DEPENDENCE COST OF INDEPENDENT UPDATES IN MASKED DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Parallel token generation makes masked diffusion a useful approach to language modeling with fewer sequential sampling steps. However, sampling tokens independently can miss dependencies, and errors in the sampling trajectory can differ from errors in the final output distribution. We study progressive unmasking with distinct data and mask states, where masking decisions are independent across tokens and independent of their values. We express the smallest conditional approximation error exactly through the dependence among tokens revealed together, after conditioning on previously revealed tokens. For sequences built from independent pairs of correlated tokens, we analyze optimal schedules starting from the true fully masked distribution. We prove that the path error in relative entropy decreases inversely with the number of steps. For the sampler using exact conditional marginals, the output error in the same metric decreases with the inverse square of that number. For this family, exact representation by a finite mixture of product distributions requires exponentially many components as sequence length grows, while a suitable ordering permits exact sampling in two conditional rounds. Our results connect conditional dependence to sampling error and representation requirements.

1 INTRODUCTION

Masked diffusion models generate discrete sequences by progressively replacing mask states with tokens predicted from the available context (Austin et al., 2021; Sahoo et al., 2024). A single model evaluation can predict many positions, making parallel token generation a useful approach to reducing the number of sequential sampling steps. However, the central challenge is to preserve dependencies among tokens as more positions are generated together. In language, choices made at different positions can remain related after conditioning on the text already revealed. An independent update samples each token from its predicted conditional distribution, so the new tokens are generated without observing one another. Such an update can miss dependencies among the new tokens even when each conditional distribution is predicted accurately (Figure 1). As a result, the number and timing of updates affect both the dependence captured during generation and the accuracy of the final sequence distribution.

Existing analyses of factorized sampling separate learning error from the error due to conditional dependence, giving exact decompositions of accumulated error and upper bounds on generation error. They also characterize this dependence cost and optimize unmasking schedules using information about the data distribution (Lavenant & Zanella, 2025; Chen et al., 2026). These analyses already distinguish the accumulated conditional error from the output divergence after averaging over random reveal patterns. Mixture models represent dependence through a shared latent component (Hayakawa et al., 2024), and stochastic mask states provide shared random inputs to the token predictors (Yoo et al., 2026). However, the accumulated conditional error does not fully describe the accuracy of the final output. Averaging over the sampling process can reduce the error in the output distribution. Therefore, a precise analysis is needed to clarify the relation between conditional approximation and final generation error.

In this paper, we characterize the cost of independent token updates through the conditional dependence among tokens revealed together. We study progressive unmasking with distinct data and mask states and masking randomness that is independent across positions and independent of token values.

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.

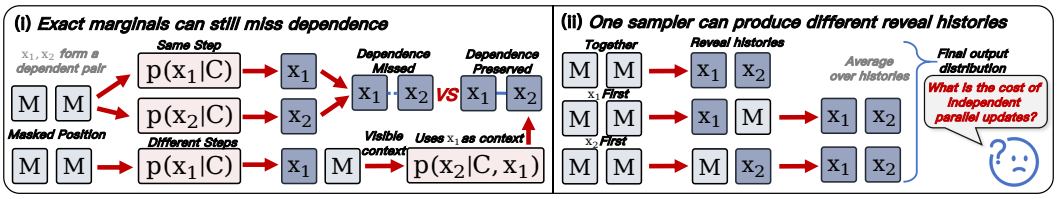

Figure 1: **Exact marginals and reveal histories.** For a dependent pair x_1, x_2 with visible context C, M denotes a masked position. (i) A simultaneous update factorizes the prediction as $p(x_1 \mid C)p(x_2 \mid C)$, but a sequential update can use $p(x_1 \mid C)$ followed by $p(x_2 \mid C, x_1)$. (ii) The same sampler can reveal the pair together, reveal x_1 first, or reveal x_2 first.

Our analysis groups each transition by the tokens already known and the tokens newly revealed. For each group, we measure the dependence remaining among the new tokens after conditioning on the known ones. Averaging this dependence over reveal patterns gives the exact minimum conditional approximation error for any finite data distribution. This expression separates the contribution of the data from the weights assigned by the schedule. For correlated token pairs, a cost is incurred when both tokens are revealed in the same step. Revealing one earlier provides context for generating the other. We compute the final output distribution by averaging these two cases. The calculation shows how averaging reveal histories changes the error rate and the sampling budget for the same sampler. We also derive optimal schedules for noisy parity blocks and compare the mixture components and conditional rounds needed to represent pair dependence. Our contributions are as follows.

(i) We quantify the gap between path error and output error for independent correlated pairs. For the sampler initialized from the true fully masked distribution and using exact conditional marginals, we compute the exact output distribution under any fixed schedule. With fixed positive noise and dependence, the optimal schedule yields path error decreasing inversely with the number of steps and output error decreasing with its inverse square.

(ii) We derive the unique schedules minimizing cumulative conditional approximation error for independent correlated pairs and noisy parity blocks. Equal increments in reveal probability are optimal for pairs. For parity blocks, an explicit recurrence determines the optimal schedule and minimum error for any number of steps. The optimum is nonuniform when each block contains more than two tokens and at least two steps are used, showing how dependence structure changes optimal scheduling.

(iii) For the pair family, we prove that exact representation by a finite mixture of product distributions requires exponentially many components as sequence length grows, and give a construction attaining the minimum. Revealing one token from each pair before its partner allows exact generation in two conditional rounds.

2 Preliminaries

Progressive Unmasking and Product Transitions. Let $X = (X_1, \ldots, X_L) \sim p$ be a clean sequence in the finite product alphabet $\mathcal{X} = \prod_{i=1}^{L} \mathcal{X}_i$, with $L \geq 1$. Let $\mathcal{M}$ be a standard Borel mask space disjoint from every $\mathcal{X}_i$. At retention level α_t, each coordinate of Z_t equals X_i with probability α_t and otherwise takes a value in $\mathcal{M}$. The retention indicators and mask values are independent across coordinates and independent of X. Observing Z_t therefore identifies which tokens have been revealed and their values.

A reverse step from t to s increases retention from α_t to α_s, where $0 \leq \alpha_t < \alpha_s \leq 1$. Set $\Delta = \alpha_s - \alpha_t$, and copy every revealed token. Each masked coordinate is independently revealed with probability $\beta_{s,t} = \Delta/(1-\alpha_t)$, using decisions independent of X and all mask values. A coordinate that remains masked is drawn from $\kappa^i_{s,t}(\cdot \mid Z^i_t)$, supported on $\mathcal{M}$. These draws are conditionally independent across coordinates and independent of X given the current state and reveal decisions. Each kernel depends only on its own current mask value. Write S for the known positions and D for the newly revealed positions. The disjoint data and mask alphabets make S observable from Z_t and D observable from (Z_t, Z_s).

Let P_t be the true law of Z_t and $P_{s|t}(\cdot \mid Z_t)$ the corresponding reverse conditional. We approximate this conditional with a measurable product kernel $R(dz_s \mid z_t) = \bigotimes_{i=1}^{L} r_i(dz_s^i \mid z_t)$. Each factor r_i may depend on the entire observed state z_t, including all revealed tokens and mask values. Independence is imposed on the outputs conditional on that shared state.

Conditional Dependence and Error Criteria. We use natural logarithms, write H for Shannon entropy, and allow relative entropy D_{KL} to take the value $+\infty$. Our error criteria place the true law in the first argument. For $A \subseteq [L] = \{1, \dots, L\}$, write $X_A = (X_i)_{i \in A}$. For fixed disjoint subsets $S, D \subseteq [L]$, the dependence among X_D given X_S is measured by conditional total correlation (Watanabe, 1960),

$$\mathrm{TC}(X_D \mid X_S) = \sum_{i \in D} H(X_i \mid X_S) - H(X_D \mid X_S). \tag{1}$$

The conditional entropies already average over X_S. This quantity is nonnegative and vanishes exactly when the coordinates in X_D are conditionally independent given X_S almost surely and it is zero for $|D| \leq 1$.

The minimum conditional approximation error at a step is

$$\Phi_{s,t}^{p} := \inf_{R} \mathbb{E}_{Z_t \sim P_t} D_{\mathrm{KL}}\big(P_{s|t}(\cdot \mid Z_t) \,\big\|\, R(\cdot \mid Z_t)\big)\,. \tag{2}$$

The infimum ranges over all measurable product kernels defined above. The expectation uses the true current state distribution, and both conditional laws are evaluated at the same observed state.

For a full schedule $1 = \alpha_0 > \cdots > \alpha_T = 0$, generation starts from the fully masked state at $t = T$ and ends at $t = 0$. Let P and Q be the true reverse Markov path law and a sampler's path law, with time t marginals P_t and Q_t. The sampler uses product transitions and an initial distribution Q_T, which may differ from P_T. We distinguish the path error $D_{\mathrm{KL}}(P\|Q)$ from the output error $D_{\mathrm{KL}}(p\|Q_0)$, where $P_0 = p$. The former compares complete trajectories and the latter compares final outputs after intermediate states have been marginalized.

3 Quantifying the Output Error

In this section, we first characterize the minimum conditional approximation error and decompose path KL (Section 3.1). We then compute the exact output distribution of the canonical sampler for correlated pairs (Section 3.2). Finally, we optimize the schedule and establish the resulting path and output error rates (Section 3.3).

3.1 Evaluating Conditional Approximation

We first evaluate the minimum conditional error $\Phi_{s,t}^{p}$ in equation 2 to quantify the cost of a single product update. We group the possible updates by the coordinates already visible at time t and those revealed during the transition. These random sets are $S = \{i \in [L] : Z_t^i \in \mathcal{X}_i\}$, $D = \{i \in [L] \setminus S : Z_s^i \in \mathcal{X}_i\}$. For a fixed reveal pattern, the observed values X_S provide the conditioning information for generating X_D. Independent retention and reveal decisions give $\Pr(i \in S) = \alpha_t$, $\Pr(i \in D) = \Delta$, and $\Pr(i \notin S \cup D) = 1 - \alpha_s$. Conditional on Z_t and D, the joint law of the new tokens is $p(X_D \in \cdot \mid X_S)$. The error of its best product approximation determines the contribution of this reveal pattern.

Theorem 1 (Exact conditional error)**.** *For every finite target distribution p, the minimum conditional approximation error satisfies*

$$\Phi_{s,t}^{p} = \mathbb{E}_{S,D}\big[\mathrm{TC}(X_D \mid X_S)\big] \quad = \sum_{\substack{S,D \subseteq [L] \\ S \cap D = \varnothing}} \alpha_t^{|S|} \Delta^{|D|} (1 - \alpha_s)^{L - |S| - |D|} \mathrm{TC}(X_D \mid X_S). \tag{3}$$

Proof sketch. The product projection identity gives the minimizing kernel $\bigotimes_i P(Z_s^i \in \cdot \mid Z_t)$ (Cover et al., 1991). Fix the sets S, D and the observed values X_S. The true conditional and its product projection share the reveal probabilities, copied coordinates, and local mask laws. Their

divergence within this branch reduces to that between $p(X_D \in \cdot \mid X_S)$ and the product of its coordinate marginals. Averaging over X_S gives $\mathrm{TC}(X_D \mid X_S)$ (Lavenant & Zanella, 2025; Chen et al., 2026). Averaging over the independent coordinate assignments gives the weights in equation 3. Appendix B.1 gives the complete proof. □

A reveal pattern has zero cost when the newly revealed coordinates are conditionally independent given the observed ones. In particular, a single newly revealed coordinate contributes no dependence cost. For larger reveal sets, the formula weights the remaining dependence by the probability of generating those coordinates together. The schedule controls these probabilities.

Independent blocks allow the transition cost to be evaluated on smaller subsets of coordinates. For a partition $\mathcal{B}$ of $[L]$, suppose $p = \bigotimes_{B\in\mathcal{B}} p_B$, where p_B is the law of X_B. Conditioning on X_S preserves independence across blocks, so conditional entropy additivity gives $\Phi^p_{s,t} = \sum_{B\in\mathcal{B}} \Phi^{p_B}_{s,t}$. Appendix B.2 gives the full calculation.

We now pass from the cost of one transition to the error of a complete sampling trajectory. The sampler's coordinate kernels may also differ from the true conditional marginals, introducing an additional error at each step. For $R_t = \bigotimes_i r_{t,i}$ in the transition from t to $t-1$, define the true coordinate conditional and its estimation error by

$$\mu_{t,i}(\cdot \mid z_t) = P(Z^i_{t-1} \in \cdot \mid Z_t = z_t), e_{t,i} \quad = \mathbb{E}_{Z_t\sim P_t} D_{\mathrm{KL}}(\mu_{t,i}(\cdot \mid Z_t) \,\|\, r_{t,i}(\cdot \mid Z_t))\,.$$

Applying the KL chain rule (Cover et al., 1991) and the decomposition into dependence and marginal estimation errors (Lavenant & Zanella, 2025) gives

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(P_T\|Q_T) + \sum_{t=1}^{T} \Phi^p_{t-1,t} + \sum_{t=1}^{T}\sum_{i=1}^{L} e_{t,i}. \tag{4}$$

The three terms measure initialization error, accumulated conditional approximation error, and marginal estimation error. Choosing $Q_T = P_T$ and $r_{t,i} = \mu_{t,i}$ makes the first and third terms zero. Since both are nonnegative, this choice minimizes path KL at a fixed schedule and leaves exactly the cumulative dependence cost $\sum_{t=1}^{T} \Phi^p_{t-1,t}$. Marginalizing the trajectory gives $D_{\mathrm{KL}}(p\|Q_0) \le D_{\mathrm{KL}}(P\|Q)$. The output law averages over the intermediate states, so its divergence can be smaller than the accumulated path error.

