 of the 640 target–depth decisions; correlation has lower cost than weighting in 17.

Across the 160 dense targets and four depths, mean selected-KL/optimal-KL ratios are 1.009 for weighting, 1.788 for correlation and 3.831 for equal batches.

Finite covariance estimation. Use $\varepsilon \in \{.2, .5\}$, $L \in \{1, 4\}$, and target seeds $0{:}4$, giving twenty targets. For each target, draw $n \in \{128, 512, 2048\}$ observations in three repetitions and form $\widehat{\Sigma} = .99S + .01\,\mathrm{diag}(S)$ from the unbiased sample covariance S. All estimated costs use this same matrix; evaluation uses the true covariance and its population field. Standard normal draws are reused across sample sizes and target parameters, with three independent draws at each fixed setting. The 720 decisions per method form a matched grid. Table 5 reports the complete comparison.

Cost construction. One suffix Schur covariance and one synchronous suffix-flow solve supply all intervals with the same start. Exact costs use float64 DOP853 with relative/absolute tolerances $10^{-9}/10^{-11}$.

Weighting by conditional scale improves population KL over correlation at larger scale contrasts. With estimated covariance, larger samples bring both rules closer to the population optimum.

Table 5: **Estimated costs evaluated on the true population flow.** Mean true KL divided by the optimum at the same depth. Each row averages two strengths, five target seeds, three draws, and four depths. "Gaussian cost" integrates the flow for the estimated covariance; the population oracle has ratio one.

L	n	Correlation	Weighted	Gaussian cost	Equal
1	128	2.221	2.236	2.217	1.846
1	512	1.592	1.552	1.556	1.846
1	2048	1.206	1.214	1.211	1.846
4	128	5.310	2.571	2.586	4.680
4	512	3.355	1.874	1.884	4.680
4	2048	2.516	1.345	1.354	4.680

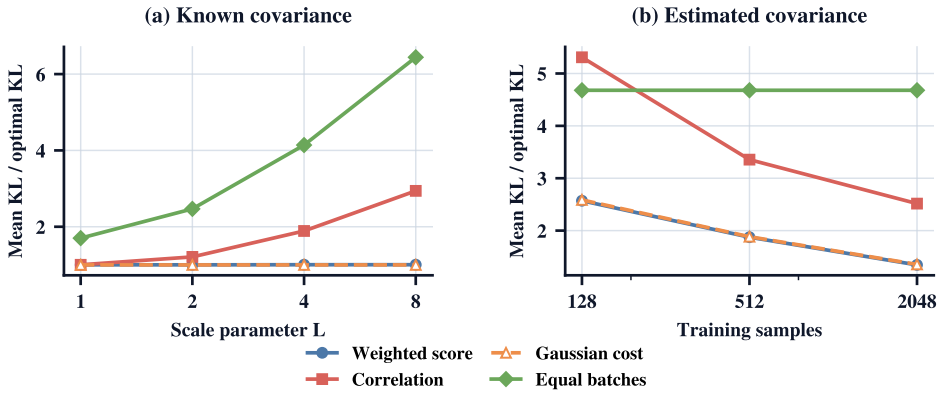


Figure 4: **Conditional scales affect batch selection with known and estimated covariance.** True population KL relative to the optimal partition at the same depth. The left panel varies marginal scale at correlation strength .5; the right varies estimation sample size at scale $L = 4$. Both panels execute the population field; Appendix G.4 specifies the paired averages.

H RELIABLE PARTITION SELECTION

Fixed targets and decisions. The grid uses the covariance in (15) with $s = 1$, $H \in \{16, 32, 64\}$, and $\rho \in \{.1, .2\}$. For each pair, one edge profile has $e_i = \rho$ and three use independent uniform draws in $[\rho/2, \rho]$ from NumPy's default generator with seeds $11, 22, 33$. Each of the 24 covariances is evaluated at $D = H/8, H/4, H/2$. All methods receive the same known covariance, isolating selection from estimation and learning error.

