# WHAT DOES A DISCRETE DIFFUSION MODEL LEARN? ONE REVERSE PROCESS IN THREE COORDINATES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

What does a discrete diffusion model actually learn? The literature trains denoisers, bridge plug-ins, and concrete scores, each deriving its own ELBO objective under different conventions. This fragmentation makes the object being learned unclear and ELBO values difficult to interpret and compare. All three methods learn the same object in different coordinates: the exact time-reversal of the noising process. We first prove the Oracle Distance identity: for any noising process, the negative ELBO is not just a bound on the likelihood; it equals the data entropy plus the path KL divergence from the reversed noising process to the learned one. Next, for token-factorizing noise, we show that denoiser, bridge plug-in, and concrete score are three coordinates of this reverse process, derive exact conversions between them, and identify the probability distribution learned by each. In particular, contrary to its standard interpretation, the bridge plug-in learns not a denoiser but the cavity distribution: the clean token given the noisy sequence excluding its own position. We confirm the theory by training all three coordinates with masked, uniform, and GIDD kernels on OpenWebText under a shared recipe, and by training uniform-diffusion solvers on four logic tasks. The learned coordinates agree after conversion, while confusing the cavity with the denoiser can severely degrade sampling quality. Together with exact boundary-term formulas, closed-form NELBO calibrations at initialization, and a clock interpretation of time schedules, these results provide a conceptual framework and a mechanical recipe for deriving, implementing, and comparing discrete diffusion processes.[1]

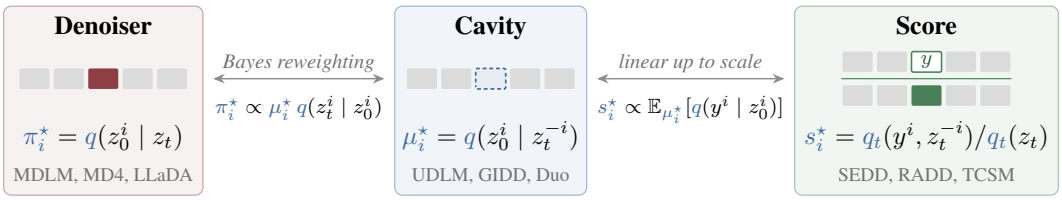


Figure 1: The true reverse rate in three coordinates, each learned at the ELBO optimum (§4). The denoiser keeps position $i$ of $z_t$, the cavity leaves it out, and the score's numerator swaps in $y^i$.

## 1 INTRODUCTION

Discrete diffusion provides a flexible route to model categorical data: slowly corrupt each position of a sequence, then learn to run the corruption backward (Hoogeboom et al., 2021; Austin et al., 2021). This unlocks a new test-time axis on which to scale computation, and any-order generation enables token-parallel decoding, self-correction, and constrained generation for language and reasoning tasks (Sahoo et al., 2024; Schiff et al., 2025; Gong et al., 2025), capabilities that have been tested at the scale of large language models (Nie et al., 2025; Ye et al., 2025). Mathematically, the corruption and denoising processes are modeled as continuous-time Markov chains (CTMCs) (Campbell et al., 2022): jump processes determined infinitesimally by their jump rates, that is, the rates of instantaneous change between states. Compared to discrete-time formulations, CTMCs are mathematically simpler and can be sampled using arbitrary time grids that are not fixed in advance.

[1]Code will be released upon publication.



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.

Discrete diffusion modeling therefore requires mapping the model's output to the denoising process through its jump rates. The literature does so in one of three seemingly different ways: the model can be a denoiser, a bridge plug-in, or a concrete score (Sahoo et al., 2024; Schiff et al., 2025; Lou et al., 2024). Each method then derives its own likelihood bound (ELBO) objective and sampling formulas, following different conventions about the boundary terms and interpreting the model differently. This is error-prone and makes the ELBO hard to interpret and compare. A better ELBO value may reflect a better generative model or a tighter variational bound, and the standard convention of dropping the boundary terms can silently invalidate ELBO comparisons. In short, a central question remains open:

*What does a discrete diffusion model actually learn?*

We answer this question in full generality: a discrete diffusion model learns the exact reversal of the noising process. For token-factorizing noise, each of the three parameterizations of the literature is a different coordinate of this same process. Our contributions are as follows:

**(i) The Oracle Distance.** For any noising process, the negative ELBO (NELBO) equals the data entropy plus the KL divergence from the oracle reverse process, the exact time-reversal of the noising process, to the learned one (Theorem 1). The NELBO is therefore a KL between two distributions over *entire trajectories*, a "path" KL. The unique jump rates that minimize it among rates depending only on the noisy sequence, not on the clean data, are a projection: the clean-conditioned reverse jump rates projected onto the noisy information, analogous to flow matching in continuous diffusion.

**(ii) Three coordinates of the reverse process.** For token-factorizing noise, the reverse process has three coordinates, denoiser, plug-in, and score, with exact conversions between them (Theorem 2). Combined, they recover the ELBO and sampling formulas of the literature and identify the distribution each method learns. Contrary to its standard description, the bridge plug-in does not train a denoiser but the *cavity* distribution: the clean-token distribution given the noisy sequence excluding its own position.

**(iii) Empirical validation.** We verify the theory by training all three coordinates with masked, uniform, and GIDD (generalized interpolating discrete diffusion; von Rütte et al., 2025) kernels on OpenWebText under one shared recipe, and by training uniform-diffusion solvers on four logic tasks (§5): the three coordinates learn the same object, ELBO training curves start at the predicted value, and confusing the cavity with the denoiser can severely degrade sampling quality.

Together with formulas for the boundary terms and an interpretation of time schedules and training weights as a curriculum over noise levels, these results give a conceptual framework and a mechanical recipe for deriving and comparing discrete diffusion processes.

## 2 Continuous-time discrete diffusion

A diffusion model defines a forward noising process $(Z_t)_{t\in[0,T]}$ on a finite space $\mathcal{X}$ whose marginals $q_t$, the distribution of $Z_t$, interpolate between the "clean"-data distribution $q_0 = p_{\text{data}}$ and a terminal distribution $q_T$ from which sampling is easy, and then learns to run the process backward.[2] In discrete diffusion the process corrupts data directly on its alphabet. This is described by a continuous-time Markov chain (CTMC), which is determined by its infinitesimal generator or jump rate $Q_t$, such that

$$\mathbb{P}(Z_{t+h} = y \mid Z_t = x) = \mathbf{1}\{x = y\} + h\, Q_t(x, y) + o(h), \qquad Q_t(x, x) = -\sum_{y \neq x} Q_t(x, y),$$

with $Q_t(x, y) \geq 0$ for $y \neq x$. Continuous time removes the need for a predefined sampling schedule, lets training times be drawn continuously, and admits a simpler formulation. We write $q_{t|s}(y \mid x) = \mathbb{P}(Z_t = y \mid Z_s = x)$ for the forward kernel, and similarly for other conditional distributions such as the reverse bridge $q_{s|t,0}$. Whenever $q_t(x) > 0$, the time reversal of a CTMC is again a CTMC. Its rate, which we call the *reverse* rate, and its clean-conditioned version obtained by conditioning on $Z_0 = z_0$, are

$$\widehat{Q}_t(z_t, y) = Q_t(y, z_t)\,\frac{q_t(y)}{q_t(z_t)}, \qquad \widehat{Q}_t(z_t, y \mid z_0) = Q_t(y, z_t)\,\frac{q_{t|0}(y \mid z_0)}{q_{t|0}(z_t \mid z_0)}, \qquad y \neq z_t. \quad (1)$$

[2] Throughout, we color quantities related to the true process in blue and those to the learned model in red.

The rate $\widehat{Q}_t(z_t, y)$ is the target of a denoising model, which has no access to clean data at generation time, while the clean-conditioned rate is the tractable training target. Implementations typically truncate the interval to $[t_1, t_2] \subset (0, T)$, since the clean-conditioned ratios are singular or high-variance near the endpoints. A denoising model thus consists of model reverse rates $\widehat{Q}_t^\theta$ matching $\widehat{Q}_t$, a reconstruction decoder $p^\theta_{0|t_1}$ aiming at $q_{0|t_1}$, and a terminal prior $p^\theta_{t_2}$ from which generation starts. Likelihood-based training then supervises the model rates via the continuous-time ELBO (Campbell et al., 2022; Benton et al., 2024), which relies on the local *rate divergence*:

$$\Phi(a, b) := a \log \frac{a}{b} - a + b, \qquad a, b \geq 0, \tag{2}$$

with $\Phi(0, b) = b$ and $\Phi(a, 0) = +\infty$ for $a > 0$: the Bregman divergence of $F(u) = u \log u - u$ and the KL rate between two Poisson jump intensities.

**Proposition 1** (Finite-window CTMC ELBO). *Let $[t_1, t_2] \subset (0, T)$ and assume the conditions of Theorem B.1. Then $\log p_0^\theta(z_0) \geq \mathrm{ELBO}_{[t_1,t_2]}(\theta; z_0)$, where*

$$\begin{aligned} \mathrm{ELBO}_{[t_1,t_2]}(\theta; z_0) = \mathbb{E}_{q_{t_1|0}(z_{t_1}|z_0)}\big[\log p^\theta_{0|t_1}(z_0 \mid z_{t_1})\big] - \mathrm{KL}\big(q_{t_2|0}(\cdot \mid z_0) \,\big\|\, p^\theta_{t_2}\big) \\ - \int_{t_1}^{t_2} \mathbb{E}_{q_{t|0}(z_t|z_0)}\Big[\textstyle\sum_{y \neq z_t} \Phi\big(\widehat{Q}_t(z_t, y \mid z_0), \widehat{Q}_t^\theta(z_t, y)\big)\Big]\, dt. \end{aligned} \tag{3}$$

App. B gives the Monte Carlo estimator of the path integral, and App. E the boundary terms of each kernel. Models are trained by gradient descent on the negative ELBO, $\mathrm{NELBO}_{[t_1,t_2]}(\theta; z_0)$, and we write $\mathrm{ELBO}_{[t_1,t_2]}(\theta) := \mathbb{E}_{q_0}[\mathrm{ELBO}_{[t_1,t_2]}(\theta; z_0)]$ and $\mathrm{NELBO}_{[t_1,t_2]}(\theta)$ for the data averages.

**Takeaway 1.** The ELBO of any noising process follows a common recipe: (i) choose the corruption kernel $q_{t|s}$, (ii) differentiate it to obtain the forward rates $Q_t$, (iii) form the clean-conditioned reverse rates $\widehat{Q}_t(\cdot \mid z_0)$ as the quotient in (1), (iv) map the model's output to the reverse rates $\widehat{Q}_t^\theta$, (v) sum $\Phi(\widehat{Q}_t(\cdot \mid z_0), \widehat{Q}_t^\theta)$ over the possible jumps $z_t \to y$, (vi) integrate in $t$, $q_{t|0}$, (vii) add the boundary terms.

The boundary terms at $t_1$ and $t_2$ are usually discarded as model-independent constants, even though only the prior term is independent of the model, and dropping either hides the identities of §3 and makes reported ELBOs incomparable across processes. In the standard masked and uniform setups and the usual windows, both boundary terms are small (Table 10, App. G.3), but nothing guarantees this in general: endpoint divergences can concentrate much of the objective in the clipped region.

## 3 The Oracle Distance Identity

Averaging (3) over $q_0$ gives $\mathrm{KL}\big(q_0 \,\big\|\, p_0^\theta\big) \leq \mathrm{NELBO}_{[t_1,t_2]}(\theta) - H(q_0)$, with $H(q_0)$ the data entropy, further justifying NELBO minimization. However, a likelihood bound is hard to interpret and leaves open whether a better value reflects a better generative distribution or a tighter bound, which matters where direct sample metrics are hard to define, such as for text and music. A more satisfying interpretation requires looking not at marginals or jump rates but at the *path distribution*, that is, the distribution over entire trajectories. The two path distributions involved are the noising distribution $P^\star_{[0,T]}$, read in reverse time[3] as started at $q_T$ and run backward with the marginal reverse rates $\widehat{Q}_t$, and the model distribution $P^\theta_{[0,T]}$, started at the terminal prior and run backward with $\widehat{Q}_t^\theta$. Throughout this section we assume the regularity of Assumption C.1 and the conditions of Theorem C.1, with its two boundary mismatch terms vanishing as $t_1 \downarrow 0$ and $t_2 \uparrow T$.

**Theorem 1** (Oracle Distance). *For any noising process and model,*

$$\boxed{\mathrm{NELBO}_{[0,T]}(\theta) - H(q_0) = \mathrm{KL}\Big(P^\star_{[0,T]} \,\Big\|\, P^\theta_{[0,T]}\Big) = \int_0^T \mathbb{E}_{q_t}\left[\sum_{y \neq z_t} \Phi\big(\widehat{Q}_t(z_t, y), \widehat{Q}_t^\theta(z_t, y)\big)\right] dt.} \tag{4}$$

On a clipped time window the identity holds with the corresponding boundary mismatch terms, similar to those of Proposition 1. Theorem C.1 gives the full statement and proves it in two alternative ways. Since $\Phi$ is a Bregman divergence, the classical Pythagorean identity (Lemma C.1) gives the unique minimizer of the NELBO over $Z_t$-measurable rates as a projection:

[3] A CTMC path distribution is equivalently characterized by reading it in either direction, that is, either from its starting distribution and forward rates, or its terminal distribution and backward rates; see App. A.

**Corollary 1** (Reverse-rate projection)**.** *Among rates measurable in the noisy state $Z_t$ alone,* $\mathrm{NELBO}_{[t_1,t_2]}(\theta)$ *is minimized, uniquely up to null sets, by*

$$\widehat{Q}_t^{\theta_\star}(z_t, y) = \mathbb{E}\Big[\widehat{Q}_t(Z_t, y \mid Z_0) \;\Big|\; Z_t = z_t\Big] = \widehat{Q}_t(z_t, y), \qquad y \neq z_t. \tag{5}$$

Thus, training on the clean-conditioned rates without access to the clean data amounts to a projection onto the noisy information available at time $t$, analogous to flow matching in continuous diffusion.

Denote by $I(Z_0; Z_t) = H(Z_0) - H(Z_0 \mid Z_t)$ the mutual information between clean and noisy data, that is, the information about $Z_0$ that survives in $Z_t$, with $H(Z_0 \mid Z_t)$ the conditional entropy. Even at the oracle, the path term of the ELBO does not vanish: it equals the information the forward process destroys, in a CTMC analogue of the I-MMSE relation (Guo et al., 2005):

**Proposition 2** (Irreducible cost is information loss)**.** *The oracle integrand satisfies*

$$\mathcal{J}_t^\star = \frac{d}{dt} H(Z_0 \mid Z_t) = -\frac{d}{dt} I(Z_0; Z_t) \;\geq\; 0, \qquad \int_{t_1}^{t_2} \mathcal{J}_t^\star \, dt = I(Z_0; Z_{t_1}) - I(Z_0; Z_{t_2}), \tag{6}$$

*where* $\mathcal{J}_t^\star := \mathbb{E}_{Z_0, Z_t}\big[\sum_{y \neq Z_t} \Phi(\widehat{Q}_t(Z_t, y \mid Z_0), \widehat{Q}_t(Z_t, y))\big]$.

The Pythagorean identity for Bregman divergences then extends this result to arbitrary model rates:

$$\mathbb{E}_{Z_0, Z_t}\left[\sum_{y \neq Z_t} \Phi\big(\widehat{Q}_t(Z_t, y \mid Z_0), \widehat{Q}_t^\theta(Z_t, y)\big)\right] = \mathcal{J}_t^\star + \mathbb{E}_{Z_t}\left[\sum_{y \neq Z_t} \Phi\big(\widehat{Q}_t(Z_t, y), \widehat{Q}_t^\theta(Z_t, y)\big)\right]. \tag{7}$$

Integrated over time, $\mathcal{J}_t^\star$ telescopes into conditional entropies by (6) and the mismatch part accumulates into the path KL, thus proving Theorem 1.

Jensen's and Gibbs' inequalities give $\mathrm{NELBO}_{[0,T]}(\theta) \geq \mathbb{E}_{q_0}[-\log p_0^\theta(z_0)] \geq H(q_0)$, and this floor is always attainable at the oracle reverse rates:

**Corollary 2** (Universal attainable NELBO floor)**.** *Suppose some $\theta_\star$ realizes $\widehat{Q}_t^{\theta_\star} = \widehat{Q}_t$ on $(0, T)$. Then $p_0^{\theta_\star} = q_0$, the ELBO is tight:* $\log p_0^{\theta_\star}(z_0) = \mathrm{ELBO}_{[0,T]}(\theta_\star; z_0)$*, and*

$$\min_\theta \mathrm{NELBO}_{[0,T]}(\theta) = H(q_0). \tag{8}$$

That is, every noising process shares the same attainable NELBO floor given enough capacity: the data entropy. As a consequence, the observed gap between masked and uniform diffusion NELBOs is not intrinsic but must instead come from modeling capacity and trainability.

**Takeaway 2.** The NELBO equals the data entropy plus the path KL from the oracle reverse process to the learned one, and its unique minimizer is the projection of the clean-conditioned reverse rate onto the information present in the noisy state. This explains the looseness of the ELBO: it measures the KL divergence from one specific distribution over reverse trajectories, that of the oracle reverse process, and in doing so penalizes deviations even among trajectories whose clean endpoint also matches the data distribution.

**Time schedules.** Importance-sampling the training times from a density $p_{\mathrm{IS}}(t)$ amounts to running the process at a different speed, going slower where $p_{\mathrm{IS}}$ is high, and the ELBO is the same integral at any speed, by (B.10) in App. B.3. Thus, importance sampling and time schedules leave the ELBO unchanged and only set the resolution of training, that is, how densely each noise level is sampled. Training weights other than the ELBO's do change the objective (Kingma & Gao, 2023; Shi & Titsias, 2025), but the reverse-rate projection of Corollary 1 is solved separately at each time, so the reweighted objective keeps the NELBO minimizer. Schedules, importance densities, and training weights can thus all be seen as a choice of curriculum over noise levels, and one would focus it where the learning signal is highest (Hong et al., 2026). Generation benefits from different properties of the time schedule, hence the common use of different schedules for training and sampling (linear against cosine, for instance).

**Takeaway 3.** Importance sampling and time schedules change the speed of the process but not the ELBO; other training weights change the objective but not its per-time optimum. Thus, all of them can be seen as a choice of curriculum over noise levels.