3.2 Computing the Output Distribution

Pairs of correlated bits allow us to track how a simultaneous reveal changes the generated distribution. Let $L \ge 2$ and $m = \lfloor L/2 \rfloor$. Group coordinates $2j-1$ and $2j$ into a pair for $j = 1, \dots, m$. For $\eta \in (0,1)$, define the pair law and the target distribution by

$$g_\eta(x,y) = \frac{1-\eta}{2}\mathbf{1}\{x=y\} + \frac{\eta}{4}, \qquad (x,y)\in\{0,1\}^2, \qquad p^{\mathrm{pair}}_{\eta,L} \quad = g_\eta^{\otimes m} \otimes \mathrm{Ber}(1/2)^{\otimes(L-2m)}. \tag{5}$$

The pairs are independent, and the final coordinate is an independent fair bit when L is odd. Each pair has fair marginals and mismatch probability $\eta/2$. Writing $h_2(u) = -u\log u - (1-u)\log(1-u)$ for binary entropy, its mutual information is $c_\eta = \log 2 - h_2(\eta/2) > 0$. For $0 < \eta < 1$, every binary configuration has positive probability and each pair is dependent. We write $p = p^{\mathrm{pair}}_{\eta,L}$ for this target.

Fix an integer $T \ge 1$ and a strict schedule $1 = \alpha_0 > \cdots > \alpha_T = 0$. Its increments are $\Delta_t = \alpha_{t-1} - \alpha_t > 0$, with $\sum_{t=1}^{T} \Delta_t = 1$. We call the sampler initialized from P_T with transitions

$$Q^*_T = P_T, \qquad R^*_t(\cdot \mid z_t) = \bigotimes_{i=1}^{L} \mu_{t,i}(\cdot \mid z_t)$$

the canonical sampler. Its path law is Q^* and its final marginal is Q^*_0. It samples coordinates independently using their exact conditional distributions, so its path error contains only the conditional approximation costs in equation 4.

The schedule determines the probability of a collision, in which both members of a pair are revealed in the same step. Write $\rho = \sum_{t=1}^{T} \Delta_t^2$ for this probability. Collisions determine the departure of the output from the target.

Theorem 2 (Exact output of the canonical sampler)**.** *For the pair target and canonical sampler defined above, the output distribution is*

$$\begin{aligned} g_{\eta,\rho} &= (1-\rho)g_\eta + \rho\ \mathrm{Unif}(\{0,1\}^2), \qquad \eta_\rho = \eta + (1-\eta)\rho, \\ Q_0^* &= g_{\eta,\rho}^{\otimes m} \otimes \mathrm{Ber}(1/2)^{\otimes(L-2m)}, \qquad g_{\eta,\rho} = g_{\eta_\rho}. \end{aligned} \tag{6}$$

Here $\mathrm{Unif}(\{0,1\}^2)$ *is the uniform law on the four possible pairs.*

Proof sketch. Let $J_i \in \{1,\dots,T\}$ denote the step in which coordinate i is first revealed, so $J_i = t$ refers to the transition from t to $t-1$. Under the canonical sampler, a coordinate remains masked at time t with probability $1-\alpha_t$. Conditional on remaining masked, it is revealed with probability $\Delta_t/(1-\alpha_t)$. Consequently $\Pr(J_i = t) = \Delta_t$. The reveal decisions are independent across coordinates, giving $\Pr(J_{2j-1} = J_{2j}) = \rho$.

If the two reveal times coincide, neither bit is available as context when the pair is generated. Both conditional marginals are fair, and the product update produces a uniform pair. If the reveal times differ, the first bit is fair and the second is sampled from its true conditional given the first. This produces g_η. Mixing the two cases gives $g_{\eta,\rho}$.

The data conditional for a bit depends only on its revealed partner. Independent initialization, local mask refreshes, and independent transition draws therefore preserve independence across pairs. An unpaired bit remains fair. Appendix B.4 gives the complete proof. □

The output remains in the same pair family, with mismatch probability increased from $\eta/2$ to $\eta_\rho/2$. The entire schedule affects this law through the single collision probability ρ. We can therefore optimize its output error by determining how that error depends on ρ.

3.3 Determining the Optimal Error Rates

For one pair, the conditional dependence cost is c_η when both members are newly revealed together. Every other reveal pattern has zero cost within that pair. Their joint reveal probability in a transition from t to s is $(\alpha_s - \alpha_t)^2$. Block additivity gives $\Phi^p_{s,t} = mc_\eta(\alpha_s - \alpha_t)^2$, and hence $\Phi^p_{t-1,t} = mc_\eta\Delta_t^2$ along the schedule. This evaluates the path error. To express the output error, define the relative entropy between Bernoulli laws by $d_2(a\|b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$, $a, b \in (0,1)$. The target and generated mismatch probabilities are $\eta/2$ and $\eta_\rho/2$, respectively.

Corollary 3 (Optimal errors for correlated pairs)**.** *Under the assumptions of Theorem 2, the canonical sampler satisfies*

$$D_{\mathrm{KL}}(P\|Q^*) = mc_\eta\rho, \qquad D_{\mathrm{KL}}(p\|Q_0^*) = m\, d_2\Big(\frac{\eta}{2}\,\Big\|\,\frac{\eta_\rho}{2}\Big). \tag{7}$$

Its cumulative conditional cost obeys

$$\sum_{t=1}^{T} \Phi^p_{t-1,t} = mc_\eta \sum_{t=1}^{T} \Delta_t^2 \geq \frac{mc_\eta}{T}. \tag{8}$$

Uniform increments $\Delta_t = 1/T$*, equivalently* $\alpha_t = 1 - t/T$*, uniquely minimize both errors over strict schedules. The optimal output error satisfies*

$$\min_{\{\alpha_t\}} D_{\mathrm{KL}}(p\|Q_0^*) = m\, d_2\left(\frac{\eta}{2}\,\middle\|\,\frac{\eta}{2} + \frac{1-\eta}{2T}\right), \qquad \frac{m(1-\eta)^2}{2T^2} \leq \min_{\{\alpha_t\}} D_{\mathrm{KL}}(p\|Q_0^*) \leq \frac{m(1-\eta)^2}{\eta T^2}. \tag{9}$$

With initialization $Q_T = P_T$*, the minimum path error over all product heads and strict schedules is* mc_η/T*.*

Proof sketch. The canonical sampler has no initialization or marginal estimation error, so equation 4 and the pair cost give $D_{\mathrm{KL}}(P\|Q^*) = mc_\eta\rho$. For output KL, group the four pair configurations according to whether their bits agree. Within either group, the two conditional laws are uniform. Their divergence thus equals the binary KL between the two mismatch probabilities. Independence across pairs gives the second identity in equation 7.

Since the increments sum to one, Cauchy's inequality gives $\rho \geq 1/T$, with equality exactly at uniform increments. The output error increases strictly with ρ, because $\partial_b d_2(a\|b) = (b-a)/[b(1-b)] > 0$ for $b > a$. Both objectives therefore have the same unique optimal schedule. Arbitrary product heads add nonnegative estimation errors, so they cannot lower the minimum path KL. Appendix B.5 supplies the full proof. □

For fixed $\eta \in (0,1)$, the optimal path error is $\Theta(L/T)$ and the optimal canonical output error is $\Theta(L/T^2)$. The path law retains the reveal history, so each collision contributes the fixed dependence cost c_η and the total cost is linear in ρ. The output law averages over that history. Its mismatch probability changes by $(1-\eta)\rho/2$, and relative entropy is quadratic in this change near the target. At the optimal value $\rho = 1/T$, this difference gives distinct step requirements for a prescribed error tolerance. As $\varepsilon/L \to 0$, path KL at most ε requires $\Theta(L/\varepsilon)$ steps, while canonical output KL at most ε requires $\Theta(\sqrt{L/\varepsilon})$ steps. The path bound can therefore give a conservative budget for output accuracy even though both errors have the same optimal schedule. The path optimum ranges over all product heads with true initialization. The output rate fixes the heads to the true conditional marginals and optimizes the schedule. Allowing arbitrary measurable heads changes the output optimization problem because every head can use the same initial mask state as a source of shared randomness.

$$\mathcal{E}^{\mathrm{KL}}_{\mathrm{opt}} = \begin{cases} \dfrac{mc_\eta}{T}, & (i)\ \text{Path, all product heads,} \\ \Theta(L/T^2), & (ii)\ \text{Canonical output,} \\ 0, & (iii)\ \text{Unrestricted output, } T = 1,\ P_T \text{ atomless.} \end{cases} \tag{10}$$

For the last equality, let $U \sim P_T$ be the fully masked state. Given any finite target p, choose a measurable partition $\{A_x : x \in \mathcal{X}\}$ with $P_T(A_x) = p(x)$ (Bogachev & Ruas, 2007). When $U \in A_x$, let coordinate head i return x_i. The update is a product of point masses conditional on U, and averaging over U produces p exactly in one step. For a target with at least two support points, this construction has infinite path KL under the original coupling. Appendix B.7 proves these.

4 Preserving Dependence during Generation

4.1 Optimizing Schedules for Parity Blocks

For pairs, the transition cost depends only on the reveal increment. We now consider blocks whose dependence involves all their coordinates. For an integer $k \geq 2$ and $\lambda \in (0,1)$, define the noisy parity law

$$\gamma_{\lambda,k}(x) = 2^{-k}\left[1 + \lambda(-1)^{x_1+\cdots+x_k}\right], \qquad x \in \{0,1\}^k. \tag{11}$$

For an integer $n \geq 1$, let $p = \gamma_{\lambda,k}^{\otimes n}$ consist of n independent blocks, with total length $L = nk$. Every proper marginal of a block is uniform product, and the total correlation of the full block is $c_\lambda = \log 2 - h_2((1-\lambda)/2) > 0$, where h_2 is binary entropy.

Consider a transition from t to s, with $0 \leq \alpha_t < \alpha_s \leq 1$. A block contributes conditional total correlation c_λ exactly when at least two coordinates are newly revealed and complete the block. Every other reveal pattern has zero cost within that block. Theorem 1 and block additivity give

$$\Phi^p_{s,t} = nc_\lambda\left[\alpha_s^k - \alpha_t^k - k\alpha_t^{k-1}(\alpha_s - \alpha_t)\right]. \tag{12}$$

Appendix C.1 gives the full calculation. When $k = 2$, this reduces to $nc_\lambda(\alpha_s - \alpha_t)^2$. For $k > 2$, the cost also depends on the current retention level.

Fix an integer $T \geq 1$ and a strict schedule $1 = \alpha_0 > \cdots > \alpha_T = 0$. Write its nodes in increasing order as $x_j = \alpha_{T-j}$, so $0 = x_0 < \cdots < x_T = 1$, and define $B_k(b,a) = b^k - a^k - ka^{k-1}(b-a)$ for $0 \leq a \leq b \leq 1$. The normalized cumulative conditional error is $\frac{1}{nc_\lambda}\sum_{t=1}^T \Phi^p_{t-1,t} = \sum_{j=1}^T B_k(x_j, x_{j-1})$. For each integer $h \geq 1$, let $C_{k,h}$ denote the minimum normalized cost over strict schedules with h steps.

Corollary 4 (Optimal schedule for parity blocks). *The minimum normalized cost satisfies $C_{k,1} = 1$ and the following recurrence for $h \geq 2$.*

$$C_{k,h} = 1 - \left(\frac{k-1}{k-1+C_{k,h-1}}\right)^{k-1}, \qquad \min_{1=\alpha_0>\cdots>\alpha_T=0} \sum_{t=1}^{T} \Phi_{t-1,t}^{p} = nc_\lambda C_{k,T}. \tag{13}$$

The minimizing schedule is unique. Define $r_h = (k-1)/(k-1+C_{k,h-1})$ for $h \geq 2$. The optimal nodes are

$$x_0 = 0, \qquad x_j = \prod_{h=j+1}^{T} r_h \quad (1 \leq j \leq T), \qquad \alpha_t = x_{T-t}, \tag{14}$$

where the empty product is one. For $k > 2$ and $T \geq 2$, the schedule is nonuniform.

Proof sketch. Let $V_h(z)$ be the minimum normalized cost of an increasing partition from zero to $z \in (0,1]$ in h steps. Homogeneity of B_k gives $V_h(z) = C_{k,h} z^k$. Writing the penultimate node as yz gives the Bellman recursion (Bellman, 1954) $V_h(z) = z^k \min_{0\leq y\leq 1} \left\{1 + (C_{k,h-1} + k - 1)y^k - ky^{k-1}\right\}$. The scalar objective decreases up to $r_h \in (0,1)$ and increases afterwards, giving a unique minimum. Substitution yields the recurrence. Recursively recovering each preceding optimal partition gives equation 14 and uniqueness of the strict schedule. For $k = 2$, the recurrence gives $x_j = j/T$. For $k > 2$ and $T \geq 2$, the first two positive nodes satisfy $x_1/x_2 = (k-1)/k > 1/2$, which excludes uniform increments. Appendix C.2 gives the full proof. □

The optimal schedule depends on k and T. The number of blocks n and the parity strength λ only scale its cost. For $k > 2$, all unconditional pairwise marginals are independent, yet the optimal schedule is nonuniform for $T \geq 2$. This shows that pairwise independence alone does not determine the optimal allocation of reveals. The recurrence gives an exact optimizer for every finite step count within the specified retention channel. The objective is the cumulative conditional approximation error in equation 4. Appendix C.3 gives bounds and asymptotics for its optimal value.

4.2 Quantifying the Resources for Exact Generation

A finite mixture generates dependence through a shared component index. Conditional on that index, it draws all coordinates independently. For integers $L \geq 2$ and $K \geq 1$, define the class of distributions on $\{0,1\}^L$

$$\mathcal{Q}_K = \left\{ \sum_{\ell=1}^{K} w_\ell \prod_{i=1}^{L} q_{i,\ell}(x_i) : w_\ell \geq 0,\ \sum_{\ell=1}^{K} w_\ell = 1 \right\},$$

where w_ℓ are mixing probabilities and each $q_{i,\ell}$ is a distribution on $\{0,1\}$. We determine when this class contains the pair target $p_{\eta,L}^{\text{pair}}$ from Section 3.2.