Finite tables and refinement. Global pivots d_j are computed once and retained for every Schur complement. One coefficient recurrence per start produces its entire suffix; every shorter candidate is a leading block. Lower-order coefficients are reused during refinement. For each candidate, we choose h_B as the maximum absolute row sum of M_B and $\ell_B = \| \operatorname{tril}_{\leq} |M_B| \|_2$, evaluated from its orthogonal column supports. The tail uses these valid local bounds and whitening denominator $\sqrt{1 - h_B}$. The analytic lower cost $19 \sum e_i^2/160$ is combined with the raw interval as in Theorem 4.3. Singleton costs are exactly zero.

The method starts with order four for every suffix and computes lower and upper DPs for all three depths. It stops when all have $V_U - V_L \leq .01 V_L + 10^{-12}$. The small additive tolerance handles zero costs; all reported relative gaps are themselves below .01. Otherwise the next order is taken from $6, 8, 12, 16$. Refinement includes uncertain nonsingleton edges on the upper-cost incumbent and on competing paths whose lower total is below that incumbent's upper total. Forward and backward DPs identify these paths; updates are shared across depths. After the order cap, the predefined fallback computes an original-ODE table as a numerical reference without an interval certificate. None of the 72 decisions uses the fallback. A separate full order-12 calculation checks the uniform sufficient order; its extra work is excluded from the adaptive method's timing and choices.

For uniqueness, a second DP tracks whether any cut differs from the selected complete sequence and minimizes lower cost over the changed state. Only a strict selected-upper versus alternative-lower comparison passes. Empty alternatives have cost $+\infty$. This test is separate from the primary 1% criterion: 38 of 72 adaptive decisions pass it, and the remaining 34 retain the reported relative guarantee.

Numerical reference. The reference integrates the original population flow in its signal clock using float64 DOP853, relative tolerance 10^{-10} and absolute tolerance 10^{-12}. It also uses one ODE per suffix and one Cholesky whitening per suffix, with cumulative row costs supplying all leading intervals. Thus the comparator receives the same reuse advantage as the finite method. Equal batches, the conditional-correlation score, the weighted score, and the finite point cost are each passed to the same complete partition DP.

All 21,952 reference candidate costs lie inside the adaptive nominal intervals. Tightening the reference tolerances to $10^{-12}/10^{-14}$ changes checked costs by at most 2.27×10^{-13}. At fixed order 12, reference costs fall below lower interval endpoints by at most 3.73×10^{-14}, within this numerical resolution. The theorem applies in exact arithmetic; these float64 calculations use nominal intervals without directed rounding.

Complete decision accuracy. The first orders passing 1% are four for 42 decisions, six for 24, and eight for six. At termination, the largest computed relative gap is .5926%, while the fixed order-12 maximum is .00297%. Table 6 reports actual selected-cost regret against the numerical optimum. The rule minimizing upper costs retains a selection guarantee; the finite point cost reaches the optimum to numerical resolution in all cases. On the 54 nonuniform-profile decisions, the interval rule reaches an optimum in 52, with maximum regret .0207%; the weighted and correlation rules reach 36 and 40, respectively, each with maximum regret .8354%. Their larger 6.77% maxima on the full grid occur in the uniform controls.

Table 6: **Actual regret on all 72 complete decisions.** Relative regret is $(C(\pi) - C_*)/C_*$. Numerical optimality uses an absolute KL tolerance of 10^{-9} and can include different tied cuts.

Rule	Mean relative regret	Maximum relative regret	Numerically optimal
Upper-cost intervals	.000768%	.0207%	61/72
Finite point cost	$< 10^{-9}$%	$< 10^{-9}$%	72/72
Weighted score	.5315%	6.7703%	36/72
Correlation score	.5012%	6.7703%	40/72
Equal batches	17.5944%	58.3794%	18/72

Preprocessing time. After warmup, each target has three paired timings with alternating method order, using two CPU workers and one BLAS/OpenMP thread per worker. Both methods share one suffix-based cost table across the three depths. Timings include complete table construction, whitening, all DPs and, for finite intervals, every refinement and competitor check. The speed ratio is the median of per-target ratios of median times (Table 7). Sampling and training are excluded.

Table 7: **Complete preprocessing for three depths.** Seconds are medians across the eight fixed targets at each width. The ratio column reports median and range of their paired ratios.