## 4 THREE COORDINATES FOR TOKEN-FACTORIZING DISCRETE DIFFUSION

Sequential data are elements of a product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_L$. The standard noising processes corrupt positions independently, that is, the forward noising kernel factors over positions: $q_{t|s}(y \mid x) = \prod_i q^i_{t|s}(y^i \mid x^i)$, where we write $x^i$ for the $i$-th token of a sequence $x$ and $x^{-i}$ for the rest of the sequence excluding $x^i$. The jump rates then decompose as a sum of per-position rates, $Q_t = \sum_i (Q_t)_i$, each position running an independent process with rate $R^i_t$ on $\mathcal{X}_i$, so that $(Q_t)_i(x, y) = \mathbf{1}\{x^{-i} = y^{-i}\}\, R^i_t(x^i, y^i)$ (Theorem D.1). Nothing forces the per-position rates $R^i_t$ to be equal, though they usually are. The standard kernels corrupt each token with probability $\beta_t$, rising from 0 to 1 over $[0, T]$, and replace it by a mask state $\mathfrak{m}$ (masked diffusion), by a token drawn uniformly from the vocabulary of size $V$ (uniform diffusion), or by a draw from their mixture with weight $\lambda$ (GIDD; App. E).

Evaluating the NELBO minimizer from (5) on a one-token jump $z_t \to y$ where position $i$ jumps $z^i_t \to y^i$, $z^i_t \neq y^i$, while the rest remains unchanged, $y^{-i} = z_t^{-i}$, gives per-position reverse rates

$$(\widehat{Q}^{\theta\star}_t)_i(z_t, y) = (\widehat{Q}_t)_i(z_t, y) = R^i_t(y^i, z^i_t)\, \mathbb{E}_{q(z^i_0 \mid z_t)}\left[\frac{q^i_{t|0}(y^i \mid z^i_0)}{q^i_{t|0}(z^i_t \mid z^i_0)}\right]. \tag{9}$$

Importantly, the clean data appears only through the $i$-th token marginal of the posterior distribution, $q(z^i_0 \mid z_t)$. This is why token-level heads suffice, and why each coordinate below fully specifies the reverse CTMC: substituting it fixes the rates, which determine the process by Kolmogorov uniqueness. The posterior average in (9) can be expressed through the following three quantities:

$$\pi^\star_i(z^i_0 \mid z_t) := q(z^i_0 \mid z_t), \quad \mu^\star_i(z^i_0 \mid z_t^{-i}) := q(z^i_0 \mid z_t^{-i}), \quad s^\star_i(z^i_t, y^i \mid z_t^{-i}) := \frac{q_t(y^i, z_t^{-i})}{q_t(z_t)}, \tag{10}$$

which we respectively call: *denoiser*, the usual clean-token predictor given the noisy sequence; *cavity*, the distribution of the clean token given the noisy context but excluding its own noisy position; and *score*, a ratio of time-$t$ marginals, where $(y^i, z_t^{-i})$ is $z_t$ with its $i$-th token replaced by $y^i$. The denoiser is the clean-token predictor of the masked diffusion line of work (Campbell et al., 2022; MDLM, Sahoo et al., 2024; MD4, Shi et al., 2024; LLaDA, Nie et al., 2025), the cavity is the distribution that the bridge plug-in optimizes in uniform and GIDD diffusion (UDLM, Schiff et al., 2025; GIDD, von Rütte et al., 2025; Duo, Sahoo et al., 2025), and the score is the model target of score methods (Meng et al., 2022; SEDD, Lou et al., 2024; RADD, Ou et al., 2025; TCSM, Zhang et al., 2025) (Figure 1). Crucially, prior work describes the bridge plug-in as learning a denoiser, but it is in fact learning a cavity distribution. In short:

**Theorem 2** (Three coordinates)**.** *For a token-factorizing noising process, the NELBO is uniquely minimized by the reverse rate of Corollary 1, and on a one-token jump $z_t \to y = (y^i, z_t^{-i})$ the posterior average in* (9) *has the three equal forms*

$$\boxed{\mathbb{E}_{\pi^\star_i}\left[\frac{q^i_{t|0}(y^i \mid z^i_0)}{q^i_{t|0}(z^i_t \mid z^i_0)}\right] = \frac{\mathbb{E}_{\mu^\star_i}\left[q^i_{t|0}(y^i \mid z^i_0)\right]}{\mathbb{E}_{\mu^\star_i}\left[q^i_{t|0}(z^i_t \mid z^i_0)\right]} = s^\star_i(z^i_t, y^i \mid z_t^{-i}).} \tag{11}$$

Substituting a network head $\pi^\theta_i$, $\mu^\theta_i$, or $s^\theta_i$ for its distribution in (11) yields three head-to-rate maps which, with Proposition 1, give the three standard objectives: denoiser, bridge plug-in, and concrete score, each optimized at its oracle distribution from (10), uniquely up to the identifiability of the map. Networks are trained with access to the full sequence $z_t$, but we keep the conditioning notation of the distribution each head is trained toward, e.g. $z_t^{-i}$ for the cavity, and test empirically in §5.2 whether the network learns to ignore $z^i_t$.

**Takeaway 4.** Denoiser, cavity, and score are three coordinates of the same object, the oracle reverse rate of any token-factorizing noising process. They subsume the published CTMC training objectives, each with an exact distribution that minimizes the NELBO, unique up to the identifiability of the head-to-rate map. In particular the bridge plug-in trains a cavity distribution rather than a denoiser.

**Proposition 3** (Coordinate conversions). *Fix $(i, z_t, t)$ and write the head-induced distribution of the noisy token: $m_i^\theta(y \mid z_t^{-i}) := \sum_{z_0^i} \mu_i^\theta(z_0^i \mid z_t^{-i})\, q_{t|0}^i(y \mid z_0^i)$. Then*

$$\pi_i^\theta(z_0^i \mid z_t) \propto \mu_i^\theta(z_0^i \mid z_t^{-i})\, q_{t|0}^i(z_t^i \mid z_0^i); \qquad \mu_i^\theta(z_0^i \mid z_t^{-i}) \propto \frac{\pi_i^\theta(z_0^i \mid z_t)}{q_{t|0}^i(z_t^i \mid z_0^i)}; \tag{12}$$

$$s_i^\theta(z_t^i, y^i \mid z_t^{-i}) = \frac{m_i^\theta(y^i \mid z_t^{-i})}{m_i^\theta(z_t^i \mid z_t^{-i})}; \qquad m_i^\theta = \mu_i^\theta\, q_{t|0}^i, \text{ with } m_i^\theta(\cdot \mid z_t^{-i}) \propto s_i^\theta(z_t^i, \cdot \mid z_t^{-i}). \tag{13}$$

The last linear system, required to go from score to cavity, has a solution if and only if $m_i^\theta$ lies in the image of $q_{t|0}^i$ arranged as a row-stochastic matrix and a unique one if and only if $q_{t|0}^i$ has full row rank, while the solution $\mu_i^\theta$ is a probability vector if and only if it is nonnegative.

The cavity-to-denoiser conversion in (12) is a *Bayes reweighting*: the forward kernel at the observed token, which we call the *Bayes factor*, is the evidence the local observation adds on top of the context. It thus splits the information that $z_t$ carries about $z_0^i$ into two parts: what the context $z_t^{-i}$ reveals, the cavity, which must be learned, and what $z_t^i$ adds, given analytically by the kernel. If however the noisy token carries no information about the original token, the distinction disappears (see App. E for further details on masked, uniform, and GIDD diffusion):

**Corollary 3** (Denoiser equals cavity in masked diffusion). *If the Bayes factor $q_{t|0}^i(z_t^i \mid z_0^i)$ is constant in $z_0^i$ on its support, then the Bayes reweighting is restriction to that support followed by renormalization. In masked diffusion, $q_{t|0}^i(\mathtt{m} \mid z_0^i) = \beta_t$ at a masked position for every $z_0^i$, so $\pi_i^\star(\cdot \mid z_t) = \mu_i^\star(\cdot \mid z_t^{-i})$.*

**Ancestral sampling.** The finite-step sampling formulas share the same token-factorized ansatz. In the denoiser coordinate, substituting $\pi_i^\theta$ into the exact backward jump distribution from $z_t$ to $z_s$, $s < t$, gives the standard ancestral formula

$$p_{s|t}^\theta(z_s \mid z_t) = \prod_{i=1}^{L} p^\theta(z_s^i \mid z_t), \quad p^\theta(Z_s^i{=}y \mid z_t) = q_{t|s}^i(z_t^i \mid y) \sum_{z_0^i} \pi_i^\theta(z_0^i \mid z_t)\, \frac{q_{s|0}^i(y \mid z_0^i)}{q_{t|0}^i(z_t^i \mid z_0^i)}. \tag{14}$$

Substituting the cavity-to-denoiser conversion from (12) gives the usual bridge plug-in sampler (App. D).

At the oracle, each factor of (14) is the exact marginal $q(z_s^i \mid z_t)$, but the positional marginals give no access to the joint $q(z_s \mid z_t)$, and their product ignores the correlations between sequence positions. This is the error committed during discrete-time ancestral sampling: a factorization error, from sampling the marginals independently, present even at the ELBO oracle. Equivalently, the product of marginals is exact for the CTMC process in which, during the step from $t$ to $s$, each position's denoiser keeps conditioning on the other positions as they were at $t$, so the error comes from not updating the denoiser within the step. This is why masked diffusion, whose denoiser changes only when a token is unmasked, incurs no such error when unmasking tokens one at a time.

We conclude the section with useful NELBO calibration identities for uninformative heads:

**Proposition 4** (Calibration at initialization). *Fix $[t_1, t_2] \subset (0, T)$ and the uniform-forward boundary distributions of App. D. Then, per token,*

- *a cavity head $\mu_i^\theta \equiv 1/V$ has $\mathrm{NELBO}_{[t_1,t_2]} = \log V$, for every token-factorizing process (and a score head at zero logits under the folded parameterization of App. E gives the same value);*
- *a denoiser head $\pi_i^\theta \equiv 1/V$ under uniform diffusion has $\mathrm{NELBO}_{[t_1,t_2]} = \frac{V-1}{V} \log \frac{1}{\beta_{t_1}} + O(1)$.*

These identities are a useful test for NELBO implementations, because they simultaneously check boundary terms, importance sampling schedules, and the head-to-rate map.

**Takeaway 5.** Denoiser and cavity coincide where the Bayes factor is constant, as in masked diffusion. Under general noise, the denoiser must also learn the analytic Bayes factor, and its NELBO diverges as $\log(1/\beta_{t_1})$ at initialization, delaying its training (§5). Cavity and score instead have a per-token NELBO of exactly $\log V$ at initialization with both boundary terms, providing an implementation check.

# 5 EXPERIMENTS

For our text experiments, we train a grid of {masked, GIDD $\lambda$=0.1, GIDD $\lambda$=0.2, uniform} $\times$ {denoiser, cavity, score} on OpenWebText under a shared recipe: a time-conditioned diffusion transformer (Sahoo et al., 2024) at GPT-2 Small size trained on the path term of the NELBO of Proposition 1 (full experimental details in App. G). Each configuration is trained with a single seed, which suffices for our conclusions. Generation uses the ancestral sampler of (14) at 256 network function evaluations (NFEs) unless otherwise stated. Since sampling first converts to the denoiser, temperature is applied after the conversion, so that it tempers the contextual evidence (the cavity) and the local evidence (the Bayes factor) alike. We use GPT-2 Large to score samples via generative perplexity (GenPPL) and MAUVE (Pillutla et al., 2021); all reported NELBOs include both boundary terms.

## 5.1 READING HEADS IN THE WRONG COORDINATE

To assess the effect of omitting the conversion, we read each of the cavity and denoiser heads as the other, which drops the Bayes reweighting of Proposition 3 or applies it twice. The Bayes reweighting biases the distribution toward the observed noisy token, and omitting it when reading a cavity as a denoiser raises the NELBO by about 7 bits/token and collapses sample quality on every non-masked kernel (Table 1), while applying it twice by reading a denoiser as a cavity is less severe (Table 12, App. G.4). Along the whole quality–entropy frontier, every misread head is dominated by its native counterpart (Figure 9, App. G.4).

Table 1: Each cavity head on OpenWebText, natively and read as a denoiser (NELBO in bits/token).

| | Masked | | GIDD $\lambda$=0.1 | | GIDD $\lambda$=0.2 | | Uniform | |
|---|---|---|---|---|---|---|---|---|
| | Native | Misread | Native | Misread | Native | Misread | Native | Misread |
| NELBO | 4.900 | =native | 4.990 | 12.416 | 5.017 | 12.289 | 5.070 | 11.788 |
| MAUVE | 0.869 | =native | 0.884 | 0.008 | 0.896 | 0.007 | 0.873 | 0.011 |

The distinction is also relevant on four logic tasks, Sudoku, Boxoban, maze solving, and graph coloring, on which we train uniform-diffusion cavity models with three seeds each (App. G.2): misread as denoisers, that is, without the Bayes factor, their test NELBO rises on every task (Table 2). The increase is smaller than on text because here the context alone nearly determines each token, making the Bayes factor less important.

Table 2: Test NELBO ($10^{-3}$ bits/token) of the cavity models, mean (standard error) over seeds.

| | Sudoku | Boxoban | Maze | Graph coloring |
|---|---|---|---|---|
| Native | 23.3 (1.1) | 32.9 (0.3) | 18.5 (0.2) | 123.7 (0.1) |
| Misread | 24.4 (1.2) | 42.2 (0.4) | 20.8 (0.2) | 153.2 (0.1) |

Solve rate, however, is not what the ELBO measures, and there the effect of the conversion depends on both the sampler and the task (Figure 2). The Bayes factor commits each position to its current token. The *re-corrupt* sampler, which at every step predicts a fully clean sequence and noises it back to the next noise level (App. G.1), otherwise redraws every position from its context alone: without the factor, its solve rates collapse on graph coloring and Boxoban, where color permutations and interchangeable moves leave the context undecided between many solutions. The ancestral sampler of (14) retains a commitment through its reverse bridge, which favors the current token, and its gap closes as the NFE grows. The gap is small on maze, where each move is pinned by its neighbors along the path, and reverses on Sudoku, plausibly because its constraints let the cavity determine the unique solution from the context alone, so that committing to an imperfect current guess hurts slightly more than it helps; consistently, the relative NELBO increase is smallest on these two tasks. Samplers can thus respond very differently to what the model has actually learned, pointing toward a promising research direction in sampler design.

## 5.2 HEAD AGREEMENT AND OWN-TOKEN SENSITIVITY

Table 3 compares trained heads through the Jensen–Shannon divergence (App. G.3): across coordinates, and before and after their own noisy token is redrawn. After conversion to the denoiser,

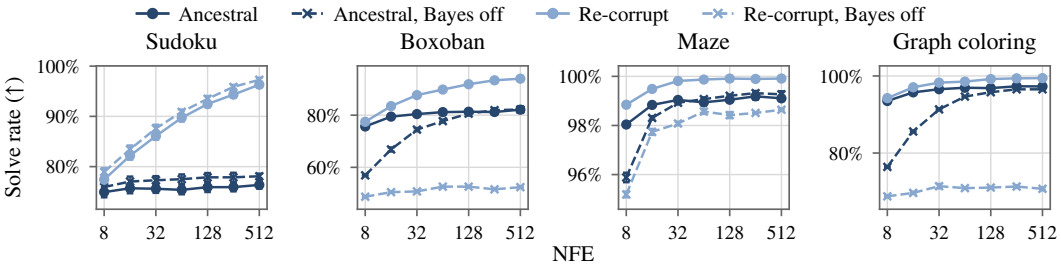


Figure 2: Single-attempt solve rate against NFE on the logic tasks, for the uniform-diffusion cavity models under the ancestral and re-corrupt samplers, with the Bayes reweighting and without it.

each pair of independently trained heads agrees to within $0.05$ nats/token, while a cavity read as a denoiser differs from its own conversion $3.5$ to $7$ times more, so the coordinates indeed learn the same object; under masked diffusion, the two coincide (Corollary 3). Since the standard transformer exposes each position's own noisy token, nothing forces a cavity head to ignore it, yet trained cavity heads largely do: they move a quarter to a third as much as the denoisers when the token is redrawn. The Bayes reweighting accounts for most of this difference, as applying it to the cavity reproduces the denoiser's own-token dependence, while removing it from the denoiser leaves about twice the cavity's. Figures 7 and 8 (App. G.4) resolve both statistics in diffusion time.

Table 3: Jensen–Shannon divergence (nats/token) between trained heads. Agreement: pairs of heads converted to the denoiser, and a cavity compared to its own denoiser conversion. Own-token sensitivity: a head before and after its noisy token is redrawn, in its own coordinate or converted ($\to$).

| | Agreement | | | | Own-token sensitivity | | | |
|---|---|---|---|---|---|---|---|---|
| | den.$\leftrightarrow$cav. | den.$\leftrightarrow$score | cav.$\leftrightarrow$score | cav. read as den. | cav. | den.$\to$cav. | den. | cav.$\to$den. |
| Masked | — | — | 0.046 | 0 | — | — | — | — |
| GIDD $\lambda$=0.1 | 0.026 | 0.030 | 0.031 | 0.184 | 0.107 | 0.200 | 0.340 | 0.338 |
| GIDD $\lambda$=0.2 | 0.027 | 0.034 | 0.035 | 0.182 | 0.083 | 0.157 | 0.298 | 0.296 |
| Uniform | 0.033 | 0.050 | 0.048 | 0.173 | 0.047 | 0.081 | 0.206 | 0.203 |

### 5.3 Trainability and calibration of the three coordinates

The three coordinates all train the same object and share the NELBO floor (Theorem 2), so any difference between them must be a difference in trainability. Figure 3 (top) shows the NELBO over training for every run, with the calibration identities of Proposition 4 visible at step 0: cavity and score start at $\log V$ for every kernel, while the denoiser starts far higher, finite on the clipped window but divergent as $t_1 \downarrow 0$. The score coordinate lags behind on every kernel in both NELBO and sample quality, plausibly because it must learn the normalization that the other two get by construction. Its quality–entropy frontier, traced by the sampling temperature, sits below the other two, worst on uniform diffusion (Figure 3, bottom), consistent with the gap between SEDD and MDLM (Lou et al., 2024; Sahoo et al., 2024).[4] Despite its higher initial NELBO, the denoiser recovers over training, and it is indistinguishable from the cavity at our resolution, in NELBO and along the frontier (further results, such as the GenPPL frontier in Figure 10, in App. G.4).