Theorem 5 (Exact mixture component count). *For $L \geq 2$, $m = \lfloor L/2 \rfloor$, and $\eta \in (0,1)$,*

$$p_{\eta,L}^{\text{pair}} \in \mathcal{Q}_K \quad \Longleftrightarrow \quad K \geq 2^m. \tag{15}$$

An explicit mixture indexed by the first bit of each pair attains the minimum.

Proof sketch. Let $A = (X_1, X_3, \ldots, X_{2m-1})$ and $B = (X_2, X_4, \ldots, X_{2m})$, omitting any unpaired bit. Their joint probability matrix is $G_\eta^{\otimes m}$, where $G_\eta = (g_\eta(a,b))_{a,b\in\{0,1\}}$. Since $\det G_\eta = (1-\eta)/4 > 0$, the pair matrix has rank two and $G_\eta^{\otimes m}$ has rank 2^m. Each coordinate product induces a matrix of rank at most one across this split. A mixture of K such products has rank at most K by the standard rank bound (Cohen & Rothblum, 1993), proving $K \geq 2^m$.

To attain the bound, draw A uniformly from $\{0,1\}^m$, together with an independent fair bit if L is odd. Conditional on A, generate the even coordinates by

$$X_{2j} = X_{2j-1} \oplus N_j, \qquad N_j \overset{\text{iid}}{\sim} \text{Ber}(\eta/2), \qquad j = 1, \ldots, m, \tag{16}$$

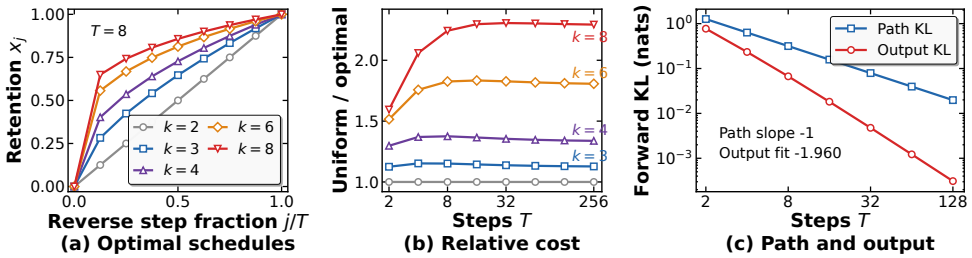

Figure 2: Exact schedule and error calculations. (a) Optimal retention nodes at $T = 8$. (b) Uniform cost divided by optimal cost for parity blocks. (c) Path and output KL for the canonical pair sampler with $\eta = 0.25$, $L = 16$, and uniform increments and the output fit uses $T \geq 16$.

where $\oplus$ denotes addition modulo two and the noise bits are independent of all draws in the first stage. For each value $A = a$, this construction is a coordinate product with the odd bits fixed. Averaging over the 2^m equally likely values of a recovers the independent pair laws. Zero weights permit every larger K. Appendix C.4 gives the full proof. □

The two stages also give an exact sampler using two conditional product rounds. The explicit mixture has 2^m components, while the conditional sampler draws one context and uses it in the next update. This sampler uses known pair structure and exact target conditionals. Its prescribed reveal order changes the independent random reveal schedule of Section 2. Appendix C.6 verifies the output law.

5 Experiments

In this section, we conduct experiments to address the following research questions: **RQ1:** How does dependence shape the cost and design of parallel generation? **RQ2:** How do path and output errors scale with the number of sampling steps? **RQ3:** How do reveal policies affect prediction in pretrained denoisers?

5.1 Experimental Setup

Evaluation metrics. We measure conditional TC costs, path KL, and output KL (Watanabe, 1960; Lavenant & Zanella, 2025; Chen et al., 2026). Total variation (TV) measures the approximation error of product mixtures. For pretrained denoisers, we report token NLL and the rate at which linked spans are revealed in the same round. Token NLL evaluates the probability assigned to reference tokens in the context available when they are revealed.

Datasets and models. Synthetic evaluations use correlated pairs, noisy parity blocks, and a coupling construction with shared mask randomness. Controlled span infilling uses constructed texts for copy, entity agreement, and repeated article tags. We evaluate MDLM (Sahoo et al., 2024) and the Qwen3 diffusion checkpoint from dLLM (Zhou et al., 2026) on these tasks.

Baseline methods. For the synthetic schedule experiments, we compare analytic optimal schedules with uniform increments. For pretrained denoisers, dependency selection is compared with confidence and random balanced selection at matched model evaluation counts. For more detailed descriptions of the experimental settings, please refer to Appendix F.

5.2 Experimental Results

Schedule Design under Dependence (RQ1). We evaluate how dependence changes the cost of parallel updates and the schedule that minimizes it. For the pair and noisy parity families, we evaluate the exact conditional cost under the analytic optimal schedule and uniform increments. Figure 2(a,b) shows that uniform increments attain the pair optimum. Parity blocks favor a larger initial reveal followed by smaller increments. At $T = 8$, adapting the schedule reduces the uniform cost by 27.3% for $k = 4$ and 55.4% for $k = 8$. As the block size grows, the benefit of adapting the schedule becomes larger.

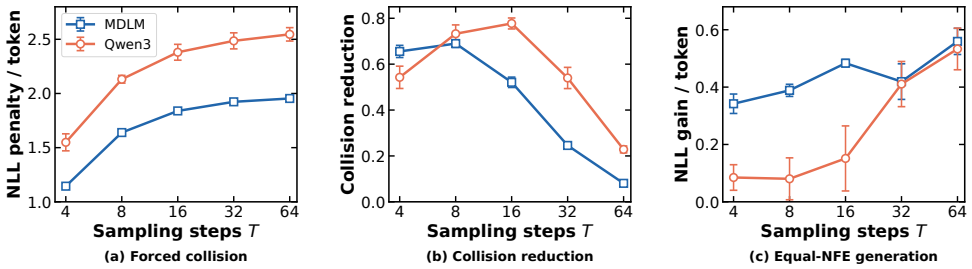

Figure 3: Pretrained-denoiser diagnostics across sampling steps. (a) Token NLL differences between forced simultaneous revelation and dependency-aware separation across sampling steps. (b) Same-round collision-rate differences between random balanced selection and dependency-aware separation. (c) Token NLL differences between confidence selection and dependency-aware selection under matched model evaluation counts.

As shown in Figure 2(b), the schedule difference remains visible as the number of steps grows. At $T = 256$, the products $TC_{k,T}$ differ from their limiting constants by at most 0.014 across the tested block sizes. As more coordinates from a parity block are already visible, later transitions retain more conditional dependence, and the optimal schedule assigns them smaller increments.

The same dependence structure also appears in the representation diagnostics. For the paired family, regularized EM fits with fewer than 2^m components leave total variation at least 0.222578 and the analytic construction with 2^m components has maximum numerical total variation 2.01×10^{-12}. The shared-randomness diagnostic gives marginal KL 0.00033 ± 0.00006, compared with conditional KL 8.587 for the corresponding smoothed construction and 0.9491 for the product projection (Appendix F.4).

Error Rates across Sampling Steps (RQ2). We evaluate path KL and output KL for the same canonical pair sampler as the number of sampling steps increases. We fix $\eta = 0.25$ and $L = 16$, initialize from the true fully masked distribution, use exact conditional marginals, and apply uniform retention increments. Figure 2(c) shows that the two errors decrease at different rates. Increasing T from 16 to 128 reduces path KL by a factor of 8 and output KL by a factor of 58.8.

At $T = 128$, the scaled output divergence $T^2 D_{\mathrm{KL}}(p \| Q_0^*)/m$ reaches 0.6344, close to the limiting value $9/14 \approx 0.6429$ (Appendix F.3). A reveal collision contributes directly to path KL because the two dependent coordinates are generated independently in that transition. The output distribution averages over reveal histories, so the resulting change in pair mismatch probability is of order $1/T$, and output KL is quadratic in this change.

Reveal Policies in Pretrained Denoisers (RQ3). We examine how reveal policies affect token prediction in two pretrained denoisers. Under teacher-forced contexts, forced simultaneous revelation increases token NLL by 1.899 ± 0.049 nats per scored token over $25{,}459$ matched pairs. Dependency-aware separation reduces the rate at which linked spans are revealed together by 0.4879 ± 0.0060 against random balanced selection over $27{,}102$ matched pairs.

We then compare dependency-aware selection with confidence-native selection. Dependency-aware selection reduces token NLL by 0.3831 ± 0.0196 nats per scored token over $22{,}495$ matched pairs. Figure 3 shows the stepwise comparisons for both checkpoints, together with the forced-revelation and collision diagnostics.

6 Conclusion

In masked diffusion, independent parallel updates can miss dependencies among tokens generated together, introducing approximation error even when their conditional marginals are exact. In this paper, we expressed the minimum conditional error under the studied masking channel as an average of conditional total correlation and derived optimal schedules for correlated pairs and noisy parity blocks. For the pair family with fixed positive noise and dependence, optimizing the canonical sampler's schedule gives path KL of inverse order and output KL of inverse square order in the number of steps.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing the writing. We 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*, volume 2025, pp. 50726–50753, 2025.

Koenraad MR Audenaert. A sharp continuity estimate for the von neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Richard Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas S Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. *Advances in Neural Information Processing Systems*, 38:55981–56007, 2026.

Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure theory*, volume 1. Springer, 2007.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11305–11315. IEEE, 2022.

Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *Journal of Machine Learning*, 4(2):108–127, 2025. doi: 10.4208/jml.240812.

Sitan Chen, Kevin Cong, and Jerry Li. Optimal inference schedules for masked diffusion models. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336, pp. 1279–1311. PMLR, 29 Jun–03 Jul 2026.

Joel E Cohen and Uriel G Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168, 1993.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast llms via self-distillation through time. In *International Conference on Learning Representations*, volume 2025, pp. 5007–5045, 2025.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 6112–6121, 2019.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*, 2018.

Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. Distillation of discrete diffusion through dimensional correlations. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

Michael Hersche, Nicolas Menet, Ronan Tanios, and Abbas Rahimi. Locally coherent parallel decoding in diffusion language models. In *Forty-third International Conference on Machine Learning*, 2026.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

Fei Huang, Tianhua Tao, Hao Zhou, Lei Li, and Minlie Huang. On the learning of non-autoregressive transformers. In *International conference on machine learning*, pp. 9356–9376. PMLR, 2022.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. Accelerating diffusion llms via adaptive parallel decoding. *Advances in neural information processing systems*, 38:52870–52888, 2026.

Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

Gowtham Ramani Kumar, Cheuk Ting Li, and Abbas El Gamal. Exact common information. In *2014 IEEE International Symposium on Information Theory*, pp. 161–165, 2014. doi: 10.1109/ISIT.2014.6874815.

Hugo Lavenant and Giacomo Zanella. Error bounds and optimal schedules for masked diffusions with factorized approximations. *arXiv preprint arXiv:2510.25544*, 2025.

Yuchen Liang, Renxiang Huang, Lifeng Lai, Ness Shroff, and Yingbin Liang. Absorb and converge: Provable convergence guarantee for absorbing discrete diffusion models. *Advances in Neural Information Processing Systems*, 38:20283–20318, 2026.

Anji Liu, Oliver Broadrick, Mathias Niepert, and Guy Van den Broeck. Discrete copula diffusion. In *International Conference on Learning Representations*, volume 2025, pp. 88953–88979, 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32819–32848, 2024.

Guido Montúfar. Mixture decompositions of exponential families using a decomposition of their sample spaces. *arXiv preprint arXiv:1008.0204*, 2010.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations*, volume 2025, pp. 64972–65009, 2025.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *Proceedings of machine learning research*, 267: 52584, 2025.

Igal Sason and Sergio Verdú. f -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Guanghan Wang, Yair Schiff, Subham Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. *Advances in Neural Information Processing Systems*, 38:147282–147339, 2026.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

Aaron Wyner. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, 21(2):163–179, 1975.

Chenxiao Yang, Cai Zhou, David Wipf, and Zhiyuan Li. On powerful ways to generate: Autoregression, diffusion, and beyond. In *International Conference on Learning Representations*, volume 2026, pp. 117754–117810, 2026.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *URL https://arxiv. org/abs/2508.15487*, 2025.

Jaehoon Yoo, Wonjung Kim, Chanhyuk Lee, and Seunghoon Hong. Infinite mask diffusion for few-step distillation. In *Forty-third International Conference on Machine Learning*, 2026.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *International Conference on Learning Representations*, volume 2025, pp. 63186–63227, 2025.

Yangyang Zhong, Yanmei Gu, Zhengqing Zang, Xiaomeng Li, Yuqi Ding, Xibei Jia, Yuting Shen, Zhenzhong Lan, Liwang Zhu, Weiping Liu, et al. Parallelism and generation order in masked diffusion language models: Limits today, potential tomorrow. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7178–7218, 2026.

Zhanhui Zhou, Lingjie Chen, Hanghang Tong, and Dawn Song. dllm: Simple diffusion language modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 78–88, 2026.

A Extended Related Work

Masked diffusion models. Parallel token generation includes translation with latent fertilities and iterative masked prediction (Gu et al., 2018; Ghazvininejad et al., 2019). Discrete diffusion extends denoising to categorical data through multinomial corruption, structured transitions, and continuous time Markov chains (Hoogeboom et al., 2021; Austin et al., 2021; Campbell et al., 2022). MDLM, LLaDA, and Dream apply masked prediction to language modeling (Sahoo et al., 2024; Nie et al., 2026; Ye et al., 2025). Score entropy models probability ratios, while generalized masked diffusion uses weighted cross entropy (Lou et al., 2024; Sason & Verdú, 2016). For absorbing diffusion, Ou et al. relate the concrete score to clean conditional probabilities (Ou et al., 2025). These formulations provide coordinate predictions whose joint use during sampling determines which token dependencies are retained.