H	Finite intervals (s)	Original ODE (s)	ODE/finite ratio
16	.0268	.1654	6.48 [2.95, 8.73]
32	.1593	.6096	4.38 [1.77, 5.97]
64	1.0242	3.7546	4.07 [1.42, 6.12]

The finite implementation uses dense matrix arithmetic and a standard partition DP at the stated widths.

I SELECTION FOR TRAINED DENOISERS

I.1 SELECTION USING THE LEARNED FINITE-STEP MODEL

This experiment reuses the four targets and twelve final coefficient networks from Sections G.1 and G.2. At six batches, NFE 96 and 192 mean eight and sixteen explicit-midpoint steps per batch. Each interval map is computed from the trained noise-conditioned coefficients at exactly this grid. All methods share the checkpoint; the validation reference is the known-zero-mean second moment $n^{-1}\sum_i x_i x_i^\mathsf{T}$ of 8,192 independent validation observations. Gaussian-cost and weighted selection retain the training-covariance inputs of the original comparison.

We compute κ_B for every legal interval, solve the six-batch dynamic program, and evaluate the selected proposal by the generated-prefix recurrence (81). Enumerating all $\binom{11}{5} = 462$ partitions provides evaluation optima for the supplied reference and the true target; these rankings are not selector inputs. A subsequent scan of single-boundary moves accepts no change in these cases, so the reported choices coincide with the initial DP proposals. They attain the optimum for their supplied validation reference in all 24 cases. Their true-target performance is reported in Table 8.

Table 8: **Actual learned-model selection across the complete budget grid.** Relative excess true-target KL (%) above the best partition for that model and budget, averaged over three data/model repetitions per target. The same models are used at both budgets.

ε	L	NFE	Gaussian cost	Weighted	Learned cost
.30	2	96	2.09	2.09	.72
.30	2	192	2.07	2.07	.73
.30	4	96	3.51	2.56	.00
.30	4	192	2.74	1.69	.20
.30	8	96	.86	3.17	.00
.30	8	192	2.76	5.65	.64
.45	4	96	81.52	2.52	.18
.45	4	192	1.68	7.68	2.19

The mean and maximum individual true-target regrets are $.583\%$ and 4.117%. The stronger-correlation target shows the largest change with budget. At NFE 96, learned cost and weighting both outperform Gaussian cost; at NFE 192, Gaussian cost narrowly leads (Table 8). Each comparison holds the model fixed and includes its actual fitting and integration errors.

Cost of selection. Suffix reuse supplies all shorter intervals with the same start. Including model queries, stage construction, conditional preparation and selection, median preprocessing is 8.82ms, with range 6.98–15.47ms on one GPU and its host CPU at horizon twelve. Model loading is excluded. Selected-partition moment propagation evaluates full cost without constructing all candidate endpoints.

I.2 How integration changes the ranking

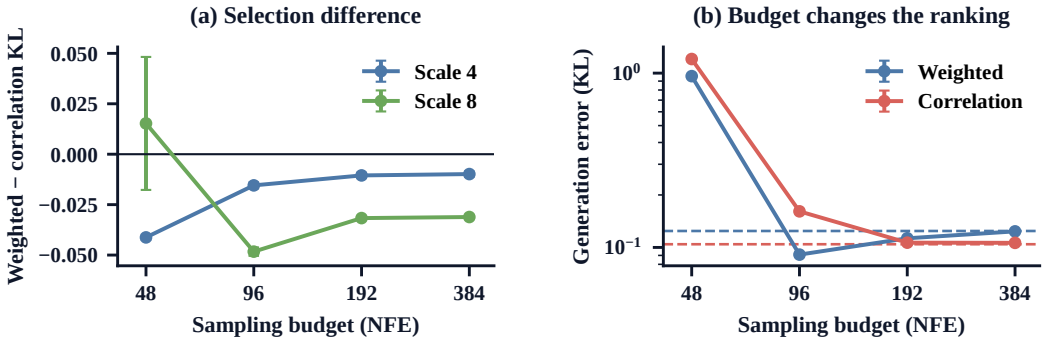


Figure 5: **The decision advantage and its numerical cancellation.** All curves use depth six and the same final learned models as Table 3; bars show sample standard deviations across three repetitions. (a) Weighted-minus-correlation KL across the fixed NFE grid. (b) In the stronger-correlation setting, coarse integration hides the weighted rule's population disadvantage. Dashed lines show the mean continuous population KL for each method's data-selected cuts. Finer integration removes much of the cancellation.