## 6 Related work

The CTMC bound is due to Campbell et al. (2022), with path-KL readings in the SDE and general-Markov lines (Song et al., 2021; Benton et al., 2024); these absorb the boundary terms into constants or send them to zero, while Proposition 1 keeps them and Theorem C.1 is exact for arbitrary model rates on clipped windows. The method lineages follow the three coordinates: the score line (Lou et al., 2024; Ou et al., 2025; Zhang et al., 2025) and the plug-in line (Austin et al., 2021; Schiff et al., 2025; von Rütte et al., 2025) are the score and cavity coordinates, and the masked cross-entropy

[4] Score samples also pass through the clipped inverse of App. G.1, which on test data moves 4–5% of the noisy-token probability mass on average. It cannot explain the lag alone, which also appears in the NELBO, computed from the raw score, and under masked diffusion, where the inverse is exact.

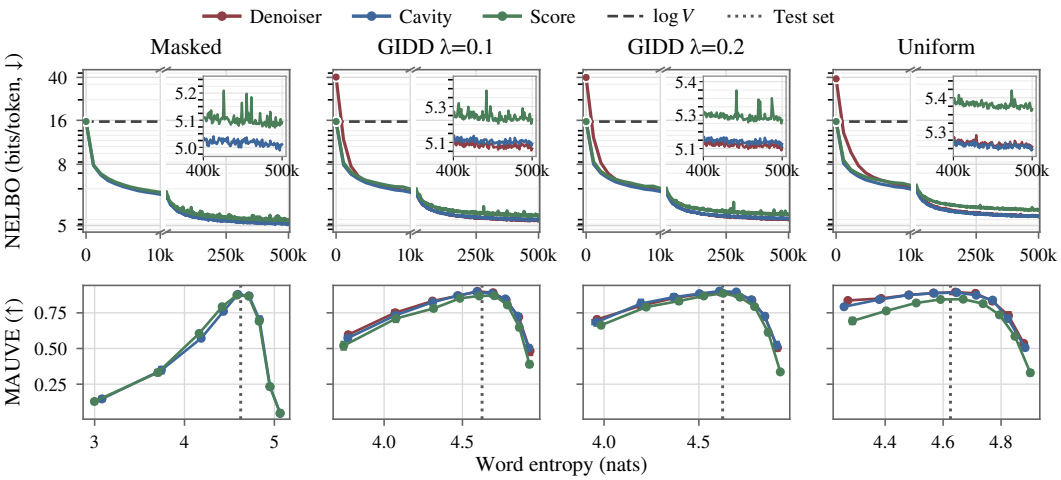


Figure 3: Top: online validation NELBO over training, where cavity and score start at $\log V$ on every kernel and the denoiser far higher (Proposition 4). Bottom: quality–entropy frontiers traced by the sampling temperature, MAUVE scored with GPT-2 Large against word entropy (App. G.3).

objectives (Sahoo et al., 2024; Shi et al., 2024) are the collapsed case of Corollary 3. Proposition 2 with (7) extends the discrete estimation–information identities (Jeon et al., 2025) to arbitrary model rates, finite windows, and every coordinate, and the invariance of the ELBO to time schedules (§3) is the CTMC counterpart of Kingma & Gao (2023). Sun et al. (2023) introduce the cavity for uniform diffusion, with its plug-in sampler, but train it by ratio matching rather than on the ELBO. The literature had not recognized that the ELBO of the bridge plug-in is optimized by the cavity rather than by the clean-data predictor: in D3PM (Austin et al., 2021), for instance, the auxiliary cross-entropy trains the denoiser while the plug-in loss and sampler assume the cavity; they report that this term helps their absorbing models and hurts their uniform ones (further details in App. H). Li et al. (2026) read the ELBO as a path KL up to an unspecified constant and list the three coordinates, but without the boundary terms, the Oracle Distance identity and its corollaries, the optimality of each coordinate, or the cavity finding. Concurrently and independently, Gourevitch et al. (2026) make the same observation for uniform diffusion. We obtain the cavity minimizer, the conversions, and the calibration identities for every token-factorizing process from one projection principle, and evaluate all three coordinates at GPT-2 scale across the standard noising processes.

## 7 Conclusion

We have answered the question from our title: a discrete diffusion model learns the exact time-reversal of the noising process. Denoiser, cavity, and concrete score are three coordinates of this same reverse process, related by exact conversions; in particular, the bridge plug-in learns the cavity rather than a denoiser. The Oracle Distance identity gives the corresponding global picture: the NELBO is exactly the data entropy plus the path KL from the oracle reverse process to the learned one, and its minimizer is the projection of the clean-conditioned reverse rates onto the noisy state. Together with explicit boundary terms and calibration identities at initialization, this yields a conceptual framework and a mechanical recipe for deriving, implementing, and comparing discrete diffusion processes.

Our coordinate theory currently assumes token-factorizing corruption; extending it to forward processes with joint multi-token transitions is a natural direction. Within the factorized setting, our framework unifies the forward noising process and the coordinate in which its reverse is represented and learned, while the sampler remains a separate approximation layer. Bridging the learned continuous-time reverse process to the different finite-step sampling schemes used in practice is therefore an important remaining problem. In this sense, the learned reverse process is unified; its realization by practical samplers is not.

### Ethics statement

This work is theoretical and methodological: it characterizes the training objective of discrete diffusion models and validates the analysis with small research models trained on OpenWebText, a public web corpus with known biases. We release no models or systems intended for deployment; the usual considerations for language-model research apply.

### Reproducibility statement

App. G specifies the training recipe, the data pipeline and splits, and every evaluation procedure and operating point. App. E gives the per-kernel loss, conversion, and sampler formulas in closed form. Proofs of every stated result are in Apps. B–E. The code used for the experiments will be released with the paper.

### AI use statement

We used large language models to aid and polish writing (editing and reviewing drafts), for retrieval and discovery (searching for and summarizing related work), and for research ideation and execution (discussing ideas and derivations, and implementing, running, and analyzing experiments). The authors reviewed all AI-assisted work, verified every technical claim and result, and take full responsibility for the content of this paper.

## References

Alan N. Amin, Nate Gruver, and Andrew Gordon Wilson. Why masking diffusion works: Condition on the jump schedule for improved discrete diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=vx6bPQWpmL.

Rami Atar and Tsachy Weissman. Mutual information, relative entropy, and estimation in the Poisson channel. *IEEE Transactions on Information Theory*, 58(3):1302–1318, 2012. doi: 10.1109/TIT.2011.2172572. URL https://doi.org/10.1109/TIT.2011.2172572.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 17981–17993, 2021. URL https://openreview.net/forum?id=h7-XixPCAL.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005a. doi: 10.1109/TIT.2005.850145. URL https://doi.org/10.1109/TIT.2005.850145.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005b. URL https://jmlr.org/papers/v6/banerjee05b.html.

Joe Benton, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. From denoising diffusions to denoising Markov models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):286–301, 2024. doi: 10.1093/jrsssb/qkae005. URL https://doi.org/10.1093/jrsssb/qkae005.

Pierre Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer Series in Statistics. Springer, 1981. ISBN 978-0-387-90536-5.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://openreview.net/forum?id=DmT862YAieY.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design.

In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5453–5512. PMLR, 2024. URL https://openreview.net/forum?id=kQwSbv0BR4.

Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: Exact implementation through uniformization, 2024. URL https://arxiv.org/abs/2402.08095.

Adrià Garriga-Alonso, Mohammad Taufeeque, and Adam Gleave. Planning behavior in a recurrent neural network that plays Sokoban. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=T9sB3S2hok.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://openreview.net/forum?id=GTDKo3Sv9p.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, Hao Peng, and Lingpeng Kong. Scaling diffusion language models via adaptation from autoregressive models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=j1tSLYKwg8.

Samson Gourevitch, Yazid Janati, Dario Shariatian, Umut Simsekli, Eric Moulines, Eric P. Xing, and Alain Durmus. Uniform diffusion models revisited: Leave-one-out denoiser and absorbing state reformulation, 2026. URL https://arxiv.org/abs/2605.22765.

Robert M. Gray. *Entropy and Information Theory*. Springer, 2 edition, 2011. doi: 10.1007/978-1-4419-7970-4. URL https://doi.org/10.1007/978-1-4419-7970-4.

Arthur Guez, Mehdi Mirza, Karol Gregor, Rishabh Kabra, Sébastien Racanière, Théophane Weber, David Raposo, Adam Santoro, Laurent Orseau, Tom Eccles, Greg Wayne, David Silver, Timothy Lillicrap, and Victor Valdes. An investigation of model-free planning: Boxoban levels. https://github.com/deepmind/boxoban-levels/, 2018.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. doi: 10.1109/TIT.2005.844072. URL https://doi.org/10.1109/TIT.2005.844072.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary Markov processes. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=RuP17cJtZo.

Chunsan Hong, Sanghyun Lee, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, Yuki Mitsufuji, Seungryong Kim, and Jong Chul Ye. Understanding and accelerating the training of masked diffusion language models, 2026. URL https://arxiv.org/abs/2605.13026.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 12454–12465, 2021. URL https://openreview.net/forum?id=6nbpPqUCIi7.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=Lm8T39vLDTE.

Chin-Wei Huang, Jae Hyun Lim, and Aaron Courville. A variational perspective on diffusion-based generative models and score matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://openreview.net/forum?id=bXehDYUjjXi.

Jean Jacod and Albert N. Shiryaev. *Limit Theorems for Stochastic Processes*, volume 288 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2 edition, 2003. doi: 10.1007/978-3-662-05265-5. URL https://doi.org/10.1007/978-3-662-05265-5.

Dominik Janzing, Rafael Chaves, and Bernhard Schölkopf. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(9):093052, 2016. doi: 10.1088/1367-2630/18/9/093052. URL https://doi.org/10.1088/1367-2630/18/9/093052.

Moongyu Jeon, Sangwoo Shin, Dongjae Jeon, and Albert No. Information-theoretic discrete diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=B2iPEX5A9c.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://openreview.net/forum?id=NnMEadcdyD.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 21696–21707, 2021. URL https://openreview.net/forum?id=2LdBqxc1Yv.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=UvmDCdSPDOW.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=X6tNkN6ate.

Jingyuan Li, Xiaoyi Jiang, Fukang Wen, Wei Liu, Renqian Luo, Yi Zhu, Zuoqiang Shi, and Pipi Hu. Neural continuous-time Markov chain: Discrete diffusion via decoupled jump timing and direction, 2026. URL https://arxiv.org/abs/2604.15694.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training, 2025. URL https://arxiv.org/abs/2502.16982.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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 (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 2024. URL https://openreview.net/forum?id=CNicRIVIPA.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://openreview.net/forum?id=_RL7wtHkPJK.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=KnqiC0znVF.

J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. doi: 10.1017/CBO9780511810633. URL https://doi.org/10.1017/CBO9780511810633.

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 (ICLR)*, 2025. URL https://openreview.net/forum?id=sMyXP8Tanm.

Vincent Pauline, Tobias Höppe, Kirill Neklyudov, Alexander Tong, Stefan Bauer, and Andrea Dittadi. Foundations of diffusion models in general state spaces: A self-contained introduction, 2025. URL https://arxiv.org/abs/2512.05092.

Jonas Peters, Dominik Janzing, Arthur Gretton, and Bernhard Schölkopf. Detecting the direction of causal time series. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pp. 801–808. ACM, 2009. doi: 10.1145/1553374.1553477. URL https://doi.org/10.1145/1553374.1553477.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pp. 154–162, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/47d1e990583c9c67424d369f3414728e-Abstract.html.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 4816–4828, 2021. URL https://openreview.net/forum?id=Tqx7nJp7PR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Gabriel Raya, Bac Nguyen, Georgios Batzolis, Yuhta Takida, Dejan Stancevic, Naoki Murata, Chieh-Hsin Lai, Yuki Mitsufuji, and Luca Ambrogioni. Noise scheduling as information-guided allocation in diffusion training, 2026. URL https://arxiv.org/abs/2602.18647.

Yinuo Ren, Haoxuan Chen, Grant M. Rotskoff, and Lexing Ying. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=6awxwQEI82.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286. PMLR, 2014. URL https://proceedings.mlr.press/v32/rezende14.html.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, 2 edition, 2004. doi: 10.1007/978-1-4757-4145-2. URL https://doi.org/10.1007/978-1-4757-4145-2.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T. Chiu, and Volodymyr Kuleshov. The diffusion duality. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52584–52619. PMLR, 2025. URL https://openreview.net/forum?id=9P9Y8FOSOk.

Subham Sekhar Sahoo, Jean-Marie Lemercier, Zhihan Yang, Justin Deschenaux, Jingyu Liu, John Thickstun, and Ante Jukic. Scaling beyond masked diffusion language models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=xKJG1CNtQM.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dallatorre, Bernardo P. de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=i5MrJ6g5G1.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1255–1262. Omnipress, 2012. URL https://icml.cc/2012/papers/625.pdf.

Neta Shaul, Itai Gat, Marton Havasi, Daniel Severo, Anuroop Sriram, Peter Holderrieth, Brian Karrer, Yaron Lipman, and Ricky T. Q. Chen. Flow matching with general discrete paths: A kinetic-optimal perspective. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=tcvMzR2NrP.

Jiaxin Shi and Michalis K. Titsias. Demystifying diffusion objectives: Reweighted losses are better variational bounds, 2025. URL https://arxiv.org/abs/2511.19664.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 103131–103167, 2024. URL https://openreview.net/forum?id=xcqSOfHt4g.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://openreview.net/forum?id=AklttWFnxS9.

Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=BYWWwSY2G5s.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/neco_a_00142. URL https://doi.org/10.1162/neco_a_00142.

Dimitri von Rütte, Janis Fluri, Yuhui Ding, Antonio Orvieto, Bernhard Schölkopf, and Thomas Hofmann. Generalized interpolating discrete diffusion. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61810–61843. PMLR, 2025. URL https://openreview.net/forum?id=rvZv7sDPV9.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model, 2025. URL https://arxiv.org/abs/2506.21734.

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

Ruixiang Zhang, Shuangfei Zhai, Yizhe Zhang, James Thornton, Zijing Ou, Joshua M. Susskind, and Navdeep Jaitly. Target concrete score matching: A holistic framework for discrete diffusion. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76716–76753. PMLR, 2025. URL https://openreview.net/forum?id=ZMrdvSm7xi.

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 (ICLR)*, 2025. URL https://openreview.net/forum?id=CTC7CmirNr.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=PEQFHRUFca.

## APPENDIX CONTENTS


**A Continuous-time Markov chain discrete diffusion** **17**

A.1 Variational formulation of diffusion models . . . . . 17

A.2 CTMC formulation . . . . . 17

**B The finite-window CTMC ELBO** **19**

B.1 First derivation: the infinitesimal-KL limit . . . . . 20

B.2 Second derivation: CTMC Girsanov formula . . . . . 21

B.3 Importance sampling, time clocks, and the empirical ELBO . . . . . 22

**C The Oracle Distance identity** **23**

C.1 The ELBO as a projection problem and the Pythagorean decomposition . . . . . 23

C.2 The oracle ELBO is the rate of information loss . . . . . 24

C.3 ELBO as distance to the oracle reverse process . . . . . 25

C.4 The universal attainable NELBO floor . . . . . 26

C.5 A direct proof of the Oracle Distance identity . . . . . 26

C.6 The arrow of time and independent mechanisms . . . . . 27

C.7 A Bregman-divergence and generator-matching view . . . . . 27

**D Sequence modeling** **28**

D.1 Token-factorizing noising processes . . . . . 28

D.2 Three coordinates of the reverse rate . . . . . 29

**E Masked, uniform, and GIDD noising processes** **32**

E.1 Common CTMC form . . . . . 32

E.2 Masked diffusion . . . . . 33

E.3 Uniform diffusion . . . . . 34

E.4 Generalized interpolating discrete diffusion . . . . . 35

**F Exact verification on a solvable model** **37**

**G Experimental details** **39**

G.1 Implementation details . . . . . 39

G.2 Logic tasks . . . . . 40

G.3 Metric estimation details . . . . . 40

G.4 Further experiments . . . . . 42

**H Extended related work** **44**

**I Generated samples** **45**

# A CONTINUOUS-TIME MARKOV CHAIN DISCRETE DIFFUSION

This appendix collects the variational and Markov-chain background the paper rests on; see Norris (1997); Brémaud (1981); Jacod & Shiryaev (2003) for textbook references.

## A.1 VARIATIONAL FORMULATION OF DIFFUSION MODELS

We begin with the variational identity underlying VAEs (Kingma & Welling, 2014; Rezende et al., 2014). Let $p^\theta(z_0, z_1)$ be a generative model for a finite data distribution $z_0 \sim q_0$ on $\mathcal{X}$ and a latent variable $z_1$, and let $q(z_1 \mid z_0)$ be an encoder satisfying $q(z_1 \mid z_0) \ll p^\theta(z_0, \cdot)$ for every $z_0$. Jensen's inequality for the concave log gives

$$\log p_0^\theta(z_0) = \log \mathbb{E}_{q(z_1|z_0)}\left[\frac{p^\theta(z_0, z_1)}{q(z_1 \mid z_0)}\right] \geq \mathbb{E}_{q(z_1|z_0)}\left[\log \frac{p^\theta(z_0, z_1)}{q(z_1 \mid z_0)}\right] =: \mathrm{ELBO}(\theta;\, z_0), \tag{A.1}$$

the evidence lower bound, which splits as

$$\mathrm{ELBO}(\theta;\, z_0) = \mathbb{E}_{q(z_1|z_0)}\big[\log p^\theta(z_0 \mid z_1)\big] - \mathrm{KL}\big(q(\cdot \mid z_0) \,\big\|\, p^\theta_{z_1}\big)\,,$$

a reconstruction term and a prior-matching term. The gap of the inequality is exactly a KL divergence between the encoding and model posterior laws (throughout the appendix, *law* is a synonym of *distribution*, and $\mathrm{Law}(X)$ denotes the distribution of $X$):

$$\log p_0^\theta(z_0) - \mathrm{ELBO}(\theta;\, z_0) = \mathrm{KL}\big(q(\cdot \mid z_0) \,\big\|\, p^\theta(\cdot \mid z_0)\big) \geq 0. \tag{A.2}$$

Diffusion models are VAEs whose latent variable is the noising trajectory. Fix $z_0 \in \mathcal{X}$ and let $z_{1:N}$ be a discrete-time Markov trajectory with $N \geq 1$, initial law $q_0$, and forward kernels $q_{t|t-1}$, so that $q_{1:N|0}(z_{1:N} \mid z_0) = \prod_{t=1}^N q_{t|t-1}(z_t \mid z_{t-1})$. The learned reverse-time model factors as $p^\theta(z_{0:N}) = p_N^\theta(z_N) \prod_{t=1}^N p^\theta_{t-1|t}(z_{t-1} \mid z_t)$ and is also Markov in forward time. In the VAE reading, $z_{1:N}$ is the latent variable and $p_N^\theta$ its prior.