Factorized token prediction. The divergence from a joint law to the product of its marginals is total correlation (Watanabe, 1960; Cover et al., 1991). Huang et al. use its conditional form to characterize the error of independent token prediction when the coordinate marginals are exact (Huang et al., 2022). Theorem 1 applies this product projection to our reverse transition. Its contribution is the explicit average over known and newly revealed positions under the specified masking channel.

Error decomposition. Lavenant and Zanella decompose the KL between joint laws of tokens and reveal histories into conditional total correlation and marginal estimation error (Lavenant & Zanella, 2025, Proposition 2). Marginalizing the histories yields an output KL upper bound. Their framework includes selection rules depending on previously generated values and allows random batch sizes. Equation 4 uses this decomposition with an initialization term. Our pair calculation directly evaluates the output distribution of the canonical sampler. With true initialization, exact conditional marginals, and fixed $\eta \in (0, 1)$, it gives optimal path and output rates of $\Theta(L/T)$ and $\Theta(L/T^2)$, respectively.

Optimal reveal schedules. Chen, Cong, and Li express the expected KL conditional on random ordered partitions with prescribed batch sizes through an information profile (Chen et al., 2026, Theorems 4 and 16). They characterize optimal schedules and bound the KL of the output mixture. Our independent reveal times induce random batch sizes. Conditional on these sizes, the coordinate partition is uniform. This relates our objective to their formulation. Within our retention channel, the pair and parity families admit explicit optimal schedules for every finite step count. The parity recurrence shows how higher-order dependence changes the optimum even when all unconditional pairwise marginals are independent.

Sampling methods. Uniformization, τ leaping, and sampling based on first hitting times have been studied through convergence guarantees and numerical error analyses (Chen & Ying, 2025; Liang et al., 2026; Zheng et al., 2025). MaskGIT selects positions using confidence (Chang et al., 2022); entropy bounded unmasking uses predicted token entropies to choose the size of an update (Ben-Hamu et al., 2026). ReMDM allows generated tokens to be masked again (Wang et al., 2026), and Zhong et al. examine generation order and realized parallelism across language tasks (Zhong et al., 2026). Distillation through time and diffusion duality provide approaches to reducing sampling steps (Deschenaux & Gulcehre, 2025; Sahoo et al., 2025). Our transition formula assumes independent reveal decisions with probabilities determined by time. Applying it to adaptive selection or remasking requires the corresponding posterior and transition law.

Mixture distillation. Di4C models dependence through mixtures of coordinate product distributions and distills multiple denoising steps into fewer updates (Hayakawa et al., 2024). The shared mixture component couples predictions that are independent conditional on that component. Our representation theorem quantifies the number of explicit components required for exact generation of the correlated pair family. It gives the threshold 2^m and an attaining construction, where m is the number of pairs. This result concerns the stated finite mixture class; it does not impose a general lower bound on neural parameter count.

Decoding with token dependencies. Discrete Copula Diffusion supplements diffusion predictions with a copula model (Liu et al., 2025). Adaptive Parallel Decoding combines diffusion marginals with a small autoregressive model (Israel et al., 2026). Block Diffusion generates blocks autoregressively, with diffusion within each block (Arriola et al., 2025), while CoDiLA uses a local autoregressive decoder conditioned on diffusion predictions (Hersche et al., 2026). These methods provide context or joint predictions within generation. Our two conditional rounds give an exact construction for the pair family using known dependency structure and true target conditionals.

Mixture representations. Mixtures of product distributions are related to nonnegative rank and decompositions of exponential families (Cohen & Rothblum, 1993; Montúfar, 2010). Wyner's common information and exact common information study shared randomness under different approximation criteria (Wyner, 1975; Kumar et al., 2014). Our exact component count uses ordinary matrix rank. Appendix C.5 gives latent entropy and approximation bounds. Component cardinality, latent entropy, and asymptotic common randomness rates measure distinct resources. Computational expressivity results concern budgets such as context length and decoding steps (Yang et al., 2026).

Shared mask randomness. Infinite Mask Diffusion introduces continuous mask randomness for distillation with few sampling steps (Yoo et al., 2026). A shared atomless variable can support exact marginal generation through deterministic coordinate heads. The conditional approximation objective also depends on how that variable is coupled to the data. Under independent masking, the fully masked posterior is the data law and its product projection cost is total correlation. A deterministic transport induces a different joint law with the same output marginal. Appendix E compares these objects using the definitions. This separates the ability to represent an output distribution from fitting the reverse conditionals of a fixed forward process.

B Proof Details for Output Error

We give the complete proofs for Section 3, followed by weighted schedule extensions and the construction for unrestricted output generation. All conditional errors are evaluated under the true state laws specified in Section 2.

B.1 Product Projection under the Masking Channel

Lemma 6 (Product projection on standard Borel spaces)**.** *Let μ be a probability law on a finite product of standard Borel spaces, with coordinate marginals μ_i. For every product reference $r^{\otimes} = \bigotimes_i r_i$, the product projection identity (Cover et al., 1991) is*

$$D_{\mathrm{KL}}(\mu \| r^{\otimes}) = D_{\mathrm{KL}}\left(\mu \,\middle\|\, \bigotimes_i \mu_i\right) + \sum_i D_{\mathrm{KL}}(\mu_i \| r_i).$$

The identity allows infinite divergences. If its first term on the right is finite, the unique minimizing product law is $\bigotimes_i \mu_i$. The identity also holds for regular conditional laws after integration over the conditioning variable.

Proof. Write $\mu^{\otimes} = \bigotimes_i \mu_i$. If some $\mu_i \not\ll r_i$, a coordinate cylinder has positive μ mass and zero $r^{\otimes}$ mass. The joint divergence and the corresponding marginal divergence are infinite. Suppose every $\mu_i \ll r_i$. The set $G = \{x : \prod_i (d\mu_i/dr_i)(x_i) > 0\}$ has full μ measure, and $\mu^{\otimes}$ is equivalent to $r^{\otimes}$ on G. If $\mu \not\ll \mu^{\otimes}$, this equivalence implies $\mu \not\ll r^{\otimes}$, so both the joint divergence and the projection term are infinite.

In the remaining case, the density identity holds μ almost surely,

$$\log \frac{d\mu}{dr^{\otimes}}(x) = \log \frac{d\mu}{d\mu^{\otimes}}(x) + \sum_i \log \frac{d\mu_i}{dr_i}(x_i).$$

Integrating a marginal term under μ is equivalent to integrating it under μ_i. Each negative logarithmic part is integrable, so the identity remains valid when a positive part has infinite integral. Nonnegativity of marginal KL gives the minimum, and equality requires $r_i = \mu_i$ for every i when the projection term is finite. Regular conditional laws exist on standard Borel spaces (Kallenberg, 1997). Applying the identity almost surely and integrating proves its conditional form. □

Lemma 7 (Observable branches)**.** *Let $Y = (Y_i)_i$ take values in a finite product of standard Borel spaces. Let $B_i = f_i(Y_i)$ be finite branch labels and write $B = (B_i)_i$. Suppose the labels are independent and $P_{Y_i|B} = P_{Y_i|B_i}$ almost surely for every i. Then*

$$D_{\mathrm{KL}}\left(P_Y \,\middle\|\, \bigotimes_i P_{Y_i}\right) = \mathbb{E}_B D_{\mathrm{KL}}\left(P_{Y|B} \,\middle\|\, \bigotimes_i P_{Y_i|B}\right).$$

When both assumptions hold given a standard Borel variable V, the identity holds conditionally and can be averaged over V.

Proof. Under the product reference, the branch law is $\bigotimes_i P_{B_i}$ and the conditional law of Y is $\bigotimes_i P_{Y_i|B_i}$. The KL chain rule first compares the branch laws and then the conditional laws within each branch. Independence makes the first divergence zero. The marginal locality assumption identifies the conditional reference with $\bigotimes_i P_{Y_i|B}$, proving the identity. The same argument applies to regular conditional laws given V, including infinite divergences. □

Proof of Theorem 1. Fix z_t outside a P_t null set. Let S be its revealed positions and x_S their values. Masking is independent of the data, so the posterior on the remaining coordinates is $p(X_{S^c} \in \cdot \mid x_S)$. Write $\mu = P_{s|t}(\cdot \mid z_t)$, $\mu^{\otimes} = \bigotimes_i \mu_i$, and $\beta = (\alpha_s - \alpha_t)/(1 - \alpha_t)$. Both laws copy x_S and reveal each remaining coordinate independently with probability β. A set $D \subseteq S^c$ of new coordinates has weight

$$w(D \mid S) = \beta^{|D|}(1-\beta)^{L-|S|-|D|}$$

under both laws, with $0^0 = 1$.

The data and mask alphabets are disjoint, so the new coordinates identify the branch D. Conditional on a branch of positive weight, the channel assumptions give

$$\mu(\cdot \mid D) = \delta_{x_S} \otimes p(X_D \in \cdot \mid x_S) \otimes \bigotimes_{i \notin S \cup D} \kappa^i_{s,t}(\cdot \mid z^i_t),$$

$$\mu^{\otimes}(\cdot \mid D) = \delta_{x_S} \otimes \bigotimes_{i \in D} p(X_i \in \cdot \mid x_S) \otimes \bigotimes_{i \notin S \cup D} \kappa^i_{s,t}(\cdot \mid z^i_t).$$

The factors occupy their original coordinate positions. The copied tokens and remaining mask laws agree under the two measures. Applying Lemma 7 conditionally on $Z_t = z_t$ gives

$$D_{\mathrm{KL}}(\mu\|\mu^{\otimes}) = \sum_{D\subseteq S^c} w(D \mid S) D_{\mathrm{KL}}\left(p(X_D \in \cdot \mid x_S) \,\middle\|\, \bigotimes_{i\in D} p(X_i \in \cdot \mid x_S) \right).$$

The mask refresh factors have canceled, establishing independence of the cost from the admissible refresh kernels.

Averaging the branch divergence over X_S gives $\mathrm{TC}(X_D \mid X_S)$. The probability of the retained set is $\Pr(S) = \alpha_t^{|S|}(1-\alpha_t)^{L-|S|}$. With $\Delta = \alpha_s - \alpha_t$, the full branch weight becomes

$$\Pr(S) w(D \mid S) = \alpha_t^{|S|} \Delta^{|D|} (1-\alpha_s)^{L-|S|-|D|}.$$

Summing over S proves equation 3.

Regular conditional probabilities give measurable coordinate kernels $\mu_i(\cdot \mid z_t)$, and their product attains the projection cost. The clean alphabets are finite, so the expected branch divergence is at most $\sum_i \log|\mathcal{X}_i|$. Lemma 6 then shows that any product minimizer must have zero expected marginal divergence in every coordinate. Its factors therefore agree with the true conditional marginals P_t almost surely. This proves attainment and uniqueness. □

B.2 Summing Costs over Independent Blocks

Let $\mathcal{B}$ be a partition of $[L]$ and suppose $p = \bigotimes_{B\in\mathcal{B}} p_B$. For fixed disjoint reveal sets S, D, put $S_B = S \cap B$ and $D_B = D \cap B$. Conditioning on X_S preserves independence across blocks, so conditional entropy additivity gives

$$\mathrm{TC}(X_D \mid X_S) = \sum_{B\in\mathcal{B}} \mathrm{TC}(X_{D_B} \mid X_{S_B}).$$

In Theorem 1, the branch probabilities outside a given block sum to one. Summing these probabilities gives $\Phi^p_{s,t} = \sum_{B\in\mathcal{B}} \Phi^{p_B}_{s,t}$, with every block evaluated at the same retention levels. Each coordinate has three possible memberships, so the block formula contains $3^{|B|}$ branch terms. Evaluating all blocks uses $O(\sum_B 3^{|B|})$ such terms when the partition and entropy tables are supplied.

The block formula also gives a general schedule bound. Within a block, conditional total correlation is zero for $|D_B| \le 1$ and is at most $\sum_{i\in D_B} \log|\mathcal{X}_i|$ otherwise. After averaging over S, the new reveal indicators are independent with probability $\Delta = \alpha_s - \alpha_t$. For each $i \in B$,

$$\Pr\big(i \in D,\ D \cap (B \setminus \{i\}) \neq \varnothing\big) = \Delta\big[1 - (1-\Delta)^{|B|-1}\big] \le (|B|-1)\Delta^2.$$

Multiplying these probabilities by the coordinate entropy bounds yields

$$\Phi^p_{s,t} \le C_{\mathcal{B}} \Delta^2, \qquad C_{\mathcal{B}} = \sum_{B\in\mathcal{B}} (|B|-1) \sum_{i\in B} \log|\mathcal{X}_i|.$$

For a strict schedule with increments $\Delta_t = \alpha_{t-1} - \alpha_t$, the cumulative cost is at most $C_{\mathcal{B}} \sum_t \Delta_t^2$. Uniform increments give $C_{\mathcal{B}}/T$. If every block has at most b coordinates and every alphabet has size at most A, this schedule has conditional cost per coordinate at most $(b-1)\log A/T$.

B.3 Decomposing the Path Divergence

Proof of equation 4. The true and sampler path laws factor as

$$P(dz_{0:T}) = P_T(dz_T) \prod_{t=1}^{T} P_{t-1|t}(dz_{t-1} \mid z_t), \qquad Q(dz_{0:T}) = Q_T(dz_T) \prod_{t=1}^{T} R_t(dz_{t-1} \mid z_t).$$

Applying the relative entropy chain rule (Cover et al., 1991; Kallenberg, 1997) along the trajectory gives

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(P_T\|Q_T) + \sum_{t=1}^{T} \mathbb{E}_{Z_t\sim P_t} D_{\mathrm{KL}}\big(P_{t-1|t}(\cdot \mid Z_t) \,\big\|\, R_t(\cdot \mid Z_t)\big).$$

This identity holds in the extended nonnegative reals. Failure of absolute continuity at initialization or in a transition on a set of positive P_t measure makes the corresponding term and the path divergence infinite.