Figure 5(a) holds the model and partitions fixed while changing the solver budget. The signed decomposition (75) explains the change in ranking. In the stronger-correlation setting, the weighted partition has mean population KL .1243, compared with correlation's .1043. At NFE 192, its negative interaction term of $-.0142$ dominates the quadratic term .0020 and diagonal remainder .0007, leaving actual KL .1128. This cancellation hides more than half of the weighted partition's structural disadvantage. At a coarser budget it can reverse the ranking, as Figure 5(b) shows.

A refined integration of the same learned field attributes most of that negative interaction to the solver displacement ($-.01360$), with a smaller contribution from fitting ($-.00055$). Both partially cancel the population error. The reference uses 256 midpoint steps per active coordinate; for the primary weighted and correlation schedules, doubling from 128 changes the target-whitened factor by at most 3.59×10^{-4} in Frobenius norm across the four targets and three repetitions. The exact finite-step identity itself does not depend on this refined learned reference. All evaluated factors had positive diagonals.

I.3 A nonlinear denoiser on a Gaussian target

Model and evaluation. The target is $(\varepsilon, L, \nu) = (.30, 4, .1)$ from Appendix G.1. The causal Transformer has two pre-normalized attention blocks, width 96, four heads, and width-192 SiLU feed-forward layers. Tokens embed the noisy value, noise level, eight Fourier features and a learned position; a normalized linear head predicts velocity. Three independent repetitions each use 32,768 training and 8,192 validation observations. Training uses sampled flow-matching labels V_i from Appendix G.2, 20,000 AdamW updates, batch 256, learning rate 3×10^{-4}, weight decay 10^{-4} and gradient clipping at one. Noise levels mix synchronous, sequential, equal four-stage and independent profiles with probabilities .2, .2, .2, .4.

All partition rules share each final model and use midpoint at NFE 48 and 192. Each repetition shares 2,048 initial noises, 2,048 independent target samples and 256 normalized Gaussian projection directions across methods. Sliced W_2 compares sorted projections; relative covariance error is $\|\widehat{S} - \widehat{\Sigma}_{\mathrm{ref}}\|_{\mathrm{F}} / \|\widehat{\Sigma}_{\mathrm{ref}}\|_{\mathrm{F}}$. Here $\widehat{S}$ and $\widehat{\Sigma}_{\mathrm{ref}}$ are the unbiased sample covariances of the generated and reference target samples, respectively. Whitened sliced W_2 applies the true target Cholesky inverse to both sample sets.

Results. At NFE 192, weighting changes the correlation-based cuts in every repetition and lowers unwhitened sliced W_2 and covariance error in all three (Table 9). Paired differences are $-.0067 \pm .0024$ and $-.0520 \pm .0087$; the same directions hold at NFE 48. The whitened difference is $-.0009 \pm .0063$.

Table 9: **Transfer to a causal Transformer.** Mean ± sample standard deviation over three repetitions at NFE 192 on the scale-4 target. Sliced W_2 uses 2,048 generated and 2,048 independent target samples; whitening uses the true target covariance for evaluation. Sequential and synchronous schedules have depths twelve and one; the other methods have depth six. Gaussian cost uses the training covariance to select the partition.

Partition	Sliced W_2	Relative covariance error	Whitened sliced W_2
Correlation	$.0692 \pm .0067$	$.1499 \pm .0136$	$.1980 \pm .0367$
Weighted	$.0625 \pm .0081$	$.0979 \pm .0102$	$.1971 \pm .0384$
Gaussian cost	$.0625 \pm .0081$	$.0983 \pm .0101$	$.1994 \pm .0406$
Equal	$.0669 \pm .0077$	$.1353 \pm .0162$	$.1986 \pm .0349$
Sequential	$.0628 \pm .0084$	$.0965 \pm .0127$	$.1794 \pm .0341$
Synchronous	$.0977 \pm .0011$	$.2859 \pm .0031$	$.9523 \pm .0763$