**Proposition A.1** (Discrete-time diffusion ELBO)**.** *Assume the KL terms below are well defined. Then*

$$\begin{aligned}\mathrm{ELBO}(\theta;\, z_0) = \mathbb{E}_{q_{1|0}(z_1|z_0)}\big[\log p^\theta_{0|1}(z_0 \mid z_1)\big] - \mathrm{KL}\big(q_{N|0}(\cdot \mid z_0) \,\big\|\, p_N^\theta\big) \\ - \sum_{t=2}^{N} \mathbb{E}_{q_{t|0}(z_t|z_0)}\left[\mathrm{KL}\big(q_{t-1|t,0}(\cdot \mid z_t, z_0) \,\big\|\, p^\theta_{t-1|t}(\cdot \mid z_t)\big)\right],\end{aligned} \tag{A.3}$$

*and* $\log p_0^\theta(z_0) - \mathrm{ELBO}(\theta;\, z_0) = \mathrm{KL}\big(q_{1:N|0}(\cdot \mid z_0) \,\big\|\, p^\theta_{1:N|0}(\cdot \mid z_0)\big) \geq 0.$

*Proof.* Start from the VAE ELBO equality with latent $z_{1:N}$, and use the reverse factorization $q_{1:N|0}(z_{1:N} \mid z_0) = q_{N|0}(z_N \mid z_0) \prod_{t=2}^N q_{t-1|t,0}(z_{t-1} \mid z_t, z_0)$ together with the model factorization above:

$$\mathrm{ELBO}(\theta;\, z_0) = \mathbb{E}_{q_{1:N|0}(z_{1:N}|z_0)}\left[\log \frac{p_N^\theta(z_N) \prod_{t=1}^N p^\theta_{t-1|t}(z_{t-1} \mid z_t)}{q_{N|0}(z_N \mid z_0) \prod_{t=2}^N q_{t-1|t,0}(z_{t-1} \mid z_t, z_0)}\right].$$

Splitting the logarithm into boundary and transition terms gives (A.3). The gap identity is the VAE identity (A.2) with latent $z_{1:N}$. □

The discrete-time ELBO already has the structure that persists in continuous time: a reconstruction term at the left endpoint, a prior term at the right endpoint, and a sum of local discrepancies between the clean-conditioned reverse kernels and the learned reverse kernels.

## A.2 CTMC FORMULATION

**Notation.** A single letter $q$ or $p$ denotes the finite-dimensional marginals, kernels, and bridges of a process, with subscripts recording the relevant times: $q_t$ is the time-$t$ marginal, $q_{t|s}$ the transition kernel for $s < t$, $q_{t|0}(\cdot \mid z_0)$ the forward kernel from clean data, and $q_{s|t,0}(\cdot \mid z_t, z_0)$ the bridge. We identify a random variable with its value when clear, writing $q(z_0 \mid z_t)$ for $q(Z_0 = z_0 \mid Z_t = z_t)$. Jump rates take $Q, R$ (and $a, b$ for generic rates); a hat denotes time reversal. We abbreviate "almost every(where)" as a.e. and "almost surely" as a.s. For laws, $\mu \ll \nu$ denotes absolute continuity (on finite spaces, $\mathrm{supp}\,\mu \subseteq \mathrm{supp}\,\nu$), with relative density $d\mu/d\nu$. $H(Z)$ is the entropy of a discrete variable and $I(X;Y) = H(X) - H(X \mid Y)$ the mutual information.

**CTMC processes and their generators.** Let $\mathcal{X}$ be finite and $0 < T \le \infty$. A CTMC $(Z_t)_{t\in[0,T]}$ on $\mathcal{X}$ is fixed by a time-dependent rate matrix (generator) $Q_t$ with $Q_t(x,y) \ge 0$ for $x \ne y$ and $Q_t(x,x) = -\sum_{y\ne x} Q_t(x,y)$, through

$$q_{t+h|t}(y \mid x) = \mathbf{1}\{x = y\} + h\, Q_t(x,y) + o(h), \qquad h \downarrow 0. \tag{A.4}$$

Sample paths are right-continuous and piecewise constant.

**Kolmogorov equations.** Arranged as matrices with normalized rows, the kernels obey Chapman–Kolmogorov $q_{u|s} = q_{t|s}q_{u|t}$ and the forward and backward linear ODEs

$$\partial_t q_{t|s} = q_{t|s}Q_t, \qquad \partial_s q_{t|s} = -Q_s q_{t|s}, \qquad q_{s|s} = I.$$

With $q_0$ given, the marginal solves $\partial_t q_t = q_t Q_t$, and by uniqueness of solutions of linear ODEs the law of the chain is determined by $q_0$ and $Q_t$.

**Path laws and trajectory likelihoods.** Three standard facts describe the process at the path level. The first is well-posedness: a law at the time from which the process is run (the left endpoint for a forward chain, the right endpoint for a reverse-time chain), together with bounded, measurable rates on the window $[s,u]$, determines a unique CTMC *path law*, a probability measure on the space of right-continuous, piecewise-constant trajectories, just as $q_0$ and $Q_t$ determine the marginals through the Kolmogorov equations. Every process in this paper is specified this way, by an endpoint law and a jump-rate family. Path laws take an uppercase $P$, subscripted by the window as in $P_{[s,u]}$, and integration against them is written $P(dz_{[s,u]})$; finite-space factors keep plain pmf notation, so a joint law over a data point and a trajectory reads $P(z_0, dz_{[s,u]})$.

The second fact is the trajectory likelihood. On a finite state space with bounded rates on a compact window, the chain makes finitely many jumps a.s., so a path is described by its jump coordinates: the initial state $x_0$, the number of jumps $n$, the ordered jump times $s < r_1 < \cdots < r_n \le u$, and the post-jump states $x_1, \dots, x_n$. With respect to the reference measure on these jump coordinates, a CTMC with initial law $p_s$ and generator $Q_t$ has density

$$p_s(x_0)\, \exp\Big( \int_s^u Q_t(Z_t, Z_t)\, dt \Big) \prod_{k=1}^{n} Q_{r_k}(x_{k-1}, x_k)\,: \tag{A.5}$$

the initial mass, the no-jump survival factor accumulated from the exit rate, and the rate of each jump. This density is relative to the jump-coordinate reference measure; only ratios of such densities enter below, for which the reference measure cancels.

Third, the compensated jump sum $M_t^{xy} := N_t^{xy} - \Lambda_t^{xy}$, with $N_t^{xy}$ the number of $x \to y$ jumps in $[s,t]$ and $\Lambda_t^{xy} := \int_s^t Q_r(Z_r, y)\mathbf{1}\{Z_r = x\}\, dr$ its compensator, is a mean-zero martingale. In particular, for any $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}\left[ \sum_{s<r\le u} f(Z_{r^-}, Z_r)\, \mathbf{1}\{Z_{r^-} \ne Z_r\} \right] = \mathbb{E}\left[ \int_s^u \sum_{y \ne Z_t} Q_t(Z_t, y)\, f(Z_t, y)\, dt \right]. \tag{A.6}$$

**Time reversal.** The time-reversed process is again a CTMC. When $q_t(x) > 0$ its generator is $\widehat{Q}_t(x,y) = Q_t(y,x)\, q_t(y)/q_t(x)$ for $x \ne y$, by the Markov property and Bayes' rule; conditioning the reversal on the endpoint $z_0$ replaces $q_t$ by $q_{t|0}(\cdot \mid z_0)$ and gives the clean-conditioned reverse rate of (1). In terms of small-time kernels,

$$q_{t-h|t,0}(y \mid z_t, z_0) = \mathbf{1}\{z_t = y\} + h\, \widehat{Q}_t(z_t, y \mid z_0) + o(h), \qquad h \downarrow 0.$$

Direction is thus a property of the chosen factorization, an endpoint law and rates, not of the law itself: a process Markov in forward time is Markov in reverse time, with different rates.

**Definition A.1** (Noising and denoising path laws)**.** *Fix a window* $[t_1, t_2] \subset (0, T)$ *and let* $(Z_t)_{t\in\{0\}\cup[t_1,t_2]}$ *be the noising process with* $Z_0 \sim q_0$. *The* clean-conditioned noising path law *is*

$$P^{\star}_{[t_1,t_2]|0}(\cdot \mid z_0) := \mathrm{Law}\big((Z_t)_{t\in[t_1,t_2]} \mid Z_0 = z_0\big),$$

*a probability measure on the path space over* $[t_1, t_2]$.[5] *Prefixing the data law gives the* joint *law*

$$P^\star_{0,[t_1,t_2]}(z_0, dz_{[t_1,t_2]}) = q_0(z_0)\, P^\star_{[t_1,t_2]|0}(dz_{[t_1,t_2]} \mid z_0),$$

*and marginalizing out* $z_0$ *gives the* marginal noising path law $P^\star_{[t_1,t_2]} := \mathrm{Law}\big((Z_t)_{t\in[t_1,t_2]}\big)$.

*The* model (denoising) path law $P^\theta_{[t_1,t_2]}$ *is initialized at* $p^\theta_{t_2}$ *and run backward with* $\widehat{Q}^\theta_t(z_t, y)$. *Attaching the data through a reconstruction kernel extends it to the model joint*

$$P^\theta_{0,[t_1,t_2]}(z_0, dz_{[t_1,t_2]}) = P^\theta_{[t_1,t_2]}(dz_{[t_1,t_2]})\, p^\theta_{0|t_1}(z_0 \mid z_{t_1}).$$

All of these are well defined by well-posedness and, by Kolmogorov uniqueness, each noising law is equivalently characterized by running the corresponding rates from either endpoint of the interval, conditionally on $Z_0$ or not, forward or backward. For instance, the clean-conditioned law $P^\star_{[t_1,t_2]|0}$ run backward is the CTMC started at $q_{t_2|0}(\cdot \mid z_0)$ with the clean-conditioned reverse rates, and the joint law $P^\star_{0,[t_1,t_2]}$ read entirely in reverse time starts at $q_{t_2}$, runs backward with the marginal reverse rates $\widehat{Q}_t$, and attaches the data through $q_{0|t_1}(\cdot \mid z_{t_1})$, using that $Z_0$ is conditionally independent of $Z_{[t_1,t_2]}$ given $Z_{t_1}$. The joint model law is determined by the triple $(p^\theta_{t_2}, \widehat{Q}^\theta_t, p^\theta_{0|t_1})$, where $\widehat{Q}^\theta_t$ may depend on the noisy state but not on the clean data.

The practical interest of diffusion rests on the terminal mixing assumption below; most results in this paper do not need it.

**Assumption A.1** (Terminal mixing)**.** *As* $t \uparrow T$ *(including* $T = \infty$*), the forward process forgets the data:* $q_{t|0}(\cdot \mid z_0) \to q_T$ *for every* $z_0 \in \operatorname{supp} q_0$*, where* $q_T$ *does not depend on* $z_0$*. In particular the joint law of* $(Z_0, Z_t)$ *converges to the product* $q_0 \otimes q_T$ *and* $I(Z_0; Z_t) \to 0$.

## B THE FINITE-WINDOW CTMC ELBO

We state the ELBO on an interior compact window $[t_1, t_2] \subset (0, T)$, avoiding potentially singular endpoints as is done in practice. The local rate divergence $\Phi$ is defined in (2), with the conventions $\Phi(0, b) = b$ and $\Phi(a, 0) = +\infty$ for $a > 0$; it is nonnegative and vanishes iff $a = b$.

**Assumption B.1** (Rate-support condition)**.** *The learned reverse rates satisfy* $\widehat{Q}_t(z_t, \cdot \mid z_0) \ll \widehat{Q}^\theta_t(z_t, \cdot)$ *for* $dt\, dq(z_t, z_0)$*-a.e.* $(t, z_t, z_0)$*, that is, the learned process can jump wherever the true reverse process can.*

**Theorem B.1** (CTMC ELBO, complete form)**.** *Let* $\mathcal{X}$ *be finite and let the true and learned reverse CTMCs have bounded, measurable rates* $\widehat{Q}_t(\cdot \mid z_0)$ *and* $\widehat{Q}^\theta_t$ *on* $[t_1, t_2] \subset (0, T)$*. Define the per-time jump divergence*

$$\mathcal{J}^\theta_t(z_0, z_t) := \sum_{y \neq z_t} \Phi\big(\widehat{Q}_t(z_t, y \mid z_0), \widehat{Q}^\theta_t(z_t, y)\big). \tag{B.1}$$

*Assume the rate-support condition (Assumption B.1), the terminal absolute continuity* $q_{t_2|0}(\cdot \mid z_0) \ll p^\theta_{t_2}$*, integrability* $\int_{t_1}^{t_2} \mathbb{E}_{q_{t|0}(z_t|z_0)}\big[\mathcal{J}^\theta_t(z_0, z_t)\big]\, dt < \infty$*, and that the reconstruction term below is well defined. Let* $P^\theta_{0,[t_1,t_2]}(z_0, \cdot)$ *be the fixed-*$z_0$ *section of the model joint law (Definition A.1), a finite path measure of total mass* $p^\theta_0(z_0)$*. With* $z_{[t_1,t_2]}$ *as latent variable, the ELBO defined at the path-measure level by*

$$\mathrm{ELBO}_{[t_1,t_2]}(\theta;\, z_0) := -\mathrm{KL}\Big(P^\star_{[t_1,t_2]|0}(\cdot \mid z_0) \,\Big\|\, P^\theta_{0,[t_1,t_2]}(z_0, \cdot)\Big) \tag{B.2}$$

*admits the expanded form*

$$\begin{aligned}\mathrm{ELBO}_{[t_1,t_2]}(\theta;\, z_0) = \underbrace{\mathbb{E}_{q_{t_1|0}(z_{t_1}|z_0)}\big[\log p^\theta_{0|t_1}(z_0 \mid z_{t_1})\big]}_{\textit{reconstruction}} - \underbrace{\mathrm{KL}\big(q_{t_2|0}(\cdot \mid z_0) \,\big\|\, p^\theta_{t_2}\big)}_{\textit{terminal prior}} \\ - \underbrace{\int_{t_1}^{t_2} \mathbb{E}_{q_{t|0}(z_t|z_0)}\big[\mathcal{J}^\theta_t(z_0, z_t)\big]\, dt}_{\textit{path integral}}.\end{aligned} \tag{B.3}$$

[5] Formally, the Skorokhod space $D([t_1, t_2], \mathcal{X})$ of càdlàg paths with the $\sigma$-field generated by the coordinate maps. This $\sigma$-field is generated by cylinder events, so a path law is determined by its finite-dimensional marginals, which are in turn determined by an endpoint law and the jump rates through the Kolmogorov equations; existence follows from the construction of the chain from its jump times and jump chain. This is the rigorous content of the well-posedness fact above.

*Moreover the ELBO gap is the relative entropy between the clean-conditioned path laws,*

$$\log p_0^\theta(z_0) - \mathrm{ELBO}_{[t_1,t_2]}(\theta;\, z_0) = \mathrm{KL}\Big(P^\star_{[t_1,t_2]|0}(\cdot \mid z_0) \,\Big\|\, P^\theta_{[t_1,t_2]|0}(\cdot \mid z_0)\Big) \geq 0,$$

*with* $P^\theta_{[t_1,t_2]|0}(\cdot \mid z_0) := P^\theta_{0,[t_1,t_2]}(z_0,\cdot)/p_0^\theta(z_0)$ *the normalized model posterior path law.*

In (B.2) the KL symbol is a mean log-ratio against a finite measure of mass $p_0^\theta(z_0)$, well defined under the stated absolute-continuity hypotheses, the Radon–Nikodym derivative being a ratio of densities of the form (A.5); this is what makes (B.2) a lower bound on $\log p_0^\theta(z_0)$ rather than an ordinary divergence. Proposition 1 in the main text is the expanded form together with the sign of the gap.

### B.1 First derivation: the infinitesimal-KL limit

The first derivation rests on an infinitesimal expansion of one-step KLs between jump kernels.

**Lemma B.1** ($\Phi$ as the differential KL of jump kernels)**.** *Let $a_t, b_t$ be piecewise continuous CTMC rates on a finite space $\mathcal{X}$ and a compact interval $[t, t+\Delta]$, $\Delta > 0$, with transition kernels $q^a_{t,t+\Delta}$, $p^b_{t,t+\Delta}$. Assume the jump graph of $a$ is a fixed set $E$ and, for constants $0 < c < C < \infty$ and every $t$,*

$$c \leq a_t(x,y), b_t(x,y) \leq C \;\; \forall (x,y) \in E; \qquad a_t(x,y) = 0 \;\; \forall (x,y) \notin E; \qquad 0 \leq b_t(x,y) \leq C. \tag{B.4}$$

*Then, uniformly in $t$ and $x$,*

$$\mathrm{KL}\big(q^a_{t,t+\Delta}(\cdot \mid x) \,\big\|\, p^b_{t,t+\Delta}(\cdot \mid x)\big) = \Delta \sum_{y \neq x} \Phi\big(a_t(x,y), b_t(x,y)\big) + o(\Delta).$$

*Proof.* We expand the KL by destination, according to whether $(x,y) \in E$.