For $R_t = \bigotimes_i r_{t,i}$, Lemma 6 decomposes each transition divergence into its product projection cost and the sum of marginal divergences. The expected projection cost is $\Phi^p_{t-1,t}$ by Theorem 1. The expected marginal remainders are exactly the $e_{t,i}$ defined in Section 3.1. Substitution proves equation 4, including infinite values. The expectations use P_t because the chain rule integrates under the true path law P.

The map from a trajectory to Z_0 sends P to p and Q to Q_0. Data processing gives $D_{\mathrm{KL}}(p\|Q_0) \le D_{\mathrm{KL}}(P\|Q)$, completing the proof. □

For probability laws μ, ν, write $\mathrm{TV}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$. Pinsker's inequality (Cover et al., 1991) gives

$$\mathrm{TV}(p, Q_0) \le \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P\|Q)}.$$

If $D_{\mathrm{KL}}(P_T\|Q_T) \le \varepsilon_T$ and $e_{t,i} \le \varepsilon_{t,i}$, the path divergence is at most $\varepsilon_T + \sum_t \Phi^p_{t-1,t} + \sum_{t,i} \varepsilon_{t,i}$. Substitution gives corresponding output KL and TV bounds. For any measurable f taking values in $[a,b]$, total variation also gives $|\mathbb{E}_p f - \mathbb{E}_{Q_0} f| \le (b-a)\,\mathrm{TV}(p, Q_0)$.

B.4 Computing the Pair Output Distribution

Proof of Theorem 2. The target $p^{\mathrm{pair}}_{\eta,L}$ has full support because $\eta \in (0,1)$. On every reveal branch of positive probability, both the true conditional and its product projection assign positive probability to every new binary configuration. They share the copied tokens and mask laws. Their conditional laws are therefore mutually absolutely continuous for P_t almost every state. Starting from $Q^*_T = P_T$, induction gives $Q^*_t \ll P_t$. Conditional kernels specified on P_t null sets do not affect the generated law.

Let $J_i = t$ mean that coordinate i is first revealed in the transition from t to $t-1$. For a masked coordinate, the reveal probability is $\beta_t = \Delta_t/(1-\alpha_t)$. These probabilities depend only on the schedule, so reveal decisions can be drawn independently across coordinates and steps, independently of the draws used for tokens and masks. The probability of remaining masked at time t is

$$\prod_{u=t+1}^{T} (1-\beta_u) = \prod_{u=t+1}^{T} \frac{1-\alpha_{u-1}}{1-\alpha_u} = 1-\alpha_t.$$

Consequently $\Pr(J_i = t) = (1-\alpha_t)\beta_t = \Delta_t$. The reveal times are independent, giving $\Pr(J_{2j-1} = J_{2j}) = \sum_t \Delta_t^2 = \rho$ for each pair.

If both bits are revealed together, neither is available as context. Their true coordinate conditionals are fair, and the product update generates a uniform pair. If the reveal times differ, the first bit is fair and the second follows its conditional under g_η. The resulting pair law is then exactly g_η. Averaging these cases gives

$$(1-\rho) g_\eta + \rho\,\mathrm{Unif}(\{0,1\}^2) = g_{\eta+(1-\eta)\rho}.$$

Each data conditional depends only on its revealed partner, and each mask refresh kernel depends only on its own current mask value. Independent initialization and independent draws preserve independence across the pair blocks. Any unpaired bit remains independent and fair. Their product is the output law in equation 6. □

B.5 Optimizing the Pair Schedule

Proof of Corollary 3. Within one pair, the conditional total correlation is c_η exactly when both coordinates are newly revealed together. This event has probability Δ_t^2 in step t, and every other branch has zero cost. Block additivity gives $\Phi^p_{t-1,t} = m c_\eta \Delta_t^2$. The canonical sampler has matching initialization and exact marginal heads, so equation 4 yields $D_{\mathrm{KL}}(P\|Q^*) = m c_\eta \rho$.

For output KL, set $a = \eta/2$ and $b = \eta_\rho/2 = a + (1-\eta)\rho/2$. Under both g_η and g_{η_ρ}, the two configurations with a given mismatch indicator are equally likely. The KL chain rule reduces the pair

divergence to $d_2(a\|b)$. Theorem 2 and independence across pairs give $D_{\mathrm{KL}}(p\|Q_0^*) = md_2(a\|b)$, proving equation 7.

The increments are positive and sum to one. Cauchy's inequality gives $\rho = \sum_t \Delta_t^2 \geq 1/T$, with equality exactly when every $\Delta_t = 1/T$. Path error is strictly increasing in ρ because $mc_\eta > 0$. Output error is also strictly increasing, since

$$\frac{\partial}{\partial b} d_2(a\|b) = \frac{b-a}{b(1-b)} > 0 \qquad (b > a).$$

Uniform increments therefore uniquely minimize both errors. Their substitution gives the exact optimal output value in equation 9. For arbitrary product heads with $Q_T = P_T$, the additional errors $e_{t,i}$ are nonnegative. The minimum path KL over these heads and all strict schedules is thus mc_η/T.

For the output bounds, the total variation between the pair laws is $b - a = (1-\eta)\rho/2$. Pinsker's inequality gives $d_2(a\|b) \geq 2(b-a)^2$. The inequality $\log u \leq u-1$ gives the upper bound $d_2(a\|b) \leq (a-b)^2/[b(1-b)]$. Since $\eta/2 \leq b \leq 1/2$, the denominator is at least $\eta/4$. Multiplying by the number of pairs gives, for every schedule,

$$\frac{m(1-\eta)^2\rho^2}{2} \leq D_{\mathrm{KL}}(p\|Q_0^*) \leq \frac{m(1-\eta)^2\rho^2}{\eta}.$$

Setting $\rho = 1/T$ completes equation 9. □

Marginalization to one pair gives $\mathrm{TV}(p, Q_0^*) \geq (1-\eta)\rho/2$. For the optimal schedule, an output KL budget $\varepsilon > 0$ requires $T \geq (1-\eta)\sqrt{m/(2\varepsilon)}$, and $T \geq (1-\eta)\sqrt{m/(\eta\varepsilon)}$ is sufficient. These conditions apply to positive integer step counts. At $T = 1$, every generated bit is independent and fair, and both errors equal mc_η. The rate constants assume fixed $\eta \in (0, 1)$.

B.6 Weighting the Pair Transition Costs

Fix the pair target of Section 3.2, a positive integer T, and weights $w_t > 0$ for $t = 1, \ldots, T$. Since $\Phi^p_{t-1,t} = mc_\eta\Delta_t^2$, the exact weighted optimum over strict schedules is

$$\min_{\{\alpha_t\}} \sum_{t=1}^{T} w_t \Phi^p_{t-1,t} = \frac{mc_\eta}{\sum_{t=1}^{T} w_t^{-1}}, \qquad \Delta_t = \frac{w_t^{-1}}{\sum_{s=1}^{T} w_s^{-1}}.$$

Indeed, Cauchy's inequality applied to $1 = \sum_t (\sqrt{w_t}\Delta_t)/\sqrt{w_t}$ gives

$$1 \leq \left(\sum_t w_t \Delta_t^2\right)\left(\sum_t w_t^{-1}\right).$$

Equality requires the displayed increments. They are positive and sum to one, so they define the unique admissible optimum.

With unit weights, the minimum cumulative conditional error per coordinate is $(m/L)c_\eta/T$. For a total conditional error budget $\varepsilon > 0$, the smallest integer step count is $\lceil mc_\eta/\varepsilon \rceil$. A budget $\bar{\varepsilon} > 0$ per coordinate requires $T \geq mc_\eta/(L\bar{\varepsilon})$. These expressions concern conditional error and the output budgets for the canonical sampler are given in Appendix B.5.

B.7 Generating the Output from Shared Randomness

Let $U \sim \nu$ be the fully masked state in a single transition, with ν atomless and independent of $X \sim p$. Enumerate the finite clean alphabet as $x^{(1)}, \ldots, x^{(N)}$. Atomless measure splitting gives a measurable partition $A_1, \ldots, A_N$ with $\nu(A_j) = p(x^{(j)})$ (Bogachev & Ruas, 2007). Define $F(u) = x^{(j)}$ on A_j and the product kernel $R_F(x \mid u) = \prod_i \mathbf{1}\{x_i = F_i(u)\}$. Every coordinate head can access the complete state u, as allowed in Section 2.

The output law is

$$\overline{R}_F(x^{(j)}) = \int R_F(x^{(j)} \mid u)\, d\nu(u) = \nu(A_j) = p(x^{(j)}).$$

The minimum output KL over measurable product kernels is therefore zero and is attained in one step. The finite measurable partition makes every coordinate head measurable.

Under the specified forward coupling, the true law is $P(dx, du) = p(dx)\nu(du)$ and the generator law is $Q_F(dx, du) = \delta_{F(u)}(dx)\nu(du)$. If p has at least two support points, the point mass $\delta_{F(u)}$ omits a point with positive target probability for every u. Hence $D_{\mathrm{KL}}(p\|\delta_{F(u)}) = +\infty$, and the conditional chain rule gives $D_{\mathrm{KL}}(P\|Q_F) = +\infty$.

The minimum path error for this problem with one transition is $\mathrm{TC}(X)$ by Lemma 6. It is attained by returning the product of data marginals at every u, whose output error is also $\mathrm{TC}(X)$. For the pair family this value is mc_η. Atomlessness guarantees a partition for every finite target and an atomic mask law can realize a particular target if its atom masses admit the required grouping.

C Proof Details for Preserving Dependence

We give the complete proofs for Section 4. The supplementary results bound the optimal parity cost, quantify approximate mixture requirements, and verify exact generation when the dependency blocks are known.

C.1 Computing the Parity Transition Cost

Derivation of equation 12. Fix an integer $k \geq 2$ and $\lambda \in (0, 1)$. For one block with law $\gamma_{\lambda,k}$, marginalizing any coordinate cancels the two signs of the parity term. Every proper marginal is therefore uniform product. The even and odd parity classes have probabilities $(1 + \lambda)/2$ and $(1 - \lambda)/2$, and each contains 2^{k-1} equiprobable states. The full entropy is $(k-1)\log 2 + h_2((1-\lambda)/2)$, giving total correlation c_λ.

Within this block, let $S, D \subseteq [k]$ denote the disjoint sets of observed and newly revealed coordinates. If $S \cup D$ is proper, its marginal is uniform product, so $\mathrm{TC}(X_D \mid X_S) = 0$. The cost is also zero whenever $|D| \leq 1$. It remains to consider $S \cup D = [k]$ and $d = |D| \geq 2$. Since the marginal of X_S is uniform, the conditional law is

$$p(x_D \mid x_S) = 2^{-d}\left[1 + \lambda(-1)^{\sum_{i\in S} x_i}(-1)^{\sum_{i\in D} x_i}\right].$$

Each coordinate marginal is fair. The entropy is $(d-1)\log 2 + h_2((1-\lambda)/2)$ for every x_S, since reversing the parity bias preserves binary entropy. Subtracting this entropy from $d \log 2$ gives conditional total correlation c_λ.

Write $\Delta = \alpha_s - \alpha_t$. A contributing branch with d new coordinates has probability $\alpha_t^{k-d}\Delta^d$, since the other $k - d$ coordinates are already observed. There are $\binom{k}{d}$ choices. Theorem 1 gives the cost of one block as

$$c_\lambda \sum_{d=2}^{k} \binom{k}{d} \alpha_t^{k-d}\Delta^d = c_\lambda\left[(\alpha_t + \Delta)^k - \alpha_t^k - k\alpha_t^{k-1}\Delta\right].$$

Applying block additivity to n independent copies and using $\alpha_t + \Delta = \alpha_s$ proves equation 12. □

C.2 Solving the Parity Schedule Problem

Proof of Corollary 4. Recall $B_k(b, a) = b^k - a^k - ka^{k-1}(b - a)$. For $z \in (0, 1]$, let $V_h(z)$ be the infimum of $\sum_{j=1}^{h} B_k(x_j, x_{j-1})$ over $0 = x_0 < \cdots < x_h = z$. We prove by induction that this infimum is attained uniquely and equals $C_{k,h}z^k$. For $h = 1$, the only partition has cost $B_k(z, 0) = z^k$, so $C_{k,1} = 1$.

Assume the assertion for $h - 1$ and write the penultimate node as yz, with $0 < y < 1$. Homogeneity gives $V_{h-1}(yz) = C_{k,h-1}y^k z^k$. Adding the final transition gives the Bellman recursion (Bellman, 1954)

$$\begin{aligned} V_h(z) &= \inf_{0<y<1}\{V_{h-1}(yz) + B_k(z, yz)\} \\ &= z^k \inf_{0<y<1} f_h(y), \qquad f_h(y) = 1 + (C_{k,h-1} + k - 1)y^k - ky^{k-1}. \end{aligned}$$

For $0 < y < 1$,

$$f_h'(y) = ky^{k-2}\big[(C_{k,h-1} + k - 1)y - (k-1)\big].$$

Since $C_{k,h-1} > 0$, the derivative changes sign exactly once, at $r_h = (k-1)/(k-1+C_{k,h-1}) \in (0,1)$. The function decreases before r_h and increases after it. Its unique minimum is

$$f_h(r_h) = 1 - r_h^{k-1} = C_{k,h} > 0.$$

The minimizing penultimate node is interior, and the preceding partition is uniquely attained by the induction hypothesis. This proves existence, strictness, uniqueness, and the recurrence for every h.

Starting from $x_T = 1$, each preceding optimal node satisfies $x_{h-1} = r_h x_h$. Iteration gives $x_j = \prod_{h=j+1}^{T} r_h$ for $1 \le j \le T$, with $x_0 = 0$. All ratios lie in $(0,1)$, so these nodes satisfy the strict schedule constraints. Reversing their order gives equation 14, and multiplying the normalized optimum by nc_λ proves equation 13.