J Batch selection on real sensor signals

J.1 Shared data, models and evaluation

Prediction task and data split. We use publicly released UCI HAPT accelerometer recordings (CC BY 4.0) (Reyes-Ortiz et al., 2016), without activity labels or activity filtering. Each example takes 256 consecutive raw observations of the first acceleration axis, originally sampled at 50 Hz, and retains every fourth value without filtering. The first 32 retained values are observed; the next 32, covering 2.56 seconds, are generated jointly. Training uses subjects 1–21 except 5, 12, and 19; these three are validation subjects, and subjects 22–30 are held out for testing. Training windows advance by 64 raw observations; validation and test windows advance by 256. The resulting splits contain 10,326, 393, and 1,364 windows. Global mean and standard deviation are estimated from raw training points, counting each point once, and applied to all splits.

Gaussian references. Let c be the observed 32-vector and y its continuation. A linear predictor $cB + b$ is fitted on training windows, with ridge $10^{-3}\,\mathrm{tr}(C_{cc})/32$, where C_{cc} is the unbiased sample covariance of training contexts. The global covariance uses training residuals $y - cB - b$. Covariances use $.99S + .01\,\mathrm{diag}(S)$, where S is the unbiased residual covariance.

Architecture and training. The conditional velocity Transformer has 157,153 parameters, with two blocks, width 96, four attention heads, and a feedforward expansion of two with SiLU activations. Each token supplies its value, noise level, observed flag, and sine/cosine noise features at frequencies $1, 2, 4, 8$ times π; positions have learned embeddings. All 32 observed tokens stay fixed. Training uses the sine/cosine path in (1), minibatches of 128, AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}, gradient norm clipping at one, and the final checkpoint after 20,000 updates.

The causal training mixture is 15% synchronous, 15% sequential, 40% uniformly random consecutive partitions, and 30% independent uniform levels. Random partitions are independent of the data and of the tested selection rules. Their active stage is uniform, with a clean prefix and pure-noise suffix. Configuration choices use validation data.

Conditional distribution scores. For a future $y \in \mathbb{R}^H$ and M generated draws $z_1, \ldots, z_M$, the primary energy score (Gneiting & Raftery, 2007) is

$$\mathrm{ES} = \frac{1}{\sqrt{H}} \left[\frac{1}{M} \sum_m \|z_m - y\|_2 - \frac{1}{M(M-1)} \sum_{m<n} \|z_m - z_n\|_2 \right]. \tag{85}$$

The ensemble estimate excludes diagonal pairs. CRPS replaces the Euclidean norm and its normalization by coordinate-averaged absolute differences. The whitened energy score applies the lower Cholesky inverse of the global training residual covariance to both draws and target. The variogram score (Scheuerer & Hamill, 2015) compares $|y_i - y_j|^{1/2}$ with its ensemble expectation, squares the difference, and averages lags one through eight with inverse-lag weights. The latter uses a finite-ensemble plug-in expectation. These scores evaluate actual conditional samples; their values are distinct from the Gaussian flow KL.

Training and inference use GPUs; covariance calculations and statistical aggregation use CPUs.

J.2 Response-based selection at NFE 32 and 64

Models and data for response-based selection. Three independently trained causal Transformers use the shared setup. Future attention logits are masked to $-\infty$; the partition component of the training mixture chooses depths two, four and eight uniformly. All selectors share each final model. Selection fixes eight batches; NFE 32 and 64 give two and four midpoint steps per batch. Each model and NFE uses a fixed pool of 32 partitions, shared by all selectors. Ordinary search uses 72 validation prefixes, balanced across its three subjects; 96 separate prefixes provide development evaluation, excluding same-recording windows within 256 raw steps of search prefixes.