For jumps $x \neq y$ with $(x,y) \in E$, integrating the uniformly continuous rates gives $q^a_{t,t+\Delta}(y \mid x) = \Delta a_t(x,y) + o(\Delta)$ and $p^b_{t,t+\Delta}(y \mid x) = \Delta b_t(x,y) + o(\Delta)$. Because both rates are bounded below and above on $E$,

$$q^a_{t,t+\Delta}(y \mid x) \log \frac{q^a_{t,t+\Delta}(y \mid x)}{p^b_{t,t+\Delta}(y \mid x)} = \Delta a_t(x,y) \log \frac{a_t(x,y)}{b_t(x,y)} + o(\Delta), \qquad (x,y) \in E. \tag{B.5}$$

For $x \neq y$ with $(x,y) \notin E$: if $y$ is unreachable from $x$ in the graph $E$ then $q^a_{t,t+\Delta}(y \mid x) = 0$. Otherwise let $d \geq 2$ be the length of a shortest $E$-path from $x$ to $y$. The uniform rate bounds and the finiteness of $\mathcal{X}$ bound the total exit rate from any state, so the number of jumps $N_\Delta$ in $[t, t+\Delta]$ is dominated by a Poisson variable and $\mathbb{P}(N_\Delta \geq k) = O(\Delta^k)$ uniformly. Reaching $y$ needs at least $d$ jumps, so $q^a_{t,t+\Delta}(y \mid x) = O(\Delta^d)$; following one shortest path and making no other jumps has probability at least a constant times $\Delta^d$ under both processes, while any single transition probability is $O(\Delta)$, so $|\log(q^a_{t,t+\Delta}/p^b_{t,t+\Delta})| = O(|\log \Delta|)$ wherever $q^a_{t,t+\Delta}(y \mid x) > 0$. Hence each such term is $O(\Delta^d |\log \Delta|) = O(\Delta^2 |\log \Delta|)$ and

$$\sum_{y \neq x,\, (x,y) \notin E} q^a_{t,t+\Delta}(y \mid x) \log \frac{q^a_{t,t+\Delta}(y \mid x)}{p^b_{t,t+\Delta}(y \mid x)} = o(\Delta). \tag{B.6}$$

For the self-transition, returning to $x$ after a jump needs another jump, of probability $O(\Delta^2)$, so with $A_t(x) := \sum_{y \neq x} a_t(x,y)$ and $B_t(x) := \sum_{y \neq x} b_t(x,y)$, $q^a_{t,t+\Delta}(x \mid x) = 1 - \Delta A_t(x) + o(\Delta)$ and likewise for $p^b$, giving the contribution

$$(1 - \Delta A_t(x) + o(\Delta)) \log \frac{1 - \Delta A_t(x) + o(\Delta)}{1 - \Delta B_t(x) + o(\Delta)} = \Delta(B_t(x) - A_t(x)) + o(\Delta). \tag{B.7}$$

Adding (B.7), (B.5), and (B.6), and collecting through $\Phi(a,b) = b - a + a\log(a/b)$, proves the claim, uniformly in $t$ and $x$. □

The additional hypotheses of Lemma B.1 (fixed jump graph, piecewise continuity, uniform bounds) are needed only for this first derivation; they hold for the masked, uniform, and fixed-support GIDD processes of App. E.

*Proof 1 of Theorem B.1 (infinitesimal KL).* Impose the conditions of Lemma B.1 on each continuity piece. Apply the discrete-time ELBO (Proposition A.1) to a finite partition $t_1 = s_0 < s_1 < \cdots < s_N = t_2$ with $\Delta_k := s_{k+1} - s_k$ and $|\Delta| := \max_k \Delta_k$:

$$\mathrm{ELBO}_{(s_k)_k}(\theta;\, z_0) = \mathbb{E}_{q_{t_1|0}(z_{t_1}|z_0)}\big[\log p^{\theta}_{0|t_1}(z_0 \mid z_{t_1})\big] - \mathrm{KL}\big(q_{t_2|0}(\cdot \mid z_0) \,\big\|\, p^{\theta}_{t_2}\big) - \sum_{k=0}^{N-1} \mathbb{E}_{q_{s_{k+1}|0}}\Big[\mathrm{KL}\Big(q_{s_k|s_{k+1},0}(\cdot \mid z_{s_{k+1}}, z_0) \,\Big\|\, p^{\theta}_{s_k|s_{k+1}}(\cdot \mid z_{s_{k+1}})\Big)\Big].$$

The two kernels in each summand are the exact transitions of the true and learned reverse CTMCs over an interval of length $\Delta_k$. Lemma B.1, applied in the backward orientation, turns each into $\Delta_k \sum_{y\neq x} \Phi(\widehat{Q}_{s_{k+1}}(x,y \mid z_0), \widehat{Q}^{\theta}_{s_{k+1}}(x,y)) + r_k(x)$ with uniform remainders $|r_k(x)| \le \Delta_k\, \eta(\Delta_k)$, $\eta(\delta) \to 0$. Taking expectations and summing, the main sum is a Riemann sum of the piecewise continuous integrand $t \mapsto \mathbb{E}_{q_{t|0}}\big[\mathcal{J}^{\theta}_t(z_0, z_t)\big]$ and converges to the path integral, while $|\sum_k \mathbb{E}[r_k]| \le (t_2 - t_1) \sup_{\delta \le |\Delta|} \eta(\delta) \to 0$. This yields (B.3) as the vanishing-mesh limit of the discrete-time ELBOs; to identify the limit with the path-level definition (B.2), take nested partitions whose union is dense in $[t_1, t_2]$: the skeletons generate the path $\sigma$-field and the skeleton KLs increase to the path KL (Gray, 2011, Lemma 5.4.1). The same skeleton limit applied to the discrete-time gap identity yields the path-level gap. □

### B.2 Second derivation: CTMC Girsanov formula

The second derivation rests on the change-of-measure formula for finite-state jump processes, stated for two generic CTMCs.

**Theorem B.2** (Girsanov formula and relative entropy for finite-state CTMCs)**.** *Let $\mathcal{X}$ be finite and $[s,u]$ compact. Let $P, P'$ be the path laws of two CTMCs specified by initial laws $p_s, p'_s$ at time $s$ and bounded, measurable rates $a_t, b_t$, with $p_s \ll p'_s$ and $a_t(x,\cdot) \ll b_t(x,\cdot)$ for every $x$ and a.e. $t$. Then $P \ll P'$ and*

$$\mathrm{KL}(P \,\|\, P') = \mathrm{KL}(p_s \,\|\, p'_s) + \int_s^u \mathbb{E}_{Z_t \sim P}\left[\sum_{y \neq Z_t} \Phi\big(a_t(Z_t,y), b_t(Z_t,y)\big)\right] dt, \tag{B.8}$$

*where both sides are finite if and only if the integral is.*

A KL between two CTMC path laws is not a pure rate mismatch: it also carries the KL between the two laws at the starting time.

*Proof.* Write $A_t(x) = \sum_{y\neq x} a_t(x,y)$ and $B_t(x) = \sum_{y\neq x} b_t(x,y)$ and apply the trajectory likelihood (A.5) to $P$ and $P'$. Under $a_t \ll b_t$ the jump-coordinate reference measure cancels in the density ratio, which is therefore well defined $P$-a.s., so $P \ll P'$ and

$$\frac{dP}{dP'} = \frac{p_s(Z_s)}{p'_s(Z_s)} \exp\Big(\int_s^u \big(B_t(Z_t) - A_t(Z_t)\big)\, dt\Big) \prod_{s<r\le u:\, Z_{r^-} \neq Z_r} \frac{a_r(Z_{r^-}, Z_r)}{b_r(Z_{r^-}, Z_r)}$$

(Brémaud, 1981; Jacod & Shiryaev, 2003). Taking $\log$ and integrating against $P$, the jump compensator (A.6) with $f = \log(a_t/b_t)$ turns the jump sum into $\int_s^u \mathbb{E}_P\Big[\sum_{y\neq Z_t} a_t(Z_t,y) \log \frac{a_t(Z_t,y)}{b_t(Z_t,y)}\Big]\, dt$, the holding factor contributes $\int_s^u \mathbb{E}_P\Big[\sum_{y\neq Z_t}(b_t(Z_t,y) - a_t(Z_t,y))\Big]\, dt$, and the initial factor contributes $\mathrm{KL}(p_s \,\|\, p'_s)$. Collecting through $\Phi(a,b) = b - a + a\log(a/b)$ yields (B.8). □

*Proof 2 of Theorem B.1 (Girsanov).* Start from (B.2). The reverse Markov property factorizes the model $z_0$-slice as $P^{\theta}_{0,[t_1,t_2]}(z_0, dz_{[t_1,t_2]}) = p^{\theta}_{0|t_1}(z_0 \mid z_{t_1})\, P^{\theta}_{[t_1,t_2]}(dz_{[t_1,t_2]})$. Substituting and applying the Radon–Nikodym chain rule,

$$-\mathrm{ELBO}_{[t_1,t_2]}(\theta;\, z_0) = -\mathbb{E}_{q_{t_1|0}(\cdot|z_0)}\big[\log p^{\theta}_{0|t_1}(z_0 \mid Z_{t_1})\big] + \mathrm{KL}\Big(P^{\star}_{[t_1,t_2]|0}(\cdot \mid z_0) \,\Big\|\, P^{\theta}_{[t_1,t_2]}\Big),$$

since the time-$t_1$ marginal of $P^{\star}_{[t_1,t_2]|0}(\cdot \mid z_0)$ is $q_{t_1|0}(\cdot \mid z_0)$. Apply Theorem B.2 to the backward descriptions of $P^{\star}_{[t_1,t_2]|0}(\cdot \mid z_0)$ and $P^{\theta}_{[t_1,t_2]}$, started at $q_{t_2|0}(\cdot \mid z_0) \ll p^{\theta}_{t_2}$ with rates $\widehat{Q}_t(z_t, \cdot \mid$

$z_0) \ll \widehat{Q}^\theta_t(z_t, \cdot)$, to expand the KL term into the terminal prior KL plus the path integral, which gives (B.3). For the gap identity, substitute $P^\theta_{0,[t_1,t_2]}(z_0, \cdot) = p^\theta_0(z_0)\, P^\theta_{[t_1,t_2]|0}(\cdot \mid z_0)$ into (B.2): the constant $\log p^\theta_0(z_0)$ splits off and the KL between the clean-conditioned and model posterior path laws remains. □

### B.3 Importance sampling, time clocks, and the empirical ELBO

**The empirical estimator.** The bound becomes a training objective by estimating its time integral, optionally with an importance density $p_{\mathrm{IS}}(t)$ (equivalently a clock, §3). Drawing $t \sim p_{\mathrm{IS}}$ on $[t_1, t_2]$ and $z_t \sim q_{t|0}(\cdot \mid z_0)$ gives the unbiased estimator

$$\widehat{\mathrm{NELBO}}_{[t_1,t_2]}(\theta;\, z_0) = \frac{\mathcal{J}^\theta_t(z_0, z_t)}{p_{\mathrm{IS}}(t)} + \mathrm{KL}\big(q_{t_2|0}(\cdot \mid z_0) \,\big\|\, p^\theta_{t_2}\big) + \widehat{\mathrm{rec}}(\theta), \tag{B.9}$$

where $\widehat{\mathrm{rec}}(\theta) = -\log p^\theta_{0|t_1}(z_0 \mid z_{t_1})$ at a draw $z_{t_1} \sim q_{t_1|0}(\cdot \mid z_0)$; this term accommodates any final decoding step from $t_1$ to 0, such as the standard collapse to clean tokens. The terminal-prior KL is computed analytically from the forward kernel and the chosen prior. Retaining both boundary terms keeps per-process estimates complete and comparable across methods.

Let $\tau$ be a strictly increasing $C^1$ change of time on $[t_1, t_2]$ with $d\tau/dt > 0$ and inverse $t(\tau)$, and let $Z'_\tau := Z_{t(\tau)}$, with marginals $q'_\tau = q_{t(\tau)}$ and kernels $q'_{\tau_2|\tau_1} = q_{t(\tau_2)|t(\tau_1)}$. By the chain rule in the definition of the generator, the generator and the reverse rate each carry one Jacobian factor,

$$Q'_\tau(x, y) = \frac{dt}{d\tau}\, Q_{t(\tau)}(x, y), \qquad \widehat{Q}'_\tau(z, y) = Q'_\tau(y, z)\, \frac{q'_\tau(y)}{q'_\tau(z)} = \frac{dt}{d\tau}\, \widehat{Q}_{t(\tau)}(z, y),$$

and likewise the clean-conditioned rate $\widehat{Q}'_\tau(z, y \mid z_0)$ and the model rate $\widehat{Q}^{\theta\prime}_\tau(z, y)$. Estimating the path integral with an importance density $p_{\mathrm{IS}}$ is uniform sampling on the clock

$$\tau(t) := \tau_1 + \int_{t_1}^{t} p_{\mathrm{IS}}(u)\, du, \qquad d\tau = p_{\mathrm{IS}}(t)\, dt, \qquad \frac{dt}{d\tau} = \frac{1}{p_{\mathrm{IS}}(t)}.$$

Since $\Phi(ca, cb) = c\,\Phi(a, b)$ for $c > 0$, the Jacobian moves from the measure into both rate arguments,

$$\begin{aligned} \int_{t_1}^{t_2} \mathbb{E}_{q_{t|0}}\big[\mathcal{J}^\theta_t(z_0, z_t)\big]\, dt &= \int_{\tau_1}^{\tau_2} \frac{dt}{d\tau}\, \mathbb{E}_{q_{t|0}}\left[\sum_{y \neq z_t} \Phi\big(\widehat{Q}_t(z_t, y \mid z_0), \widehat{Q}^\theta_t(z_t, y)\big)\right] d\tau \\ &= \int_{\tau_1}^{\tau_2} \mathbb{E}_{q_{t|0}}\left[\sum_{y \neq z_t} \Phi\big(\widehat{Q}'_\tau(z_t, y \mid z_0), \widehat{Q}^{\theta\prime}_\tau(z_t, y)\big)\right] d\tau, \end{aligned} \tag{B.10}$$

with $t = t(\tau)$ throughout. So (B.9) is an unbiased estimator of the ELBO of the same pair of processes read on the clock $\tau$, and the optimizer is the same. When the generator is separable, $Q_t = \frac{d\kappa}{dt} M$ for a strictly increasing scalar $\kappa$ and a fixed matrix $M$, the clock $\tau = \kappa(t)$ makes the forward process time-homogeneous with generator $M$.

Importance sampling leaves the estimator unbiased for the ELBO, whereas reweighting the per-time integrand without the $1/p_{\mathrm{IS}}$ correction changes the objective. Monotone reweightings define valid and possibly tighter variational bounds of their own (Kingma & Gao, 2023; Shi & Titsias, 2025), and are used to trade likelihood for sample quality.

**The intrinsic clock.** One clock is intrinsic to the process: the one on which it destroys information at a constant rate,

$$\tau(t) := \frac{H(Z_0 \mid Z_t)}{H(q_0)}, \qquad \text{so that} \qquad I(Z_0; Z_\tau) = (1 - \tau)\, H(q_0). \tag{B.11}$$

By Proposition 2 the oracle ELBO integrand is exactly this information-loss rate, so $\tau$ is the clock on which the oracle cost per unit time is flat, and the per-time ELBO profile of a model is a rough estimator of the intrinsic clock density that improves with training. Under the linear schedule $\alpha_t = 1 - t$ the masked oracle cost is already constant in $t$ for unigram data and approximately so for uniform

away from $t = 1$, but other kernels can destroy information following very different schedules, which makes $\tau$ a natural reference for a new kernel, though not necessarily the best clock for training.

For a single draw $t \sim p_{\mathrm{IS}}$ the estimator variance is minimized at $p^{\star}_{\mathrm{IS}}(t) \propto \sqrt{\mathbb{E}_{q_0\, q_{t|0}}\left[\mathcal{J}^{\theta}_t(z_0, z_t)^2\right]}$ (Robert & Casella, 2004). This second-moment rule underlies adaptive timestep sampling in DDPMs, while variational diffusion models learn the noise schedule directly to reduce the variance of the bound (Nichol & Dhariwal, 2021; Kingma et al., 2021). The intrinsic clock (B.11), $p_{\mathrm{IS}} \propto \mathcal{J}^{\star}_t$, equalizes the oracle cost per unit clock instead, and the two agree only where the root second moment of the integrand is proportional to its oracle mean; continuous-diffusion work adapts the training noise distribution from the same conditional-entropy-rate profile (Raya et al., 2026).

# C The Oracle Distance identity

We state the Oracle Distance in its finite-window form, prove it twice (by assembling the projection, Pythagorean, and information-loss identities, each of independent use, a route that needs the regularity of Assumption C.1, and by a direct path-space argument that does not), and derive Theorem 1 as its vanishing-boundary limit.

## C.1 The ELBO as a projection problem and the Pythagorean decomposition

The model rate cannot depend on the unknown clean data, so the path term of the ELBO poses a projection problem: given a random target rate depending on $(Z_0, Z_t)$, choose the best rate that sees only $Z_t$. Its solution is governed by a Pythagorean identity for $\Phi$.

**Lemma C.1** (Bregman–Pythagoras identity for $\Phi$)**.** *Let $A \geq 0$ with $\mathbb{E}[A \log A] < \infty$, let $G$ be an observable, and let $B = b(G) \geq 0$ be a $G$-measurable predictor for which the displayed expectations are finite. With $B^{\star} := \mathbb{E}[A \mid G]$,*

$$\mathbb{E}[\Phi(A, B)] = \mathbb{E}[\Phi(A, B^{\star})] + \mathbb{E}[\Phi(B^{\star}, B)].$$

*Proof.* Condition on $G$ and write $m = B^{\star}$, $b = B$. For $m, b > 0$, expanding $\Phi(a, b) = b - a + a \log \frac{a}{b}$ and using $\mathbb{E}[A \mid G] = m$ (with $0 \log 0 = 0$),

$$\begin{aligned}
\mathbb{E}[\Phi(A, b) \mid G] &= b - m + \mathbb{E}[A \log A \mid G] - m \log b, \\
\mathbb{E}[\Phi(A, m) \mid G] &= \mathbb{E}[A \log A \mid G] - m \log m, \\
\Phi(m, b) &= b - m + m \log m - m \log b.
\end{aligned}$$

Adding the last two reproduces the first, giving the conditional identity $\mathbb{E}[\Phi(A, B) \mid G] = \mathbb{E}[\Phi(A, B^{\star}) \mid G] + \Phi(B^{\star}, B)$; the boundary cases follow by continuity with the conventions in the definition of $\Phi$. Take expectations over $G$. □

Since $\mathbb{E}[\Phi(B^{\star}, B)] \geq 0$ vanishes iff $B = B^{\star}$ a.e., the conditional mean is the unique minimizer of $\mathbb{E}[\Phi(A, B)]$ over nonnegative $G$-measurable predictors. (The same follows from convexity: $b \mapsto \mathbb{E}[\Phi(A, b) \mid G = g]$ is strictly convex with derivative $1 - \mathbb{E}[A \mid G = g]\,/b$.)