For $k = 2$, induction in the recurrence gives $C_{2,h} = 1/h$ and $r_h = (h-1)/h$, hence $x_j = j/T$. For $k > 2$ and $T \ge 2$, the first two positive nodes satisfy $x_1/x_2 = r_2 = (k-1)/k > 1/2$. A uniform partition would have $x_1/x_2 = 1/2$, proving nonuniformity. □

C.3 Bounding the Optimal Parity Cost

For every integer $k \ge 2$ and positive integer T, the optimal normalized cost satisfies

$$\frac{1}{T} \le C_{k,T} \le \frac{k(k-1)}{2T}, \qquad \lim_{T\to\infty} TC_{k,T} = \frac{2(k-1)}{k} \quad \text{for fixed } k.$$

Multiplication by nc_λ converts these statements to cumulative conditional error for n parity blocks.

For the lower bound, we first show $B_k(b,a) \ge (b^{k/2} - a^{k/2})^2$ for $0 \le a \le b \le 1$. The cases $b = 0$ and $k = 2$ are immediate. For $b > 0$ and $k > 2$, put $r = a/b$. Dividing the difference by b^k gives

$$r^{k/2}\big[2 - kr^{(k-2)/2} + (k-2)r^{k/2}\big].$$

The bracket is zero at $r = 1$ and has derivative $k(k-2)r^{(k-4)/2}(r-1)/2 \le 0$ on $(0,1)$. It is nonnegative on $[0,1]$. For any partition from zero to one, Cauchy's inequality now gives

$$\sum_{j=1}^{T} B_k(x_j, x_{j-1}) \ge \sum_{j=1}^{T} (x_j^{k/2} - x_{j-1}^{k/2})^2 \ge \frac{1}{T}.$$

For the upper bound, the second derivative of u^k is at most $k(k-1)$ on $[0,1]$. The integral remainder formula gives

$$B_k(b,a) = \int_a^b (b-u)k(k-1)u^{k-2}\,du \le \frac{k(k-1)}{2}(b-a)^2.$$

Using $x_j = j/T$ and summing over the T transitions yields the stated upper bound.

To obtain the limit, define $g(c) = (1 + c/(k-1))^{-(k-1)}$. Strict convexity and the tangent at zero give $g(c) > 1 - c$ for $c > 0$. The recurrence therefore satisfies $0 < C_{k,h} < C_{k,h-1}$. Its only nonnegative fixed point is zero, so $C_{k,h} \to 0$. Taylor expansion of the recurrence at zero gives

$$C_{k,h} = C_{k,h-1} - \frac{k}{2(k-1)} C_{k,h-1}^2 + O(C_{k,h-1}^3).$$

It follows that

$$C_{k,h}^{-1} - C_{k,h-1}^{-1} \longrightarrow \frac{k}{2(k-1)}.$$

Summing these reciprocal increments and dividing by T gives $C_{k,T}^{-1}/T \to k/[2(k-1)]$. Taking reciprocals proves the limit.

C.4 Counting Exact Mixture Components

Proof of Theorem 5. Fix $L \geq 2$, $m = \lfloor L/2 \rfloor$, and $\eta \in (0, 1)$. Marginalize any unpaired coordinate and let $A = (X_1, X_3, \ldots, X_{2m-1})$ and $B = (X_2, X_4, \ldots, X_{2m})$. One pair has probability matrix

$$G_\eta = \frac{1}{4} \begin{pmatrix} 2-\eta & \eta \\ \eta & 2-\eta \end{pmatrix}.$$

Its determinant is $(1-\eta)/4 > 0$, so its rank is two. The joint probability matrix of (A, B) is $G_\eta^{\otimes m}$ and has rank 2^m.

For a distribution in $\mathcal{Q}_K$, define $u_\ell(a) = \prod_{j=1}^m q_{2j-1,\ell}(a_j)$ and $v_\ell(b) = \prod_{j=1}^m q_{2j,\ell}(b_j)$. Marginalizing any extra coordinate leaves the matrix $\sum_{\ell=1}^K w_\ell u_\ell v_\ell^\top$. Each summand has rank at most one, so the matrix has rank at most K (Cohen & Rothblum, 1993). Exact representation of the pair target requires $K \geq 2^m$.

For attainment, index components by $a \in \{0,1\}^m$, each with weight 2^{-m}. Define

$$q_a(x) = 2^{-(L-2m)} \prod_{j=1}^m \left[\mathbf{1}\{x_{2j-1} = a_j\}\, 2 g_\eta(a_j, x_{2j}) \right].$$

The odd coordinates are fixed, and $2g_\eta(a_j, \cdot)$ is the conditional distribution of the partner given the fair first bit. The factor $2^{-(L-2m)}$ supplies an independent fair bit when L is odd. Thus each q_a is a normalized coordinate product. For every x, only the index $a = (x_1, x_3, \ldots, x_{2m-1})$ contributes to the mixture, yielding

$$2^{-m} \sum_{a \in \{0,1\}^m} q_a(x) = 2^{-(L-2m)} \prod_{j=1}^m g_\eta(x_{2j-1}, x_{2j}) = p_{\eta,L}^{\text{pair}}(x).$$

This attains 2^m components. Zero weights permit every larger K, completing the equivalence. □

C.5 Bounding Approximate Mixture Size

For an exact discrete latent representation of the pair target, let C be a shared index given which the coordinates are independent. Use the odd and even vectors A, B from Appendix C.4. Conditional independence gives $I(A; B) \leq I(A; C) \leq H(C)$. Independence across the target pairs gives $I(A; B) = mc_\eta$, hence $H(C) \geq mc_\eta$. If C has at most K values, then $\log K \geq mc_\eta$. Theorem 5 gives the sharper exact cardinality requirement through matrix rank.

For approximation, suppose $q \in \mathcal{Q}_K$ satisfies $\mathrm{TV}(p_{\eta,L}^{\text{pair}}, q) \leq \delta \leq 1/2$, with $m \geq 1$. Marginalize any unpaired bit. Under q, the shared component index gives $I_q(A; B) \leq \log K$. Entropy continuity on an alphabet of size d gives $|H(\mu) - H(\nu)| \leq \delta \log d + h_2(\delta)$ at total variation at most δ (Audenaert, 2007). Applying this bound to A, B, and (A, B) uses alphabet sizes 2^m, 2^m, and 2^{2m}. Since marginalization contracts total variation,

$$|I_p(A; B) - I_q(A; B)| \leq 4m\delta \log 2 + 3h_2(\delta), \qquad \log K \geq mc_\eta - 4m\delta \log 2 - 3h_2(\delta).$$

If $D_{\mathrm{KL}}(p_{\eta,L}^{\text{pair}} \| q) \leq \varepsilon \leq 1/2$, Pinsker's inequality gives total variation at most $\sqrt{\varepsilon/2}$. The continuity correction is increasing on $[0, 1/2]$, so

$$\log K \geq mc_\eta - 4m\sqrt{\varepsilon/2} \log 2 - 3h_2(\sqrt{\varepsilon/2}).$$

For fixed η, a fixed TV tolerance $\delta < c_\eta/(4 \log 2)$ leaves a positive coefficient of m in this lower bound. A fixed KL tolerance $\varepsilon < c_\eta^2/[8(\log 2)^2]$ has the same property. These bounds certify exponential component requirements as m grows under the stated approximation criteria.

C.6 Generating Known Dependency Blocks

Let $\mathcal{B}$ partition the coordinates, and let each head generate the coordinates in one block jointly. Applying Lemma 6 to the block variables gives

$$\inf_{\{r_B\}} D_{\mathrm{KL}}\left(p \,\middle\|\, \prod_{B \in \mathcal{B}} r_B \right) = \sum_{B \in \mathcal{B}} H(X_B) - H(X).$$

The minimum is attained by the true block marginals. For the pair target, choosing each pair as a head block makes the cost zero. A singleton head generates any extra fair bit. Independence across the target blocks then gives exact generation in one round.

The two conditional rounds in equation 16 also generate the pair target exactly. The first round draws the odd vector A uniformly and samples any unpaired bit independently. Conditional on $A = a$, the second round has distribution $2g_\eta(a_j, \cdot)$ for partner j and copies the odd bits. These are precisely the components q_a in Appendix C.4, selected with weights 2^{-m}. Their mixture is the target, and each round is a coordinate product conditional on its input state.

For $p = \gamma_{\lambda,k}^{\otimes n}$, every proper marginal of a target block is uniform product. If each head block has size at most $k-1$, its intersection with every target block is proper. Independence of the target blocks makes each head marginal uniform product. The optimal block projection cost is then $L \log 2 - H(p) = nc_\lambda$. Choosing head blocks that coincide with the target blocks gives cost zero.

Two conditional rounds suffice for the parity target as well. First sample the initial $k-1$ bits of every target block independently and fairly. Given these values $a \in \{0,1\}^{k-1}$ in one block, sample its last bit b with probability

$$\Pr(b \mid a) = \frac{1}{2}\left[1 + \lambda(-1)^{a_1+\cdots+a_{k-1}+b}\right].$$

Multiplying by $2^{-(k-1)}$ gives $\gamma_{\lambda,k}(a,b)$. The last bits are independent across blocks conditional on the first round, so the full output law is p. These constructions use known dependency blocks and exact target conditionals. Their prescribed reveal order differs from the independent reveal decisions in Section 2.

D Additional Results on Conditional Dependence

D.1 Product Projection with Side Information

Let X take values in the finite product alphabet $\mathcal{X}$ and let U be a standard Borel variable with an arbitrary joint law with X. Write $p_u = P_{X|U=u}$ and $p_{i,u} = P_{X_i|U=u}$. Side information changes the product approximation cost through the conditional dependence that remains after observing U.

Proposition 8 (Conditional product projection)**.** *For measurable coordinate kernels $r_i(\cdot \mid U)$,*

$$\mathbb{E}_U D_{\mathrm{KL}}\left(p_U \,\middle\|\, \prod_i r_i(\cdot \mid U)\right) = \mathrm{TC}(X \mid U) + \sum_i \mathbb{E}_U D_{\mathrm{KL}}(p_{i,U} \| r_i(\cdot \mid U)), \tag{17}$$

where $\mathrm{TC}(X \mid U) = \sum_i H(X_i \mid U) - H(X \mid U)$. The minimum equals $\mathrm{TC}(X \mid U)$ and is attained precisely by the conditional marginals, up to P_U null sets.

Proof. Apply Lemma 6 to p_u at almost every u and integrate. Finite clean alphabets bound the projection term by $\sum_i \log|\mathcal{X}_i|$, so equality at the minimum requires every nonnegative marginal remainder to vanish. Regular conditional probabilities provide measurable attaining kernels. The identity also covers a reference kernel that assigns zero probability to an outcome with positive conditional target probability. □

If U is independent of X, the minimum is $\mathrm{TC}(X)$. If $X = F(U)$ almost surely, the conditional law is a product of coordinate point masses and the minimum is zero. These two cases can have identical X marginals. Their conditional objectives differ because their joint laws differ, as detailed in Appendix E.

D.2 Conditional Total Correlation as Entropy Curvature

Define the entropy set function $h(A) = H(X_A)$, with $h(\varnothing) = 0$. For disjoint sets $S, D \subseteq [L]$, define

$$\Gamma_h(S, D) = \sum_{i \in D}\big[h(S \cup \{i\}) - h(S)\big] - \big[h(S \cup D) - h(S)\big] = \mathrm{TC}(X_D \mid X_S). \tag{18}$$

The equality follows by expanding each conditional entropy. Nonnegativity follows from conditional entropy subadditivity, and $\Gamma_h(S, D) = 0$ for $|D| \leq 1$. For two newly revealed coordinates it is their conditional mutual information. Larger reveal sets measure the gap between the sum of individual information increments and their joint increment.

Theorem 1 averages these quantities over the observable reveal branches. A branch with $|D| = d$ has a factor Δ^d in its probability. The other factors account for the already revealed and still masked coordinates. This representation explains why pair blocks give an exact quadratic collision cost, while parity blocks retain information about the current retention level. The computation for a known independent block partition is given in Appendix B.2.

D.3 Clean Reconstruction under Independent Retention

Let Z_α retain each clean coordinate independently with probability α, using mask randomness independent of X. Consider reconstruction of the complete clean vector through a product kernel, with loss

$$\mathcal{L}_\alpha(R) = \mathbb{E}_{Z_\alpha} D_{\mathrm{KL}}\left(P_{X|Z_\alpha} \,\middle\|\, \prod_i r_i(\cdot \mid Z_\alpha) \right).$$

Proposition 9 (Exact reconstruction cost)**.** *For $0 \leq \alpha \leq 1$, the minimum is attained by the true conditional coordinate marginals and equals*

$$\inf_R \mathcal{L}_\alpha(R) = \sum_{S \subseteq [L]} \alpha^{|S|}(1-\alpha)^{L-|S|} \mathrm{TC}(X_{S^c} \mid X_S). \tag{19}$$

Its values at $\alpha = 0$ and $\alpha = 1$ are $\mathrm{TC}(X)$ and zero, respectively.

Proof. The observed state identifies the retained set S and its values X_S. Independence of the mask channel leaves the posterior on the remaining coordinates equal to $P_{X_{S^c}|X_S}$. The retained coordinates are deterministic in that conditional law and contribute zero to its total correlation. Proposition 8 therefore gives the displayed sum after averaging over S. The endpoint values follow from full masking and full observation. □

Keeping only the nonnegative term $S = \varnothing$ gives $\inf_R \mathcal{L}_\alpha(R) \geq (1-\alpha)^L \mathrm{TC}(X)$. Over $\alpha \in [1/4, 3/4]$, this implies the uniform lower bound $4^{-L}\mathrm{TC}(X)$. Its exponential coefficient reflects the probability of observing no clean coordinates. Positivity of this bound alone does not imply a cost proportional to length in the interior of the retention interval.