Training references and stage maps. Two fixed anchors per training subject give 36 contexts. For each c, the 256 nearest training prefixes supply $P_c = \mathcal{N}(\mu_c, K_c)$: the mean is the fitted linear prediction plus mean neighboring residual, and $K_c = .99S + .01\,\mathrm{diag}(S)$ uses centered residual covariance. At every batch start, freeze the preceding future positions at their reference means and initialize the active suffix with zero noise. Central differences of size .01 estimate responses of its finite midpoint endpoint to each input. Causality permits restriction of whole-suffix responses to shorter intervals. In residual coordinates these define the surrogate

$$\widetilde{Y}_B = d_B + M_B\widetilde{Y}_J + N_B\xi_B, \quad X = \mu_c + \widetilde{Y}, \quad \xi_B \sim \mathcal{N}(0, I). \tag{86}$$

Here d_B is the stage endpoint minus $\mu_{c,B}$ at the anchor; M_B, N_B are prefix/noise responses. Recurrence (81), with $G_B = M_B$, composes the residual mean u_π and covariance C_π. The stages are anchored separately, so this Gaussian surrogate is assessed against actual nonlinear samples in the experiments.

Candidate pool. The ordered pool starts with equal batches and seven random partitions. Next come partitions selected by global correlation, weighted and Gaussian costs, followed by correlation and weighted partitions whose interval costs are averaged over 18 training anchors. We sample one prefix per training subject and use the references and stage maps above. After eight random partitions, we add two partitions minimizing the summed costs κ_B in (79) and $\frac{1}{2}d_B^{\mathsf{T}}K_B^{-1}d_B$, respectively, averaged over these anchors. We keep the first occurrence of each cut sequence and append random partitions until the pool contains 32 distinct partitions. Each random partition uniformly samples seven distinct cuts from $\{1, \ldots, 31\}$.

Expected-energy selection. Let $\Psi(a, S) = \mathbb{E}\|Z\|_2$ for $Z \sim \mathcal{N}(a, S)$. The expected energy of the surrogate against its Gaussian reference is

$$E_c(\pi) = \frac{\Psi(u_\pi, C_\pi + K_c) - \frac{1}{2}\Psi(0, 2C_\pi)}{\sqrt{H}}. \tag{87}$$

The two terms use the Gaussian laws of surrogate–reference and surrogate–surrogate differences. We average scores over the 36 anchors and minimize over complete candidates, since expected energy is generally nonadditive. The square-root identity and Gaussian Laplace transform give

$$\Psi(a, S) = \frac{1}{2\sqrt{\pi}} \int_0^\infty \frac{1 - \det(I + 2tS)^{-1/2} \exp\{-ta^{\mathsf{T}}(I + 2tS)^{-1}a\}}{t^{3/2}}\, dt. \tag{88}$$

We normalize by the root second moment, substitute $t = (v/(1-v))^2$, and use fixed 64-point Gauss–Legendre quadrature on $(0, 1)$ in double precision. Singular positive-semidefinite covariances are allowed. The fixed-covariance version sets $C_\pi = K_c$, retaining the same propagated mean; the KL control uses the same maps and references.

Search and cost. Full validation search generates 32 draws per selection prefix and candidate, minimizing subject-balanced energy. A preliminary comparison on 96 validation prefixes, with 32 draws per prefix, fixes a starting partition for each model and NFE. It minimizes energy averaged equally over subjects among equal batches, covariance rules and the eight random partitions above. Similar-time search starts with this partition, followed by equal batches and the remaining pool in its original order, removing duplicates. It takes the longest prefix whose generation/scoring cost fits

the response selector's measured necessary time, with a five-candidate floor: seven candidates at NFE 32 and five at NFE 64. Budgeting uses timing only. Necessary response-selection cost includes reference fitting/loading, new responses, candidate composition/scoring and selection. On the same hardware, this cost is 3.701s versus full search's 15.479s at NFE 32, and 5.079s versus 30.863s at NFE 64. Model loading and generation after selection are excluded; the confirmation uses these timings.

Held-out confirmation. Following development at NFE 32 and 64, the NFE-32 selections were fixed before reading reserved future segments. The NFE-64 development comparison gives .413% higher energy than full search. The confirmation uses all 158 reserved segments from subjects 22–30, excluding prior queries and same-recording neighbors within 256 raw steps. Each model uses 32 fresh-noise continuations per context; identical partitions share samples. Scores balance subjects, then models. Response-based selection lowers mean energy by .223% relative to full search. Mean whitened energy, CRPS and variogram scores are .318%, .788% and .420% lower, respectively.