**Corollary 1** (Reverse-rate projection)**.** *Among rates measurable in the noisy state $Z_t$ alone, the NELBO is minimized, uniquely up to null sets, by $\widehat{Q}^{\theta\,\star}_t(z_t, y) = \mathbb{E}[\widehat{Q}_t(Z_t, y \mid Z_0) \mid Z_t = z_t] = \widehat{Q}_t(z_t, y)$.*

*Proof.* Fix $t$ and a destination $y$, and apply Lemma C.1 with $G = Z_t$ and target $A = \widehat{Q}_t(Z_t, y \mid Z_0)$. The optimal $Z_t$-measurable rate is the conditional mean, computed by Bayes' rule:

$$\begin{aligned}
\mathbb{E}\left[\widehat{Q}_t(Z_t, y \mid Z_0) \mid Z_t = z_t\right] &= Q_t(y, z_t) \sum_{z_0} \frac{q_0(z_0)\, q_{t|0}(z_t \mid z_0)}{q_t(z_t)} \frac{q_{t|0}(y \mid z_0)}{q_{t|0}(z_t \mid z_0)} \\
&= Q_t(y, z_t) \frac{1}{q_t(z_t)} \sum_{z_0} q_0(z_0)\, q_{t|0}(y \mid z_0) \\
&= Q_t(y, z_t) \frac{q_t(y)}{q_t(z_t)} = \widehat{Q}_t(z_t, y).
\end{aligned}$$

Uniqueness is the equality case of the lemma. That the minimizer of the full NELBO is the same follows by summing over destinations and integrating over time, each term being minimized by the same rate. □

**Proposition C.1** (Pythagorean decomposition of the ELBO integrand)**.** *For any $Z_t$-measurable model rate $\widehat{Q}_t^\theta$ such that the model mismatch term below is finite,*

$$\mathbb{E}_{q_0\, q_{t|0}}\left[\mathcal{J}_t^\theta(z_0, z_t)\right] = \mathbb{E}_{q_0\, q_{t|0}}\left[\mathcal{J}_t^{\theta_\star}(z_0, z_t)\right] + \mathbb{E}_{q_t}\left[\sum_{y \neq z_t} \Phi\big(\widehat{Q}_t(z_t, y), \widehat{Q}_t^\theta(z_t, y)\big)\right], \tag{C.1}$$

*where $\mathcal{J}_t^{\theta_\star}$ is the integrand* (B.1) *evaluated at the marginal reverse rate $\widehat{Q}_t$.*

*Proof.* Fix $t$ and a destination $y$ and apply Lemma C.1 with $G = Z_t$, target $A = \widehat{Q}_t(Z_t, y \mid Z_0)$, and predictor $B = \widehat{Q}_t^\theta(Z_t, y)$; the conditional mean is $B^\star = \widehat{Q}_t(z_t, y)$ by the projection above. The mismatch expectation collapses to one over $z_t \sim q_t$ because both arguments are $Z_t$-measurable. Sum over $y \neq z_t$. □

### C.2 The oracle ELBO is the rate of information loss

The second ingredient evaluates the oracle cost $\mathcal{J}_t^\star = \mathbb{E}_{q_0\, q_{t|0}}\left[\mathcal{J}_t^{\theta_\star}(z_0, z_t)\right]$ under the following regularity, which the masked, uniform, and fixed-support GIDD processes satisfy on interior windows.

**Assumption C.1** (Regularity and fixed support)**.** *On the compact window $[t_1, t_2] \subset (0, T)$, the rates $Q_t$ are piecewise continuous and bounded, and the piecewise supports of $q_t$ and $q_{t|0}(\cdot \mid z_0)$ are independent of $t$ for each $z_0 \in \operatorname{supp} q_0$.*

**Proposition 2** (Irreducible cost is information loss)**.** *Under Assumption C.1, $\mathcal{J}_t^\star = \frac{d}{dt} H(Z_0 \mid Z_t) = -\frac{d}{dt} I(Z_0; Z_t) \geq 0$, and $\int_{t_1}^{t_2} \mathcal{J}_t^\star \, dt = H(Z_0 \mid Z_{t_2}) - H(Z_0 \mid Z_{t_1}) = I(Z_0; Z_{t_1}) - I(Z_0; Z_{t_2})$.*

*Proof.* Terms with zero clean-conditioned rate contribute zero under $0 \log 0 = 0$; on positive terms the fixed-support assumption makes all ratios below well defined. Inside $\Phi(\widehat{Q}_t(\cdot \mid Z_0), \widehat{Q}_t)$ the rate quotient is, by the two reverse-rate formulas and Bayes' rule,

$$\frac{\widehat{Q}_t(z, y \mid z_0)}{\widehat{Q}_t(z, y)} = \frac{q_{t|0}(y \mid z_0)}{q_{t|0}(z \mid z_0)} \frac{q_t(z)}{q_t(y)} = \frac{q_{0|t}(z_0 \mid y)}{q_{0|t}(z_0 \mid z)}.$$

In the expansion $\Phi(a, b) = b - a + a \log(a/b)$ the linear terms cancel after conditioning on $Z_t$, because $\widehat{Q}_t(z, y) = \mathbb{E}[\widehat{Q}_t(z, y \mid Z_0) \mid Z_t = z]$. Hence, using $q_{t|0}(z \mid z_0)\, \widehat{Q}_t(z, y \mid z_0) = q_{t|0}(y \mid z_0)\, Q_t(y, z)$ and adding the diagonal generator terms, whose logarithms vanish,

$$\mathcal{J}_t^\star = \sum_{z_0, y, z} q_0(z_0)\, q_{t|0}(y \mid z_0)\, Q_t(y, z) \log \frac{q_{0|t}(z_0 \mid y)}{q_{0|t}(z_0 \mid z)}. \tag{C.2}$$

Conversely, differentiate $H(Z_0 \mid Z_t) = -\sum_{z_0, z} q_0(z_0)\, q_{t|0}(z \mid z_0) \log q_{0|t}(z_0 \mid z)$. Differentiation passes through the finite sum; the derivative of the log terms cancels by normalization, $\sum_{z_0, z} q_0(z_0)\, q_{t|0}(z \mid z_0)\, \partial_t \log q_{0|t}(z_0 \mid z) = \sum_z q_t(z)\, \partial_t \sum_{z_0} q_{0|t}(z_0 \mid z) = 0$, so inserting the Kolmogorov forward equation $\partial_t q_{t|0}(z \mid z_0) = \sum_y q_{t|0}(y \mid z_0) Q_t(y, z)$ and adding the vanishing term $\sum_{z_0, y, z} q_0(z_0) q_{t|0}(y \mid z_0) Q_t(y, z) \log q_{0|t}(z_0 \mid y) = 0$ (zero because $\sum_z Q_t(y, z) = 0$) gives

$$\frac{d}{dt} H(Z_0 \mid Z_t) = \sum_{z_0, y, z} q_0(z_0)\, q_{t|0}(y \mid z_0)\, Q_t(y, z) \log \frac{q_{0|t}(z_0 \mid y)}{q_{0|t}(z_0 \mid z)},$$

which is (C.2). Nonnegativity is $\Phi \geq 0$; the fixed-support assumption makes $H(Z_0 \mid Z_t)$ absolutely continuous on the window, so integration gives the cumulative form. □

**Corollary C.1** (Monotonicity of retained information)**.** *The map $t \mapsto I(Z_0; Z_t)$ is nonincreasing, and wherever it is differentiable, $-\frac{d}{dt} I(Z_0; Z_t) = \mathcal{J}_t^\star \geq 0$.*

*Proof.* For $s < t$ the Markov chain $Z_0 \to Z_s \to Z_t$ and the data-processing inequality give $I(Z_0; Z_t) \leq I(Z_0; Z_s)$; the derivative statement is Proposition 2. □

Figure 4 shows both halves of the identity on the solvable model of App. F: the common ceiling $H(q_0)$ reached along process-dependent paths, and the rate $\mathcal{J}_t^\star$ at which each process approaches it.

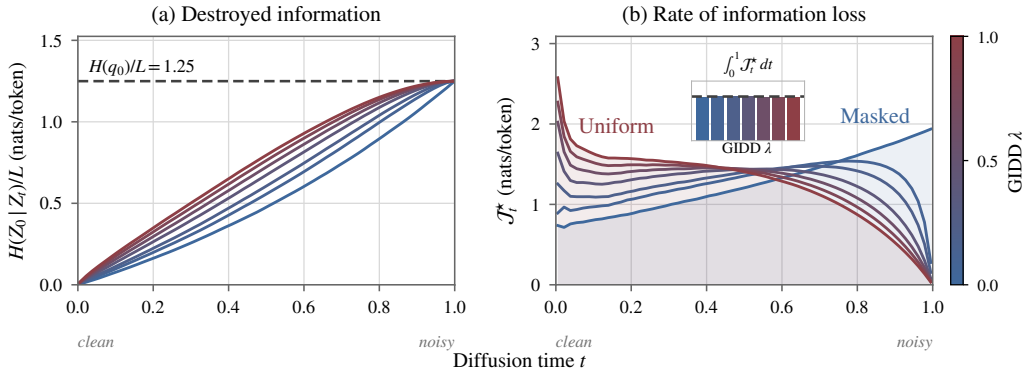


Figure 4: Every noising process destroys $H(q_0)$ information at a possibly different rate, computed at the exact oracle across the GIDD family. **(a)** Information destroyed by time $t$, $H(Z_0 \mid Z_t)/L$: different paths to the common ceiling $H(q_0)/L$. **(b)** Its rate $\mathcal{J}_t^\star$ (Proposition 2): uniform destroys information earlier than masked, with equal area (inset).

### C.3 ELBO AS DISTANCE TO THE ORACLE REVERSE PROCESS

**Theorem C.1** (Oracle Distance, finite window)**.** *Let $[t_1, t_2] \subset (0, T)$ and let $\widehat{Q}_t$ and $\widehat{Q}_t^\theta$ be bounded and measurable. Assume the support conditions $q_{0|t_1}(\cdot \mid z_{t_1}) \ll p^\theta_{0|t_1}(\cdot \mid z_{t_1})$, $q_{t_2} \ll p^\theta_{t_2}$, and $\widehat{Q}_t(z_t, \cdot) \ll \widehat{Q}_t^\theta(z_t, \cdot)$ a.e., and that the time integral in* (C.4) *is finite. Then*

$$\mathrm{NELBO}_{[t_1,t_2]}(\theta) - H(q_0) = \mathbb{E}_{q_{t_1}}\left[\mathrm{KL}\big(q_{0|t_1}(\cdot \mid z_{t_1}) \,\big\|\, p^\theta_{0|t_1}(\cdot \mid z_{t_1})\big)\right] + \mathrm{KL}\Big(P^\star_{[t_1,t_2]} \,\Big\|\, P^\theta_{[t_1,t_2]}\Big), \tag{C.3}$$

*where*

$$\mathrm{KL}\Big(P^\star_{[t_1,t_2]} \,\Big\|\, P^\theta_{[t_1,t_2]}\Big) = \mathrm{KL}\big(q_{t_2} \,\big\|\, p^\theta_{t_2}\big) + \int_{t_1}^{t_2} \mathbb{E}_{q_t}\left[\sum_{y \neq z_t} \Phi\big(\widehat{Q}_t(z_t, y), \widehat{Q}_t^\theta(z_t, y)\big)\right] dt. \tag{C.4}$$

*Assembly proof of Theorem C.1.* Averaging the negative of (B.3) over the data, $Z_0 \sim q_0$, writes $\mathrm{NELBO}_{[t_1,t_2]}(\theta)$ as reconstruction, path, and terminal-prior terms, treated in turn. *Reconstruction*: subtracting and adding $H(Z_0 \mid Z_{t_1}) = \mathbb{E}_{Z_0, Z_{t_1}}[-\log q_{0|t_1}(Z_0 \mid Z_{t_1})]$ gives

$$\mathbb{E}_{Z_0,Z_{t_1}}\left[-\log p^\theta_{0|t_1}(Z_0 \mid Z_{t_1})\right] = H(Z_0 \mid Z_{t_1}) + \mathbb{E}_{Z_{t_1}}\left[\mathrm{KL}\big(q_{0|t_1}(\cdot \mid Z_{t_1}) \,\big\|\, p^\theta_{0|t_1}(\cdot \mid Z_{t_1})\big)\right].$$

*Path*: Proposition C.1 splits the integrand into oracle plus mismatch, and Proposition 2 integrates the oracle part into $H(Z_0 \mid Z_{t_2}) - H(Z_0 \mid Z_{t_1})$. *Terminal prior*: splitting the log ratio through the marginal $q_{t_2}$,

$$\begin{aligned}\mathbb{E}_{Z_0}\left[\mathrm{KL}\big(q_{t_2|0}(\cdot \mid Z_0) \,\big\|\, p^\theta_{t_2}\big)\right] &= I(Z_0; Z_{t_2}) + \mathrm{KL}\big(q_{t_2} \,\big\|\, p^\theta_{t_2}\big) \\ &= H(Z_0) - H(Z_0 \mid Z_{t_2}) + \mathrm{KL}\big(q_{t_2} \,\big\|\, p^\theta_{t_2}\big).\end{aligned}$$

Summing the three, the conditional entropies telescope into $H(Z_0) = H(q_0)$, leaving the reconstruction KL, the terminal KL, and the integrated mismatch, which is (C.3) with the path KL expanded by Theorem B.2. □

*Proof of Theorem 1.* In Theorem C.1 let $t_1 \downarrow 0$ and $t_2 \uparrow T$: the two boundary mismatches, the reconstruction KL and the terminal KL inside (C.4), vanish by hypothesis, and the path integrals converge to the integral in (4) by monotone convergence, leaving the identity. □

Figure 5 verifies the identity and its per-time Pythagorean split on the solvable model of App. F.

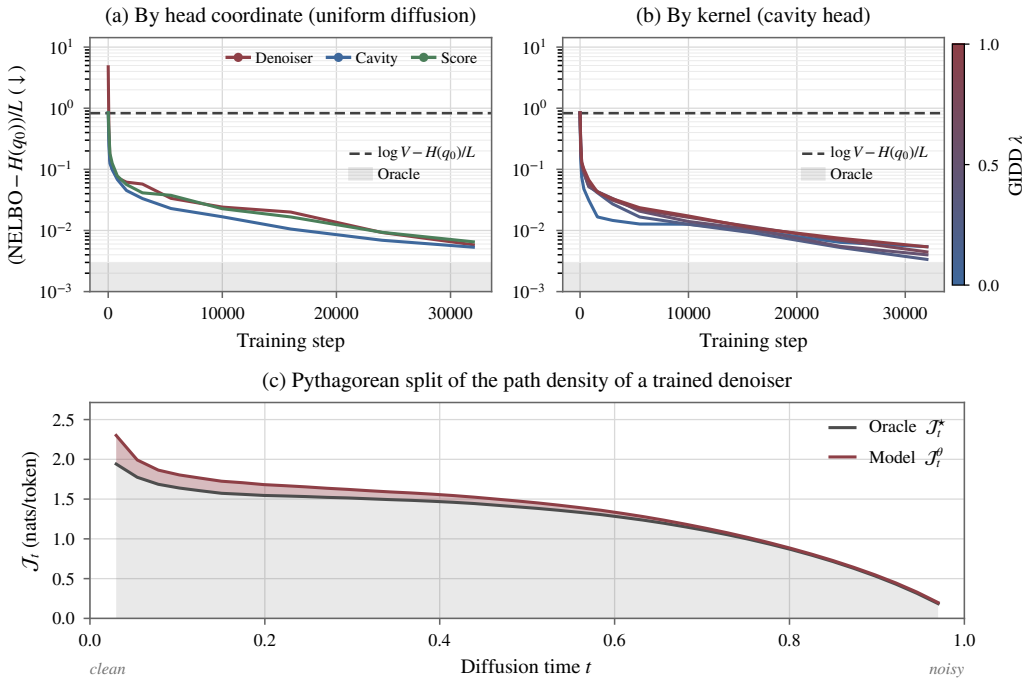


Figure 5: The NELBO is an exact distance to the oracle (Theorem 1). **(a, b)** Per-token excess $\big(\mathrm{NELBO}(\theta) - H(q_0)\big)/L$ over training falling to the oracle's own value on this window (shaded lower band), which is not zero but the residual boundary terms of the clipped window, for every head coordinate (a) and every noising process (b); the dashed line is the initialization value $\log V - H(q_0)/L$ of Proposition 4, where the cavity and score heads start. **(c)** The Pythagorean split (7) of the path density over diffusion time: path density = oracle rate $\mathcal{J}_t^\star$ + model–oracle divergence, here for a denoiser head trained for 300 steps.

### C.4 The universal attainable NELBO floor

**Lemma C.2** (NELBO floor)**.** *For every noising process and model satisfying the conditions of Theorem B.1,*

$$\mathrm{NELBO}_{[t_1,t_2]}(\theta) \geq \mathbb{E}_{q_0}\big[-\log p_0^\theta(z_0)\big] \geq H(q_0). \tag{C.5}$$

*Proof.* The first inequality is the ELBO gap of Theorem B.1; the second is Gibbs' inequality, $\mathbb{E}_{q_0}\big[-\log p_0^\theta\big] = H(q_0) + \mathrm{KL}\big(q_0 \,\big\|\, p_0^\theta\big)$. □

**Corollary C.2** (Universal floor, finite window)**.** *In the setting of Theorem C.1, if some $\theta_\star$ realizes $p^{\theta_\star}_{t_2} = q_{t_2}$, $\widehat{Q}_t^{\theta_\star} = \widehat{Q}_t$ on $[t_1, t_2]$, and $p^{\theta_\star}_{0|t_1} = q_{0|t_1}$, then $p_0^{\theta_\star} = q_0$, the ELBO is tight at $\theta_\star$, and* $\inf_\theta \mathrm{NELBO}_{[t_1,t_2]}(\theta) = \mathrm{NELBO}_{[t_1,t_2]}(\theta_\star) = H(q_0)$.

*Proof.* Evaluate Theorem C.1 at $\theta_\star$: the reconstruction term vanishes because $p^{\theta_\star}_{0|t_1} = q_{0|t_1}$, and the path KL vanishes because $P^\theta_{[t_1,t_2]} = P^\star_{[t_1,t_2]}$, both initialized at $q_{t_2}$ and run with $\widehat{Q}_t$. Hence $\mathrm{NELBO}_{[t_1,t_2]}(\theta_\star) = H(q_0)$, which with Lemma C.2 gives the infimum. The lemma also writes $\mathrm{NELBO}_{[t_1,t_2]}(\theta_\star) - H(q_0)$ as the sum of the expected ELBO gap and $\mathrm{KL}\Big(q_0 \,\Big\|\, p_0^{\theta_\star}\Big)$; both vanish, so $p_0^{\theta_\star} = q_0$ and the ELBO is tight $q_0$-a.e. Under Assumption A.1 the conclusion passes to $t_2 \uparrow T$, entropy and mutual information being continuous and bounded on the finite space, and with vanishing boundary mismatches this is Corollary 2. □

### C.5 A direct proof of the Oracle Distance identity

The assembly proof uses Assumption C.1, which the theorem itself does not need. The following proof avoids it but works with laws on the path space: the KL chain rule plays the role that the

Pythagorean split played at the rate level, and since $\Phi$ is the KL rate between Poisson intensities, Proposition C.1 may be read as the infinitesimal, jump-by-jump version of the chain rule.