For the pair family, a pair contributes c_η precisely when both of its coordinates remain masked. Independence across pairs therefore gives

$$\inf_R \mathcal{L}_\alpha(R) = m(1-\alpha)^2 c_\eta \geq \frac{m c_\eta}{16}, \qquad \alpha \in [1/4, 3/4].$$

For fixed $\eta \in (0, 1)$, this provides a cost proportional to the number of pairs throughout that interval. It follows from dependence within each masked pair and includes branches on which other coordinates are visible.

D.4 Stability under Perturbations of the Target

Proposition 10 (Continuity of total correlation)**.** *Let p, q be laws on $\mathcal{X} = \prod_{i=1}^L \mathcal{X}_i$, with $d = |\mathcal{X}| \geq 2$ and $\mathrm{TV}(p, q) = \delta \leq 1/2$. Then*

$$|\mathrm{TC}(p) - \mathrm{TC}(q)| \leq 2\delta \log d + (L+1) h_2(\delta). \tag{20}$$

Proof. The classical case of entropy continuity gives $|H(\mu) - H(\nu)| \leq \delta \log d + h_2(\delta)$ for laws on d points at total variation at most $\delta \leq 1/2$ (Audenaert, 2007). Marginalization contracts total variation. Apply the bound to the joint law and to each nonsingleton coordinate alphabet, using zero for a singleton alphabet. The joint entropy contributes $\delta \log d + h_2(\delta)$, while the sum of coordinate entropies contributes at most $\delta \sum_i \log |\mathcal{X}_i| + L h_2(\delta)$. Since $\sum_i \log |\mathcal{X}_i| = \log d$, their sum proves the claim. □

If a fully masked variable remains independent of a perturbed target q, its optimal conditional product error is consequently at least

$$\big[\mathrm{TC}(p) - 2\delta \log d - (L+1)h_2(\delta)\big]_+, \qquad [v]_+ = \max\{v, 0\}.$$

This estimate quantifies how a positive error survives distributional perturbations. Its dependence on alphabet size remains explicit and can weaken the guarantee as dimension grows.

D.5 A Globally Correlated Distribution with Full Support

Consider a smoothed distribution concentrated near two opposite binary strings,

$$p_{\epsilon,L} = (1-\epsilon)\,\mathrm{Unif}\{0^L, 1^L\} + \epsilon\,\mathrm{Unif}(\{0,1\}^L), \qquad 0 < \epsilon < 1.$$

Every coordinate is fair and the distribution has full support. With $N = 2^L$, put $a = (1-\epsilon)/2 + \epsilon/N$ and $b = \epsilon/N$. Direct entropy evaluation gives

$$\mathrm{TC}(p_{\epsilon,L}) = L\log 2 + 2a\log a + (N-2)b\log b \geq (1-\epsilon)(L-1)\log 2 - h_2(\epsilon). \tag{21}$$

To prove the inequality, introduce the Bernoulli variable selecting the mixture component. The entropy of X is at most the entropy of this variable plus the conditional entropy of X. This gives $H(X) \leq h_2(\epsilon) + (1-\epsilon)\log 2 + \epsilon L \log 2$. Subtracting from the sum $L\log 2$ of marginal entropies gives equation 21. For $\epsilon = 1/4$ and $L \geq 4$, the bound is at least $L\log 2/4$.

The unsmoothed distribution has total correlation $(L-1)\log 2$. At $L = 2$, any product law with probabilities s, t of the value one satisfies

$$D_{\mathrm{KL}}(\mathrm{Unif}\{00, 11\} \,\|\, \mathrm{Ber}(s) \otimes \mathrm{Ber}(t)) = \log 2 + d_2(1/2\|s) + d_2(1/2\|t).$$

At the fully masked endpoint with noise independent of the data, the minimal conditional cost is $\log 2$. A balanced deterministic partition of atomless noise still generates the same two strings exactly, as in Appendix B.7. The full support family above preserves the separation between conditional product approximation and marginal generation without relying on zero target probabilities.

E Coupling and Channel Assumptions

E.1 Two Joint Laws with the Same Output Marginal

The transport construction in Appendix B.7 gives a measurable map F with $F_{\#}\nu = p$. It defines two relevant joint laws,

$$P_{\mathrm{ind}}(dx, du) = p(dx)\nu(du), \qquad P_F(dx, du) = \delta_{F(u)}(dx)\nu(du). \tag{22}$$

Both have data marginal p and noise marginal ν. Their conditional data laws are p and $\delta_{F(u)}$, respectively. Their mutual informations satisfy $I_{P_{\mathrm{ind}}}(X;U) = 0$ and $I_{P_F}(X;U) = H(p)$. Proposition 8 gives conditional product minima $\mathrm{TC}(p)$ and zero under these two couplings.

If p has at least two support points, evaluating the deterministic kernel under P_{ind} gives infinite conditional KL, although its output divergence is zero. The expectation is taken under the fixed independent forward law, so matching an output marginal does not alter that conditional target. Convexity of relative entropy in its second argument gives, for every measurable family of output laws,

$$D_{\mathrm{KL}}\left(p \,\middle\|\, \int R(\cdot \mid u)\, d\nu(u)\right) \leq \int D_{\mathrm{KL}}(p\|R(\cdot \mid u))\, d\nu(u). \tag{23}$$

The inequality permits a zero left side and an infinite right side. A conditional lower bound therefore supplies no output lower bound without additional structure on the sampler.

E.2 Definitions in Infinite Mask Diffusion

We compare the fixed definitions in arXiv:2605.10518v1 (Yoo et al., 2026). Its Equation 6 places conditional KL inside an expectation. Equation 12 specifies the sitewise corruption marginal $q_\alpha(dz_i \mid x_i) = \alpha\delta_{x_i} + (1-\alpha)\pi_{\mathcal{M}}$, and Equation 13 gives copy, reveal, and refresh transitions.

Theorem 4.2 and Theorem A.3 state a zero factorization error result, with a deterministic transport construction in Appendix A.2. The following comparison concerns compatibility with the independent sitewise forward coupling.

At the fully masked endpoint, independent use of Equation 12 makes the complete mask state independent of X. A transition revealing all coordinates therefore has conditional target p at almost every mask state. For any product reverse kernel, Proposition 8 gives

$$\mathbb{E}_U D_{\mathrm{KL}}\left(p \,\middle\|\, \prod_i r_i(\cdot \mid U)\right) \geq \mathrm{TC}(p).$$

This value is strictly positive for any dependent finite target. Refresh choices cannot change this endpoint posterior under the specified independent coupling.

A deterministic map selected to have pushforward p provides exact marginal generation. Its conditional law $\delta_{F(u)}$ belongs to the coupling P_F in equation 22. Substituting that conditional for p changes the joint law used in the error objective. The construction is therefore valid as a marginal representation result, while zero conditional factorization error requires the corresponding transport coupling. This observation concerns the stated mathematical objects and leaves empirical performance to its own evaluation.

E.3 Order of Averaging in Conditional and Marginal Estimates

For noise samples $u_1, \ldots, u_n$ and the same conditional kernels $R_j = R(\cdot \mid u_j)$, the two empirical objectives are

$$\widehat{\mathcal{C}}_n = \frac{1}{n}\sum_{j=1}^{n}\sum_x p(x)\log\frac{p(x)}{R_j(x)}, \qquad \widehat{\mathcal{M}}_n = \sum_x p(x)\log\frac{p(x)}{n^{-1}\sum_j R_j(x)}. \tag{24}$$

Jensen's inequality gives $\widehat{\mathcal{M}}_n \leq \widehat{\mathcal{C}}_n$ for every realized sample. Computing KL after averaging the conditional probabilities estimates the output divergence. Averaging separately computed conditional divergences estimates the conditional objective under independent noise.

If a conditional kernel vanishes at a state with positive target probability, its conditional KL is infinite. Adding a pseudocount to a histogram of generated outputs regularizes an empirical output law. Adding coordinate noise to each deterministic head changes the conditional kernel itself. Appendix F.4 identifies the two constructions used in the archived diagnostic.

The noise averaging procedure reported for Table 1 of the cited IMDM version uses $10{,}000$ noise samples before computing the divergence (Yoo et al., 2026). That order corresponds to a marginal comparison. Relating such a measurement to an expected conditional KL requires retaining the individual conditional distributions and evaluating the logarithm before averaging.

E.4 Scope of the Channel and Representation Results

The data alphabet is finite, while the mask space may be any standard Borel space disjoint from it. Disjointness makes copy, reveal, and mask branches observable. Independent reveal labels and coordinate local refresh kernels are used in Lemma 7. Data dependent masks, hidden branch labels, or correlated refreshes require their own conditional analysis. The clean marginal at a site alone does not specify the full forward joint law.

All divergences use the direction from the true law to the approximation. Missing reference support yields an infinite divergence and remains covered by the projection and chain rule identities. Product targets and a single clean coordinate have zero structural cost. With general side information, the cost can vanish whenever the target coordinates become conditionally independent. These cases are consistent with Proposition 8.

The schedules in the main text have positive increments. A zero increment contributes zero transition cost and can be removed when counting effective steps. The weighted optimum in Appendix B.6 assumes strictly positive weights. The pair rate constants assume fixed $\eta \in (0, 1)$ and at $\eta = 1$ the target is product, while $\eta = 0$ lies outside the stated rate bounds. The direct output theorem also

uses matching initialization and exact conditional marginal heads. The path decomposition records the additional terms when either assumption changes.

Atomlessness guarantees a partition with arbitrary finite target masses. A purely atomic noise law supports a deterministic map to a given target precisely when its atom masses can be grouped to match that target. An exact finite mixture counts its explicit shared components. Unrestricted measurable heads, known joint blocks, and multiple conditional rounds provide different representation resources. The component lower bound does not impose a general bound on neural parameter count, running time, or the difficulty of learning the required heads.

F Experimental Details and Supplementary Results

F.1 Synthetic Distributions and Evaluation

The pair evaluations use $\eta \in \{0.05, 0.15, 0.25, 0.5, 0.8\}$ and $L \in \{2, 4, 8, 16, 32\}$. The four retention transitions (α_s, α_t) are $(1, 0)$, $(0.9, 0.6)$, $(0.75, 0.25)$, and $(0.55, 0.4)$. For each noise level, we generate $100{,}000$ independent pairs per seed by drawing the first bit uniformly and flipping it with probability $\eta/2$ to obtain the second. Twenty base seeds, from 4100 to 4119, are offset by $10{,}000\eta$ before sampling. Each cell of the empirical 2×2 table receives a pseudocount of $1/2$. We compute mutual information from the normalized table and scale it by $m(\alpha_s - \alpha_t)^2$.

The exact numerical comparison computes mutual information from the four probabilities of g_η and applies the same collision factor. The analytic coefficient is $\log 2 - h_2(\eta/2)$. The 100 transition settings share five sets of coefficient estimates. Table 1 reports the coefficient estimates.

For parity, we use $k \in \{2, 3, 4, 6, 8\}$ and $T \in \{2, 4, 8, 16, 32, 64, 128, 256\}$. All reported costs are divided by nc_λ, so the schedule comparison is independent of the positive scale factor. The nodes and objective values are evaluated from equation 13 and equation 14. These deterministic calculations have no seed uncertainty.

Table 1: Pair mutual information in nats. Estimates are means with sample standard deviations across 20 seeds, using 100,000 pairs per seed. The same estimates determine all length and transition comparisons at each noise level.

η	Exact c_η	Empirical estimate
0.05	0.576240	0.576001 ± 0.002123
0.15	0.426763	0.427462 ± 0.002665
0.25	0.316377	0.316182 ± 0.002089
0.5	0.130812	0.131053 ± 0.001397
0.8	0.020136	0.020225 ± 0.000666

F.2 Schedule Comparisons

The pair schedule comparison fixes $\eta = 0.25$, $m = 8$, and $T \in \{2, 4, 8, 16, 32, 64, 128\}$. Besides uniform increments, we evaluate $\Delta_t = 2t/[T(T+1)]$ and its reversal, and 12 draws from $\mathrm{Dirichlet}(1, \ldots, 1)$ for each T. A NumPy generator initialized with seed 6199 produces the random schedules. The two linear orders have identical cost because reversing the increments preserves their sum of squares. Figure 4(a) displays the costs divided by mc_η/T and the full range of the random draws.

For parity, substituting the stored optimal nodes into $\sum_j B_k(x_j, x_{j-1})$ gives the direct objective check. Uniform costs use $x_j = j/T$ in the same expression. At $T = 8$, the uniform and optimal values are 0.234375 and 0.170421 for $k = 4$, and 0.427650 and 0.190659 for $k = 8$. Figure 4(b) shows convergence of $TC_{k,T}$ to $2(k-1)/k$, as established in Appendix C.3. At $T = 256$, the largest absolute difference from the limiting constant is 0.013674 over the displayed block sizes.

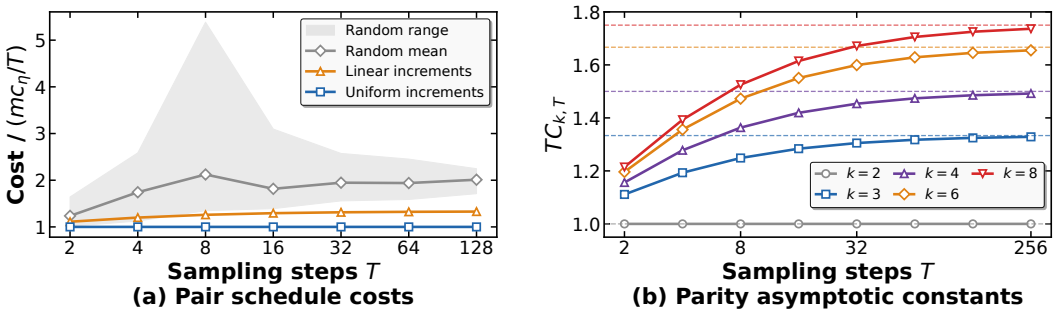

Figure 4: Supplementary schedule calculations. (a) Pair costs relative to the uniform optimum. The shaded range covers 12 random schedules for each step count and the gray curve is their mean. The two linear orders coincide. (b) Optimal parity costs multiplied by T, with dashed lines at their limiting constants. All costs are exact evaluations for the displayed schedules.