Table 2 reports the energy comparisons.

J.3 A LARGER CANDIDATE POOL AND A TRAJECTORY VARIANT

This comparison retains the three response-selection models, NFE 32, eight batches, references, 72 selection prefixes and 96 development prefixes from Section J.2. Random partitions expand the original ordered pool to 256, discarding duplicates. All rules receive the response selector's choice from the original pool of 32 partitions as their incumbent. This is development data, separate from the 158-segment confirmation. The trajectory variant runs each complete nonlinear partition from zero noise at each training anchor and scores its output with the fixed-covariance version of (87). It constructs no response matrices; its deterministic output is not the nonlinear distribution's mean. Both approaches pay for all required computations.

As a refinement check, the incumbent and four highest-ranked alternatives are evaluated on the 72 selection prefixes with 32 draws each. This leaves response-based choices unchanged.

Table 10: **Selection from 256 candidate partitions.** Subject/model-balanced energy on the 96-context development set; necessary selection seconds include all method-specific work. All rules search the same 256 candidates with the same three models at NFE 32. The first row is the already available incumbent.

Selection rule	Energy	Seconds
Incumbent	.35241	.000
Response-based selection	.35006	6.379
Zero-noise trajectories	.35025	3.171
Full validation search	.35328	125.033

K ADDITIONAL RELATED WORK

Information access and training objectives. Diffusion Forcing provides likelihood lower bounds for independently noised positions (Chen et al., 2024). Flow matching establishes distribution paths using the full state (Lipman et al., 2023); FloodDiffusion uses the full active window and identifies distributional problems under causal restrictions (Cai et al., 2026). Causal Forcing studies a fixed positive-noise prefix kernel at a clean-prefix value (Proposition 3.4 and Appendix B.2), alongside autoregressive distillation (Zhu et al., 2026a). PCD aligns pretraining with clean-prefix continuation by changing attention, corruption, and supervision (Lu et al., 2026). Jacobi Forcing progressively distills causal language models on their own parallel-decoding trajectories (Hu et al., 2025). Flux Matching matches probability-flux divergence to accommodate directed fields (Pao-Huang et al., 2026). Our fixed Gaussian regression receives each available position's noise level. We characterize its terminal law under continuous monotone schedules; changing the objective can change the field.

Block generation and within-block information. Block Diffusion combines autoregression across blocks with discrete diffusion within blocks (Arriola et al., 2025). FLARE adapts block diffusion

to hybrid-attention language models with a causal clean stream and bidirectional attention within noisy blocks (Zhu et al., 2026c). AURORA-LM generates continuous text latents blockwise, using the current noisy block and clean prefix (Liang et al., 2026). These architectures impose causality across blocks; our analysis fixes causality at every position within a block.

Choosing positions to generate together. ARMD uses a permutation-equivariant causal architecture and strided groups, motivated by approximate conditional independence of distant tokens (Karami & Ghodsi, 2026). Dynamic Chunking learns content-dependent groups and attention patterns (Zhu et al., 2026b); Learn2PD learns a filter for predictions to finalize together (Bao et al., 2026); A3 adapts autoregressive models to arbitrary groups and orders (Du et al., 2026). DEMASK selects tokens with a learned dependency predictor and, under sub-additivity, bounds a parallel update's deviation from the model joint distribution (Ringel et al., 2026). Mean-Field Parallel Decoding coordinates token commitments through pairwise interactions from predictive distributions, without retraining (Zoabi et al., 2026). Our endpoint question concerns the data distribution under a fixed continuous causal field. Under its Gaussian assumptions, Theorem 3.1 rules out full cancellation of local incompatibility later. Consecutive batches preserve order and a completed prefix, making the local residual flow independent of prefix values and enabling Theorem 4.1 and interval optimization.

Verification and parallelizing denoising steps. Speculative decoding preserves a target autoregressive model's output distribution through acceptance and residual resampling (Leviathan et al., 2023). S2D2 verifies diffusion proposals with a block-diffusion model in autoregressive mode, with routing to decide when verification is worthwhile (Han et al., 2026). ParaDiGMS uses Picard refinement to parallelize denoising time steps (Shih et al., 2023); our batches parallelize sequence positions.