*Direct proof of Theorem C.1.* The support hypotheses make every term below well defined: $P^\star_{[t_1,t_2]} \ll P^\theta_{[t_1,t_2]}$ by Theorem B.2, and with $q_{0|t_1} \ll p^\theta_{0|t_1}$ this gives $P^\star_{0,[t_1,t_2]} \ll P^\theta_{0,[t_1,t_2]}$.

Negating (B.2) and averaging over $z_0 \sim q_0$ expresses the average NELBO through the joint factorization of Definition A.1:

$$\mathrm{NELBO}_{[t_1,t_2]}(\theta) = \int \log \frac{dP^\star_{[t_1,t_2]|0}(\cdot \mid z_0)}{dP^\theta_{0,[t_1,t_2]}(z_0,\cdot)}(z_{[t_1,t_2]})\; P^\star_{0,[t_1,t_2]}(z_0, dz_{[t_1,t_2]}).$$

Since $q_0$ is the $z_0$-marginal of $P^\star_{0,[t_1,t_2]}$, subtracting $H(q_0)$ inserts the factor $q_0(z_0)$ into the numerator, and the integrand becomes the joint Radon–Nikodym derivative:

$$\mathrm{NELBO}_{[t_1,t_2]}(\theta) - H(q_0) = \mathrm{KL}\Big(P^\star_{0,[t_1,t_2]} \,\Big\|\, P^\theta_{0,[t_1,t_2]}\Big)\,.$$

By the Markov property, $\mathrm{Law}(Z_0 \mid Z_{[t_1,t_2]}) = q_{0|t_1}(\cdot \mid Z_{t_1})$, so the joint noising law also factors backward,

$$P^\star_{0,[t_1,t_2]}(z_0, dz_{[t_1,t_2]}) = P^\star_{[t_1,t_2]}(dz_{[t_1,t_2]})\, q_{0|t_1}(z_0 \mid z_{t_1}). \tag{C.6}$$

With both joint laws factored in the same order, the Radon–Nikodym derivative factorizes, and taking logs and integrating against $P^\star_{0,[t_1,t_2]}$ splits the joint divergence into a window term and an endpoint-kernel term:

$$\mathrm{KL}\Big(P^\star_{0,[t_1,t_2]} \,\Big\|\, P^\theta_{0,[t_1,t_2]}\Big) = \mathrm{KL}\Big(P^\star_{[t_1,t_2]} \,\Big\|\, P^\theta_{[t_1,t_2]}\Big) + \mathbb{E}_{q_{t_1}}\left[\mathrm{KL}\big(q_{0|t_1}(\cdot \mid z_{t_1}) \,\big\|\, p^\theta_{0|t_1}(\cdot \mid z_{t_1})\big)\right],$$

which is (C.3); Theorem B.2 expands the path KL into (C.4). □

### C.6 The arrow of time and independent mechanisms

At the path-law level there is no asymmetry between the two time directions. In the forward direction the description factorizes into two independent mechanisms, an initial condition (the data law $q_0$) and a dynamical law (the generator $Q_t$), chosen without knowledge of each other. Backward, this independence is destroyed: the reverse generator $\widehat{Q}_t(x,y) = Q_t(y,x)\, q_t(y)/q_t(x)$ requires the marginals and hence the data law. Janzing et al. (2016) derive the thermodynamic arrow of time from precisely this independence of initial state and dynamics, and the same asymmetry-through-restriction underlies causal direction detection in time series (Peters et al., 2009; 2013); see also Schölkopf et al. (2012). Discrete diffusion is an instance: the forward process belongs to the small class of data-independent, position-factorized mechanisms, whereas its reversal, whenever $q_0$ is correlated across positions, exists only in the larger class of data-dependent generators, and the purpose of training is to supply that dependence. The framework quantifies the information the reverse mechanism must restore: the forward mechanism destroys $I(Z_0; Z_t)$ at the rate $\mathcal{J}^\star_t = -\frac{d}{dt} I(Z_0; Z_t)$ of Proposition 2, the total amount destroyed is conserved, the floor $H(q_0)$ of Corollary 2, and the choice of noising process controls only when along the trajectory it is lost.

### C.7 A Bregman-divergence and generator-matching view

These results are instances of standard Bregman identities. A strictly convex differentiable potential $F$ generates $D_F(a,b) = F(a) - F(b) - F'(b)(a-b)$; the divergence $\Phi$ is generated by $F(u) = u \log u - u$. For any $D_F$, the conditional mean minimizes $\mathbb{E}[D_F(A, b(G))]$ over $G$-measurable predictors and satisfies the Pythagorean identity (Banerjee et al., 2005a); Lemma C.1 specializes this to $\Phi$, and the optimal rate is the marginal generator obtained by posterior-averaging the clean-conditioned generators, as in Generator Matching (Holderrieth et al., 2025). It is also the jump-process analogue of flow matching (Gat et al., 2024): the regressed object is a reverse intensity and the divergence is $\Phi$ rather than the squared norm. The correspondence with continuous diffusion is exact.

*Remark* C.1 (The continuous-diffusion analogue). Let the noising process be a Markov diffusion $dZ_t = f_t(Z_t)\, dt + \sigma_t(Z_t)\, dW_t$ on $\mathbb{R}^d$ with $Z_0 \sim q_0$ and $a_t := \sigma_t \sigma_t^\mathsf{T}$. Under standard smoothness and nondegeneracy assumptions the reverse-time process is again a diffusion with the same diffusion

matrix and, at forward time $t$, drift $\widehat{b}_t(z) = -f_t(z) + \nabla \cdot a_t(z) + a_t(z)\,\nabla \log q_t(z)$; conditioning on the data replaces the marginal by the forward kernel. The score $\nabla \log q_t$ plays the role of $\widehat{Q}_t$ and $\nabla \log q_{t|0}$ that of the computable clean-conditioned target. For reverse processes sharing the diffusion matrix, Girsanov's theorem gives

$$\mathrm{KL}\Big(P^{\star}_{[0,T]} \,\Big\|\, P^{\theta}_{[0,T]}\Big) = \mathrm{KL}\big(q_T \,\big\|\, p^{\theta}_T\big) + \frac{1}{2}\int_0^T \mathbb{E}_{q_t}\Big[\Big(\big(\widehat{b}_t - \widehat{b}^{\theta}_t\big)^{\mathsf{T}} a_t^{-1}\,\big(\widehat{b}_t - \widehat{b}^{\theta}_t\big)\Big)(Z_t)\Big]\,dt,$$

the projection becomes a weighted $L^2$ projection whose optimizer among $Z_t$-measurable drifts is the conditional mean, the drift form of Fisher's identity $\nabla \log q_t(z) = \mathbb{E}[\nabla \log q_{t|0}(z \mid Z_0) \mid Z_t = z]$, and with matched endpoints

$$\mathrm{NELBO}_{[0,T]}(\theta) = H(q_0) + \frac{1}{2}\int_0^T \mathbb{E}_{q_t}\Big[\big\|\sigma_t^{-1}\big(\widehat{b}_t - \widehat{b}^{\theta}_t\big)(Z_t)\big\|^2\Big]\,dt,$$

with $H(q_0)$ the differential entropy. Training on a window with a terminal prior and a reconstruction decoder gives the exact analogue of Theorem 1, the terminal mismatch again sitting inside the path KL. Table 4 collects the correspondence.

Table 4: The discrete/continuous correspondence: the same path structure throughout, with jump rates and the rate divergence $\Phi$ replaced by reverse drifts and the quadratic Girsanov energy.

| | Finite-state CTMC | Continuous diffusion |
|---|---|---|
| Forward infinitesimal object | generator $Q_t(x, y)$ | generator $f_t \cdot \nabla + \frac{1}{2} a_t : \nabla^2$ |
| True reverse object | reverse rate $\widehat{Q}_t(z_t, y)$ | reverse drift $\widehat{b}_t(z)$, or score $\nabla \log q_t(z)$ |
| Tractable conditional target | $\widehat{Q}_t(z_t, y \mid z_0)$ | $\widehat{b}_t(z \mid z_0)$, or $\nabla \log q_{t\|0}(z \mid z_0)$ |
| Path-KL divergence | $\Phi(a, b) = a \log \frac{a}{b} - a + b$ | $\frac{1}{2}(u - v)^{\mathsf{T}} a_t^{-1}(u - v)$ |
| Path-KL formula | Theorem B.2 | Girsanov's theorem |
| Population optimizer | $\mathbb{E}[\widehat{Q}_t(z_t, y \mid Z_0) \mid Z_t = z_t]$ | $\mathbb{E}\big[\widehat{b}_t(z \mid Z_0) \mid Z_t = z\big]$ |
| NELBO meaning | data entropy + reverse path KL, plus the two boundary mismatches | |

# D Sequence Modeling

## D.1 Token-factorizing noising processes

In this section $Q$ denotes sequence-level rates and $R$ single-token rates. Forward noising factors over positions; correlations in the clean-data law mean that the unconditional marginals $q_t$ do not, and only the clean-conditioned kernels factor.

**Theorem D.1** (Product CTMCs)**.** *Suppose position $1 \le i \le L$ has generator $R^i_t$ on $\mathcal{X}_i$ with transition kernel $q^i_{t|s}$, and define the product kernel $q_{t|s}(y \mid x) := \prod_{i=1}^L q^i_{t|s}(y^i \mid x^i)$ on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_L$. Then its unique infinitesimal generator is*

$$Q_t(x, y) = \sum_{i=1}^{L} (Q_t)_i(x, y), \qquad (Q_t)_i(x, y) = \mathbf{1}\{x^{-i} = y^{-i}\}\, R^i_t(x^i, y^i).$$

*Conversely, any process with a generator of this form has the product transition kernel, that is, it factors over positions.*

*Proof.* First, $Q_t$ is a generator: off-diagonal entries are nonnegative, and

$$\sum_{y \in \mathcal{X}} Q_t(x, y) = \sum_{i=1}^{L} \sum_{y \in \mathcal{X}} \mathbf{1}\{x^{-i} = y^{-i}\}\, R^i_t(x^i, y^i) = \sum_{i=1}^{L} \sum_{u \in \mathcal{X}_i} R^i_t(x^i, u) = 0.$$

Second, the product kernel satisfies the forward Kolmogorov equation for $Q_t$: since each position's kernel satisfies its own, $\partial_t q^i_{t|s}(y^i \mid x^i) = \sum_{u \in \mathcal{X}_i} q^i_{t|s}(u \mid x^i) R^i_t(u, y^i)$, differentiating the product

gives

$$\partial_t q_{t|s}(y \mid x) = \sum_{i=1}^{L} \sum_{u \in \mathcal{X}_i} q^i_{t|s}(u \mid x^i) R^i_t(u, y^i) \prod_{j \neq i} q^j_{t|s}(y^j \mid x^j),$$

and expanding $\sum_z q_{t|s}(z \mid x) Q_t(z, y)$ with the definition of $Q_t$ produces the same expression. The initial condition $q_{s|s}(y \mid x) = \mathbf{1}\{x = y\}$ holds, so by uniqueness of solutions of linear ODEs the product kernel is the kernel of $Q_t$. For the converse, a process with a generator of the displayed form solves the same forward equation, whose solution is unique and equal to the product kernel. □

The clean-conditioned reverse process still factors: for $s < t$, Bayes' rule gives

$$q_{s|t,0}(z_s \mid z_t, z_0) = \frac{q_{s|0}(z_s \mid z_0)\, q_{t|s}(z_t \mid z_s)}{q_{t|0}(z_t \mid z_0)} = \prod_{i=1}^{L} q^i_{s|t,0}(z^i_s \mid z^i_t, z^i_0).$$

Applying Theorem D.1 to this factorized backward process gives the token-wise clean-conditioned reverse rates: for a one-token jump $z_t \to y$ with $y^i \neq z^i_t$,

$$a_i(t, z_0, z_t, y) := (\widehat{Q}_t)_i(z_t, y \mid z_0) = \mathbf{1}\{z_t^{-i} = y^{-i}\}\, R^i_t(y^i, z^i_t)\, \frac{q^i_{t|0}(y^i \mid z^i_0)}{q^i_{t|0}(z^i_t \mid z^i_0)}, \tag{D.1}$$

the target rates inside the token-wise ELBO.

### D.2 Three coordinates of the reverse rate

Fix a position $i$ and write $e_\ell$ for the vocabulary elements. Bayes' rule expands the denoiser (10) so as to isolate the Bayes factor from the rest of the sequence:

$$\pi^\star_i(e_\ell \mid z_t, t) = \frac{q^i_{t|0}(z^i_t \mid e_\ell) \sum_{z_0^{-i}} q_0(e_\ell, z_0^{-i}) \prod_{j \neq i} q^j_{t|0}(z^j_t \mid z^j_0)}{\sum_r q^i_{t|0}(z^i_t \mid e_r) \sum_{z_0^{-i}} q_0(e_r, z_0^{-i}) \prod_{j \neq i} q^j_{t|0}(z^j_t \mid z^j_0)}, \tag{D.2}$$

where $(e_\ell, z_0^{-i})$ is the sequence with $i$-th token $e_\ell$. The common inner sums are the cavity law: expanding $\mu^\star_i$ (10) the same way,

$$\mu^\star_i(e_\ell \mid z_t^{-i}, t) = \frac{\sum_{z_0^{-i}} q_0(e_\ell, z_0^{-i}) \prod_{j \neq i} q^j_{t|0}(z^j_t \mid z^j_0)}{\sum_r \sum_{z_0^{-i}} q_0(e_r, z_0^{-i}) \prod_{j \neq i} q^j_{t|0}(z^j_t \mid z^j_0)},$$

so the denoiser is the cavity reweighted by the local forward kernel,

$$\pi^\star_i(e_\ell \mid z_t, t) \propto \mu^\star_i(e_\ell \mid z_t^{-i}, t)\, q^i_{t|0}(z^i_t \mid e_\ell) : \tag{D.3}$$

the local kernel is exactly the Bayes factor for observing $z^i_t$ on top of the context. Equivalently, for any $f$,

$$\mathbb{E}_{\pi^\star_i(e_\ell \mid z_t, t)}[f(e_\ell)] = \frac{\mathbb{E}_{\mu^\star_i(e_\ell \mid z_t^{-i}, t)}\big[f(e_\ell)\, q^i_{t|0}(z^i_t \mid e_\ell)\big]}{\mathbb{E}_{\mu^\star_i(e_\ell \mid z_t^{-i}, t)}\big[q^i_{t|0}(z^i_t \mid e_\ell)\big]}, \tag{D.4}$$

and the denominator is the local noisy-token likelihood given the context:

$$\mathbb{E}_{\mu^\star_i(e_\ell \mid z_t^{-i}, t)}\big[q^i_{t|0}(w \mid e_\ell)\big] = q(Z^i_t = w \mid z_t^{-i}). \tag{D.5}$$

*Remark* D.1 (Support of the denoiser law). Here and throughout, terms with both $\pi^\star_i(e_\ell \mid z_t, t) = 0$ and $q^i_{t|0}(z^i_t \mid e_\ell) = 0$ contribute zero. Defining the reverse support $\mathcal{S}_i(z^i_t, t) := \{\ell : q^i_{t|0}(z^i_t \mid e_\ell) > 0\}$, Bayes' rule gives $\operatorname{supp} \pi^\star_i(\cdot \mid z_t, t) \subseteq \mathcal{S}_i(z^i_t, t)$ at every reachable state.

**Theorem 2** (Three coordinates)**.** *For a token-factorizing noising process the NELBO is uniquely minimized by the marginal reverse rate, whose posterior average on a one-token jump has the three equal forms of* (11)*.*

*Proof of Theorem 2.* Uniqueness of the optimizer is Corollary 1. For the three forms, average the clean-conditioned rate (D.1) over $q(z_0 \mid z_t)$: the product kernels cancel at every position except $i$, leaving

$$(\widehat{Q}_t)_i(z_t, y) = \mathbf{1}\{z_t^{-i} = y^{-i}\}\, R^i_t(y^i, z^i_t) \sum_\ell \pi^\star_i(e_\ell \mid z_t, t)\, \frac{q^i_{t|0}(y^i \mid e_\ell)}{q^i_{t|0}(z^i_t \mid e_\ell)},$$

the denoiser form of (11). Applying (D.4) with $f(e_\ell) = q^i_{t|0}(y^i \mid e_\ell)/q^i_{t|0}(z^i_t \mid e_\ell)$ turns it into the cavity form, and (D.5) identifies the ratio of cavity-averaged kernels with the ratio of noisy conditionals, the score form. □

**Proposition 3** (Coordinate conversions)**.** *The three heads are related by* (12) *and* (13).

*Proof of Proposition 3.* The cavity-to-denoiser map is the Bayes reweighting (D.3); inverting it gives $\mu^\theta_i \propto \pi^\theta_i / q^i_{t|0}(z^i_t \mid e_\ell)$ wherever that kernel is positive. By the definition of the score and (D.5), the score is the ratio $m^\theta_i(y^i)/m^\theta_i(z^i_t)$ of the cavity-induced marginal $m^\theta_i = \mu^\theta_i\, q^i_{t|0}$, so either distribution head determines it. Conversely a score fixes $m^\theta_i$ up to scale, recovered as a probability vector by normalizing $s^\theta_i(z^i_t, \cdot)$, and $\mu^\theta_i$ solves the linear system $m^\theta_i = \mu^\theta_i\, q^i_{t|0}$, which has a solution iff $m^\theta_i$ lies in the image of the channel, unique iff the channel has full row rank. For the GIDD kernel, $q^i_{t|0} = \alpha_t I + \beta_t \mathbf{1}\omega_t$ decouples across $\ell$ (the noise floor is source-independent), giving the affine inverse $\mu^\theta_i(e_\ell) = \big(m^\theta_i(e_\ell) - \beta_t\omega_t(e_\ell)\big)/\alpha_t$, with full row rank for $\alpha_t > 0$. □

The maps into the score are always available; recovering a distribution coordinate from a score needs an invertible channel and a nonnegative solution. Where $q^i_{t|0}(z^i_t \mid e_\ell)$ vanishes for some $\ell$, the visible states of masked diffusion, the denoiser cannot recover the corresponding cavity mass. The inversion is the discrete analogue of empirical-Bayes deconvolution: the prior (cavity) is recovered by inverting the noising channel. Table 5 collects the generic maps; Table 6 instantiates them for GIDD.