Table 2: Exact optimal parity costs $C_{k,T}$, normalized by nc_λ. These are deterministic recurrence evaluations. For $k = 2$, the cost equals $1/T$.

T	$k = 2$	$k = 3$	$k = 4$	$k = 6$	$k = 8$
2	0.500000	0.555556	0.578125	0.598122	0.607304
4	0.250000	0.298279	0.319443	0.338998	0.348240
8	0.125000	0.156079	0.170421	0.184072	0.190659
16	0.062500	0.080239	0.088698	0.096909	0.100927
32	0.031250	0.040784	0.045423	0.049981	0.052231
64	0.015625	0.020586	0.023029	0.025447	0.026647
128	0.007813	0.010348	0.011605	0.012855	0.013478
256	0.003906	0.005190	0.005828	0.006465	0.006783

F.3 Canonical Path and Output Calculations

The rate comparison fixes $\eta = 0.25$, $m = 8$, and the same step grid as the pair schedule comparison. We use $\rho = 1/T$ in equation 7, with the true fully masked initialization and exact marginal heads. The output values are calculated from binary relative entropy. Ordinary least squares on the logarithms of the four values at $T = 16, 32, 64, 128$ gives the displayed slope -1.959526.

For fixed $\eta \in (0, 1)$, write $a = \eta/2$ and $\delta_T = (1 - \eta)/(2T)$. Expanding the binary divergence at a gives

$$d_2(a\|a + \delta_T) = \frac{\delta_T^2}{2a(1-a)} + O(\delta_T^3), \qquad \frac{T^2}{m} D_{\mathrm{KL}}(p\|Q_0^*) \longrightarrow \frac{(1-\eta)^2}{2\eta(2-\eta)}.$$

At $\eta = 0.25$, this constant is $9/14$. Figure 2(c) compares the two exact divergences.

Table 3: Canonical sampler with $m = 8$, $\eta = 0.25$, and uniform increments. Both divergences are exact formula evaluations in nats. The last column converges to $9/14$.

T	Path KL	Output KL	$T^2\,\mathrm{KL}_{\mathrm{out}}/m$
2	1.26550808	0.77184367	0.38592183
4	0.63275404	0.23368501	0.46737002
8	0.31637702	0.06696471	0.53571767
16	0.15818851	0.01820667	0.58261333
32	0.07909425	0.00477129	0.61072559
64	0.03954713	0.00122312	0.62623524
128	0.01977356	0.00030977	0.63439955

A finite example checks the output calculation directly. For one pair with $\eta = 1/2$ and two uniform steps, enumerate the four outcomes in the order $00, 01, 10, 11$. The target and generated probabilities

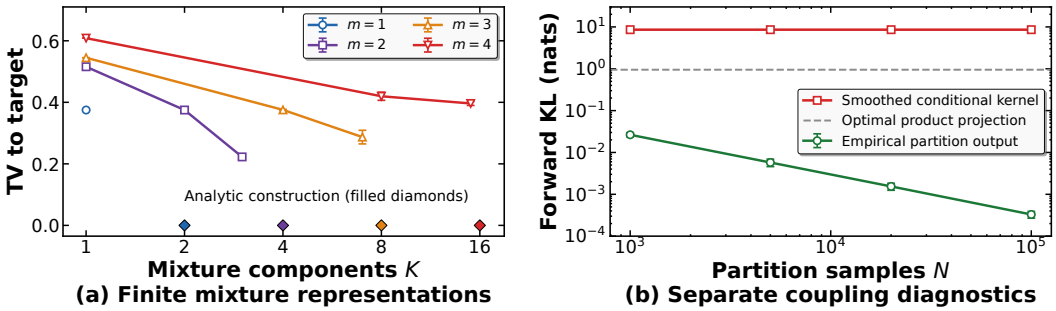

Figure 5: Representation diagnostics for pairs with $\eta = 0.25$. (a) Regularized EM fits with fewer than 2^m components, shown as mean TV with standard deviations over 12 restarts. Filled diamonds show the separate analytic construction at $K = 2^m$. (b) The empirical law of deterministic partition outputs and the conditional divergence of a separately smoothed kernel. Bars on the empirical output curve show standard deviations across 20 seeds. These curves evaluate distinct constructions.

are

$$p = (3/8, 1/8, 1/8, 3/8), \qquad Q_0^* = (5/16, 3/16, 3/16, 5/16).$$

The path KL is $c_{1/2}/2 = 0.065406$, the output KL is 0.035375, and the output TV is $1/8$. This finite calculation uses the two reveal time collision cases.

F.4 Representation Diagnostics

The mixture comparison fixes $\eta = 0.25$ and uses $m \in \{1, 2, 3, 4\}$. For each m, the component counts are $K = 1$, $\max(1, 2^{m-1})$, and $2^m - 1$, with duplicate values removed. Each configuration uses 12 initializations and 600 regularized EM iterations over the enumerated target law. We initialize all mixture weights at $1/K$ and draw each Bernoulli parameter independently from $\mathrm{Unif}(0.1, 0.9)$. The random seed is $b + 100m + K$, with $b = 7400, \ldots, 7411$.

Let $r_k(x)$ be the posterior responsibility of component k at state x under the current mixture. The updates use the exact target weights $p(x)$,

$$u_k = \sum_x p(x) r_k(x) + 10^{-12}, \qquad w_k = \frac{u_k}{\sum_{\ell=1}^{K} u_\ell}, \qquad \theta_{ki} = \frac{\sum_x p(x) r_k(x) x_i + 0.01}{u_k + 0.02}.$$

The 0.01 terms regularize both binary outcomes. When evaluating probabilities, we floor each Bernoulli outcome probability at 10^{-12} and each mixture weight at 10^{-300}, then normalize the resulting law over the enumerated states. All runs use the fixed iteration count.

Figure 5(a) reports the mean TV and sample standard deviation across restarts. The attaining construction at $K = 2^m$ is evaluated separately. Its largest numerical TV is 2.01×10^{-12}, while the smallest TV among the fitted configurations is 0.222578.

The coupling diagnostics use $m = 3$, $\eta = 0.25$, and a partition of a uniform variable into cells of masses $p(x)$. For $N \in \{1000, 5000, 20000, 100000\}$, we sample the deterministic partition outputs and estimate their distribution using a pseudocount of $1/2$ on each of the 64 states. The 20 base seeds range from 8500 to 8519 and are offset by N. At $N = 100{,}000$, forward KL to this empirical law has mean 0.000330 and standard deviation 0.000060.

A separate calculation replaces each coordinate of the partition output by a Bernoulli kernel with error probability 0.02 and evaluates conditional KL before integrating over the partition variable. Its value is 8.586925, compared with 0.949131 for the optimal product projection. Figure 5(b) labels these evaluations separately. The exact deterministic partition result and its infinite path divergence are given in Appendix B.7.

Table 4: All fitted mixture configurations at $\eta = 0.25$. TV is summarized by the mean and sample standard deviation over 12 regularized EM restarts. The minimum is the best observed restart and provides no certificate of global optimality.

m	K	Mean TV	Minimum TV
1	1	0.375000 ± 0.000000	0.375000
2	1	0.515625 ± 0.000000	0.515625
2	2	0.375000 ± 0.000000	0.375000
2	3	0.222578 ± 0.000000	0.222578
3	1	0.544922 ± 0.000000	0.544922
3	4	0.375462 ± 0.001813	0.374685
3	7	0.286958 ± 0.022141	0.259492
4	1	0.608643 ± 0.000000	0.608643
4	8	0.419710 ± 0.013182	0.401945
4	15	0.396312 ± 0.006407	0.384017

Table 5: Forward KL from the target to the empirical output law of the deterministic partition. Each of the 64 histogram cells receives a pseudocount of $1/2$. Values are means with standard deviations over 20 seeds and concern the empirical law.

N	Empirical output KL
1000	0.026457 ± 0.003201
5000	0.005757 ± 0.001163
20000	0.001543 ± 0.000294
100000	0.000330 ± 0.000060

F.5 Pretrained Denoisers and Evaluation

Tasks and models. Nominal length settings are $128, 256, 512, 1024$, dependency densities are $0.05, 0.1, 0.2, 0.4$, and replacement probabilities are $0, 0.1, 0.2, 0.3$. Length is a construction parameter and does not define a token count independently of the tokenizer.

Masking and decoding. The runner tokenizes each target without adding special tokens and masks positions that overlap marked dependency spans. Unmarked text remains visible. At each step, the model scores the selected positions in one model call and fills them with argmax predictions. The step budget is $T \in \{4, 8, 16, 32, 64\}$. MDLM receives a timestep determined by the round index and total budget.

Token NLL. For reference tokens $x^\star$ and positions D_t selected at step t, we define

$$\mathcal{L}^a_{\mathrm{tok}} = -\frac{1}{N_{\mathrm{scored}}} \sum_t \sum_{i \in D_t} \log q_{\theta,i}(x^\star_i \mid z^a_t, t), \qquad N_{\mathrm{scored}} = \sum_t |D_t|, \quad a \in \{\mathrm{gen}, \mathrm{TF}\}. \tag{25}$$

Here, $q_{\theta,i}$ is the softmax of the model logits at position i. Each selected occurrence contributes once to the negative sum, and the denominator counts those occurrences. The measure uses natural logarithms, in nats per scored token. The state z^{gen}_t contains previously generated tokens, while z^{TF}_t contains previously revealed reference tokens. We use the former for infilling comparisons and the latter for supplementary diagnostics.

Reveal policies and matching. Random balanced selection samples positions without replacement within each round. Confidence selection ranks remaining positions by their maximum predicted token probability. Dependency selection uses supplied span roles. The role policy uses precedence based on the reference dependency structure, while the conflict policy avoids revealing linked positions together while preserving round quotas. The positive comparisons use the role policy for MDLM and the available role and conflict cohorts for Qwen3. Pairwise differences are averaged within each of the eight seeds before computing the reported normal 95% interval.

Positive diagnostics. With reference tokens supplied as context, forced simultaneous revelation increases token NLL by 1.899 ± 0.049 nats per scored token over $25{,}459$ matched pairs. Separating linked spans reduces the collision rate within a round by 0.488 ± 0.006 relative to random balanced

selection over 27,102 matched pairs. During generation, the dependency policy reduces token NLL by 0.383 ± 0.020 relative to confidence selection over 22,495 matched pairs. Figure 3 shows the corresponding model and step comparisons.

F.6 Schedules Estimated from Finite Samples

The comparison estimates the full law on six binary coordinates from N independent draws, adding a pseudocount of $1/2$ to each of the 64 states. Thus the estimated probability of state x is $(n_x + 1/2)/(N + 32)$, where n_x is its observed count. We use sample sizes $N = 256$, 1024, and 4096, with 24 replicates for each target and sample size. Each replicate uses the same estimated law for all methods that select a schedule from samples. Conditional entropy tables of this law define transition costs through equation 3. Dynamic programming selects eight strict transitions on the retention grid $\{0, 1/16, \ldots, 1\}$. All selected schedules are evaluated using the true law on that grid.

The three target laws are pairs with $\eta = 0.35$, an Ising law with overlapping interactions, and a law with additional dependence across distant coordinates. Writing $s_i = 2x_i - 1$ and $s_7 = s_1$, the latter two are

$$p_{\mathrm{Ising}}(x) \propto \exp\left(0.52 \sum_{i=1}^{6} s_i s_{i+1} + 0.20(s_1 s_4 + s_3 s_6 + s_5 s_2)\right),$$

$$p_{\mathrm{mixed}}(x) \propto \exp\left(0.28 \sum_{i=1}^{6} s_i s_{i+1} + 0.78 s_1 s_2 s_3 s_4 + 0.65 s_1 s_3 s_5 s_6\right).$$

We compare dynamic programming on the estimated law with uniform increments, a power schedule selected using estimated costs, and dynamic programming on the product of its marginals over $\{1, 2\}$, $\{3, 4\}$, and $\{5, 6\}$. For the pair, Ising, and mixed targets, respectively, set $d = 1, 2, 3$. Replicate $r = 0, \ldots, 23$ at sample size N uses a NumPy generator with seed $709100 + 100000d + 1000N + r$.

The power schedules have retention levels $\alpha_t = (1 - t/8)^\gamma$ for $t = 0, \ldots, 8$. The candidate exponents are 0.55, 0.70, 0.85, 1, 1.2, 1.45, and 1.8. We round each level to its nearest grid point, using the smaller grid index on ties, and exclude schedules that lose strictness. The retained schedule minimizes cumulative cost under the same estimated full law. Reported regret is the selected schedule's true cumulative cost minus the minimum true cost over the same grid and eight transitions.

To specify the optimization, let $g_j = j/M$ with $M = 16$, and let $\widehat{\Phi}(g_i, g_j)$ be the transition cost computed from the estimated law for $i > j$. Define $V_k(j)$ as the least estimated cost of reaching g_j from $g_M = 1$ in exactly k strict transitions. The recurrence is

$$V_0(j) = \begin{cases} 0, & j = M, \\ +\infty, & j < M, \end{cases} \qquad V_k(j) = \min_{i>j} \{V_{k-1}(i) + \widehat{\Phi}(g_i, g_j)\}. \tag{26}$$

Stored predecessors recover the minimizing schedule from $V_8(0)$. Conditioning on the predecessor of each final transition proves the recurrence by induction. Given transition costs, optimization takes $O(TM^2)$ operations and constructing them from the entropy table takes $O(M^2 3^L)$ branch terms. We use this enumeration at $L = 6$, with the true law reserved for evaluation and the grid oracle.