**The three parameterizations, their rates, and their ancestral steps.** Substituting a head for its law in Theorem 2 defines the three model rates; each determines the reverse CTMC in full through the Kolmogorov equation. The finite-step samplers share one template, the token-factorized ansatz $p^\theta_{s|t}(z_s \mid z_t) := \prod_{i=1}^L p^\theta(z^i_s \mid z_t)$, whose factors are the exact per-position marginals $q(z^i_s \mid z_t)$ at the oracle laws and which is approximate only through the forced product over positions: the per-position posteriors are frozen within the step, so simultaneous jumps are sampled independently, an $O((t-s)^2)$ effect.

*Denoiser* (Austin et al., 2021; Hoogeboom et al., 2021; Campbell et al., 2022; Sahoo et al., 2024; Shi et al., 2024). A head $\pi^\theta_i$ gives the rate

$$a^\theta_{i,\mathrm{den}}(t, z_t, y) := \mathbf{1}\{z_t^{-i} = y^{-i}\}\, R^i_t(y^i, z^i_t) \sum_\ell \pi^\theta_i(e_\ell \mid z_t, t)\, \frac{q^i_{t|0}(y^i \mid e_\ell)}{q^i_{t|0}(z^i_t \mid e_\ell)}. \tag{D.6}$$

Bayes' rule on the single-token bridge gives the exact marginal $q(z^i_s{=}y \mid z_t) = q^i_{t|s}(z^i_t \mid y) \sum_\ell \pi^\star_i(e_\ell \mid z_t, t)\, q^i_{s|0}(y \mid e_\ell)/q^i_{t|0}(z^i_t \mid e_\ell)$, hence the finite-step sampler

$$p^\theta_{\mathrm{den}}(Z^i_s = y \mid z_t) := q^i_{t|s}(z^i_t \mid y) \sum_\ell \pi^\theta_i(e_\ell \mid z_t, t)\, \frac{q^i_{s|0}(y \mid e_\ell)}{q^i_{t|0}(z^i_t \mid e_\ell)}, \tag{D.7}$$

the marginalization form. For a learned head the inclusion of Remark D.1 is not guaranteed: a denoiser head must learn to vanish off $\mathcal{S}_i(z^i_t, t)$, or be masked to it, to keep the rate finite. In masked diffusion the restriction is automatic (the support at a masked position is the whole vocabulary, and a revealed position admits no reverse jump), while in uniform diffusion the support collapses only at $t = 0$, so a denoiser head must concentrate on the true token exactly as $t \downarrow 0$, which is where an uninformative one diverges (Proposition 4).

*Cavity* (Sun et al., 2023; von Rütte et al., 2025; Schiff et al., 2025; Sahoo et al., 2025). A head $\mu^\theta_i$ induces the noisy laws at times $t$ and $s$,

$$\begin{aligned} m^\theta_i(y \mid z_t^{-i}, t) &:= \sum_\ell \mu^\theta_i(e_\ell \mid z_t^{-i}, t)\, q^i_{t|0}(y \mid e_\ell), \\ m^\theta_{i,t\to s}(y \mid z_t^{-i}, t) &:= \sum_\ell \mu^\theta_i(e_\ell \mid z_t^{-i}, t)\, q^i_{s|0}(y \mid e_\ell), \end{aligned} \tag{D.8}$$

hence the rate and the plug-in step

$$a^\theta_{i,\mathrm{cav}}(t, z_t, y) := \mathbf{1}\{z_t^{-i} = y^{-i}\}\, R^i_t(y^i, z^i_t)\, \frac{m^\theta_i(y^i \mid z_t^{-i}, t)}{m^\theta_i(z^i_t \mid z_t^{-i}, t)},$$
$$p^\theta_{\mathrm{cav}}(Z^i_s = y \mid z_t) := q^i_{t|s}(z^i_t \mid y)\, \frac{m^\theta_{i,t\to s}(y \mid z_t^{-i}, t)}{m^\theta_i(z^i_t \mid z_t^{-i}, t)}. \tag{D.9}$$

The rate never divides by a per-token likelihood, and out-of-support tokens vanish on their own through the Bayes reweighting, so a cavity head is automatically confined to $\mathcal{S}_i$. Although often presented as a clean-token denoiser, this head is ELBO-optimized by the cavity law.

*Score* (Meng et al., 2022; Sun et al., 2023; Lou et al., 2024; Ou et al., 2025; Zhang et al., 2025). A head $s^\theta_i$ gives the rate

$$a^\theta_{i,\mathrm{score}}(t, z_t, y) := \mathbf{1}\{z_t^{-i} = y^{-i}\}\, R^i_t(y^i, z^i_t)\, s^\theta_i(z^i_t, y^i \mid z_t^{-i}, t), \tag{D.10}$$

whose path loss is exactly SEDD's diffusion-weighted denoising score entropy, with the forward rate a forced rather than auxiliary weight (App. E). For a finite step, a score head is converted to a distribution coordinate (Proposition 3) and sampled as one of those.

Table 5: Conversions between coordinates (Proposition 3), with $m^\theta_i(y^i) = \sum_\ell \mu^\theta_i(e_\ell)\, q^i_{t|0}(y^i \mid e_\ell)$. *Wherever $q^i_{t|0}(z^i_t \mid e_\ell) > 0$. †Recover $m^\theta_i(\cdot) \propto s^\theta_i(z^i_t, \cdot)$ by normalization, then solve the linear system; the inverse is unique iff the channel $q^i_{t|0}$ has full row rank.

| From \ to | denoiser $\pi^\theta_i$ | cavity $\mu^\theta_i$ | score $s^\theta_i$ |
|---|---|---|---|
| denoiser $\pi^\theta_i$ | — | $\propto^*$ $\dfrac{\pi^\theta_i(e_\ell)}{q^i_{t\|0}(z^i_t \mid e_\ell)}$ | $\sum_\ell \pi^\theta_i(e_\ell)\, \dfrac{q^i_{t\|0}(y^i \mid e_\ell)}{q^i_{t\|0}(z^i_t \mid e_\ell)}$ |
| cavity $\mu^\theta_i$ | $\propto \mu^\theta_i(e_\ell)\, q^i_{t\|0}(z^i_t \mid e_\ell)$ | — | $m^\theta_i(y^i)/m^\theta_i(z^i_t)$ |
| score $s^\theta_i$ | via $\mu^\theta_i$, then the cavity row | solve† $m^\theta_i = \mu^\theta_i\, q^i_{t\|0}$ | — |

**Calibration of the cavity and score heads.** Define the *uniform-forward* boundary laws: $p_0^{\text{unif-fwd}}$ is the uniform law on $\mathcal{Y}^L$, $p_t^{\text{unif-fwd}}$ its image under $q_{t|0}$, and the boundary choices are $p^\theta_{t_2} = p_{t_2}^{\text{unif-fwd}}$ and $p^\theta_{0|t_1} = p_{0|t_1}^{\text{unif-fwd}}$. These are the support-matched laws referenced in Proposition 4.

**Proposition 4** (Calibration at initialization)**.** *With the support-matched uniform-forward boundary laws, per token: a cavity head $\equiv 1/V$ has NELBO $\log V$ for every token-factorizing process and window; a score head at zero logits under the folded parameterization represents the same model and gives $\log V$; and a denoiser head $\equiv 1/V$ under uniform diffusion gives $\frac{V-1}{V}\log(1/\beta_{t_1}) + O(1)$, divergent as $t_1 \downarrow 0$.*

*Proof of Proposition 4, cavity head.* It suffices to treat one position. The uniform cavity head induces the noisy marginal $m^\theta(y \mid z_t^{-i}, t) = \frac{1}{V}\sum_\ell q_{t|0}(y \mid e_\ell) = p_t^{\text{unif-fwd}}(y)$, so by (D.9) the model rate equals the oracle reverse rate of the noising process started from $p_0^{\text{unif-fwd}}$. With the matching boundary laws, Corollary C.2 applied to that process makes the ELBO tight: $\mathrm{NELBO}_{[t_1,t_2]}(\theta;\, z_0) = -\log p_0^{\text{unif-fwd}}(z_0) = L\log V$ for every $z_0 \in \mathcal{Y}^L$. With different boundary laws the path-rate calibration still holds but the full NELBO need not equal $L\log V$; under terminal mixing the usual terminal law coincides with $p_{t_2}^{\text{unif-fwd}}$ for an i.i.d. token-wise kernel. □

*Proof of Proposition 4, score head.* By the folded parameterizations (E.8) and (E.25) of App. E, the score head represents $s^\theta_i(z^i_t, y) = s^u_i(z^i_t, y)\, e^{f^\theta_y}$, where $s^u_i$ is the score induced by the uniform cavity through (D.9) (App. E gives $s^u_i$ per kernel). At $f^\theta \equiv 0$ the head equals $s^u_i$ exactly, which by Theorem 2 is the score coordinate of the same reverse process as the uniform cavity head; the rates coincide, the boundary laws are shared, and the NELBO is again $L\log V$. □

The proof for the denoiser head is kernel-specific and given in App. E (Proposition E.2 and Remark E.1). Figure 6 confirms the calibration identities numerically, together with the boundary-term split of the oracle NELBO.

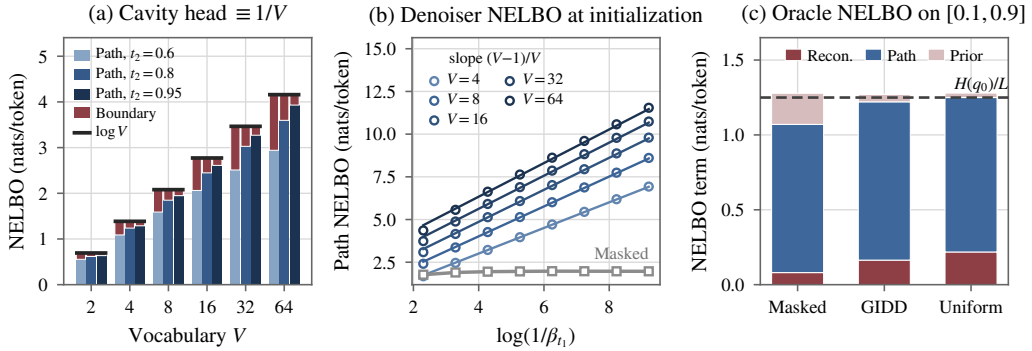


Figure 6: Calibration at initialization, exact on the solvable model. **(a)** A cavity head $\equiv 1/V$ has per-token NELBO $\log V$ once boundary terms are kept, at every $V$ and every $t_2$ (Proposition 4). **(b)** A $1/V$ denoiser head diverges in $\log(1/\beta_{t_1})$ under uniform noising, with asymptotic slope $\frac{V-1}{V}$, while masked stays finite. **(c)** The oracle NELBO splits as reconstruction + path + terminal prior, the latter taken against the support-matched uniform-forward prior of Proposition 4 (at masked it is exactly $\alpha_{t_2} \log V$ by (E.11), whereas an absorbing $\delta_{\mathtt{m}}$ prior would make it infinite on any window with $t_2 < T$). The three terms sum to $H(q_0)$ for every process up to the finite-window residual (window $[0.1, 0.9]$, chosen to exaggerate the boundary terms).

# E MASKED, UNIFORM, AND GIDD NOISING PROCESSES

The noisy alphabet $\mathcal{Z}$ includes a mask state $\mathtt{m}$ for masked and GIDD kernels; uniform diffusion uses the clean vocabulary $\mathcal{Y}$, $V = |\mathcal{Y}|$, shared across positions. All processes take $T = 1$ and use a differentiable, decreasing signal schedule with $\alpha_0 = 1$, $\alpha_1 = 0$,

$$\beta_t := 1 - \alpha_t, \qquad \rho_t := -\frac{\alpha_t'}{\alpha_t}.$$

Table 7 collects the resulting rates, integrands, and boundary terms.

## E.1 COMMON CTMC FORM

All forward kernels factor over positions (Theorem D.1); we use the true reverse rate $a_i$ of (D.1) and the model rates of (D.6)–(D.10), suppress unchanged positions, and write the candidate token $y^i$ as $y$. Leaving the boundary terms implicit, the path integrand is

$$\mathcal{J}_t^\theta(z_0, z_t) = \sum_{i=1}^{L} \sum_{y \in \mathcal{N}_i(z_t)} \Phi\big(a_i(t, z_0, z_t, y),\, a_i^\theta(t, z_t, y)\big)\,, \qquad \mathcal{N}_i(z) := \{y : y^{-i} = z^{-i},\ y^i \neq z^i\}. \tag{E.1}$$

**Proposition E.1** (Cavity ELBO for source-independent jump rates)**.** *Consider a product CTMC whose off-diagonal rates at position $i$ depend only on time and destination, $R_t^i(y, x) = \gamma_i(t, x)$ for $x \neq y$ (the relevant value below being at the current noisy token). Fix $z_0, z_t$ and define the true and model noisy marginals*

$$d_i(y) := q_{t|0}^i(y \mid z_0^i), \qquad m_i^\theta(y) := m_i^\theta(y \mid z_t^{-i}, t).$$

*If $d_i(z_t^i), m_i^\theta(z_t^i) > 0$ and Assumption B.1 holds, the per-position cavity-coordinate integrand is*

$$\mathcal{J}_{i,t}^{\mathrm{cav}} = \frac{\gamma_i(t, z_t^i)}{d_i(z_t^i)} \left[\mathrm{KL}\big(d_i \,\|\, m_i^\theta\big) + D_{\mathrm{IS}}\big(d_i(z_t^i) \,\|\, m_i^\theta(z_t^i)\big)\right], \qquad D_{\mathrm{IS}}(p \,\|\, q) := \frac{p}{q} - \log\frac{p}{q} - 1. \tag{E.2}$$

*Proof.* Write $x := z_t^i$, $d := d_i$, $m^\theta := m_i^\theta$. Equations (D.1) and (D.9), positive homogeneity of $\Phi$, and source-independence give $\mathcal{J}_{i,t}^{\mathrm{cav}} = \gamma_i(t, x) \sum_{y \neq x} \Phi\big(d(y)/d(x),\, m^\theta(y)/m^\theta(x)\big)$. Since $d$ and

$m^\theta$ are probability vectors, $\sum_{y\neq x} d(y) = 1 - d(x)$ and likewise for $m^\theta$, so

$$\begin{aligned}\sum_{y\neq x} \Phi\left(\frac{d(y)}{d(x)}, \frac{m^\theta(y)}{m^\theta(x)}\right) &= \frac{1-m^\theta(x)}{m^\theta(x)} - \frac{1-d(x)}{d(x)} + \frac{1}{d(x)}\sum_{y\neq x} d(y)\log\frac{d(y)\,m^\theta(x)}{d(x)\,m^\theta(y)} \\ &= \frac{1}{d(x)}\left[\sum_y d(y)\log\frac{d(y)}{m^\theta(y)} + \frac{d(x)}{m^\theta(x)} - \log\frac{d(x)}{m^\theta(x)} - 1\right] \\ &= \frac{\mathrm{KL}\big(d \,\|\, m^\theta\big) + D_{\mathrm{IS}}\big(d(x) \,\|\, m^\theta(x)\big)}{d(x)}.\end{aligned}$$

Multiplying by $\gamma_i(t,x)$ proves the claim; summing over $i$ gives the sequence integrand. □

### E.2 Masked diffusion

Let $\mathcal{Z} = \mathcal{Y} \cup \{\mathfrak{m}\}$ with $\mathfrak{m}$ absorbing. The single-position marginal, kernel, and the only forward jump are

$$\begin{aligned} q^i_{t|0}(y \mid z^i_0) &= \alpha_t \mathbf{1}\{y = z^i_0\} + \beta_t \mathbf{1}\{y = \mathfrak{m}\}, \qquad R^i_t(e_\ell, \mathfrak{m}) = \rho_t, \\ q^i_{t|s}(y \mid x) &= \frac{\alpha_t}{\alpha_s}\mathbf{1}\{y = x\} + \big(1 - \frac{\alpha_t}{\alpha_s}\big)\mathbf{1}\{y = \mathfrak{m}\}. \end{aligned} \tag{E.3}$$

The only nonzero true reverse jumps go from the mask to the clean token,

$$a_i(t, z_0, z_t, e_\ell) = -\frac{\alpha'_t}{\beta_t}\mathbf{1}\{e_\ell = z^i_0,\ z^i_t = \mathfrak{m}\}, \tag{E.4}$$

and the cavity head induces $m^\theta_i(\mathfrak{m} \mid z^{-i}_t, t) = \beta_t$ and $m^\theta_i(e_\ell \mid z^{-i}_t, t) = \alpha_t \mu^\theta_i(e_\ell \mid z^{-i}_t, t)$, so

$$a^\theta_{i,\mathrm{cav}}(t, z_t, e_\ell) = -\frac{\alpha'_t}{\beta_t}\,\mu^\theta_i(e_\ell \mid z^{-i}_t, t)\,\mathbf{1}\{z^i_t = \mathfrak{m}\}. \tag{E.5}$$

**Corollary 3** (Denoiser equals cavity in masked diffusion)**.** *Wherever the Bayes factor $q^i_{t|0}(z^i_t \mid e_\ell)$ is constant in $\ell$ on its support, the Bayes reweighting* (12) *is restriction to that support followed by renormalization; at a masked position the support is the whole vocabulary and $\pi^\star_i = \mu^\star_i$.*

*Proof of Corollary 3.* Where $q^i_{t|0}(z^i_t \mid e_\ell)$ is constant in $\ell$ on its support, the Bayes reweighting (12) multiplies $\mu^\star_i$ by the indicator of the support and renormalizes. At a masked position $q_{t|0}(\mathfrak{m} \mid e_\ell) = \beta_t$ for every $\ell$: the support is all of $\mathcal{Y}$, the update is the identity, and $\pi^\star_i = \mu^\star_i$. □

A denoiser head gives the same masked rate with $\pi^\theta_i$ in place of $\mu^\theta_i$; at an unmasked position the denoiser is a point mass rather than the cavity law, but the masked CTMC has no reverse jump out of an unmasked token, so the distinction enters neither the ELBO nor the sampler. At a masked position the jump divergence reduces to cross-entropy: using (E.4), (E.5), and homogeneity of $\Phi$,

$$\mathcal{J}^{\mathrm{mask}}_{i,t}(z_0, z_t) = -\frac{\alpha'_t}{\beta_t}\mathbf{1}\{z^i_t = \mathfrak{m}\}\left[-\log \mu^\theta_i(z^i_0 \mid z^{-i}_t, t)\right], \tag{E.6}$$

for both distribution heads: only masked positions contribute.

**Score coordinate: SEDD and RADD as special cases.** Denoting the clean-conditioned score $\sigma_\ell := q^i_{t|0}(e_\ell \mid z^i_0)/q^i_{t|0}(z^i_t \mid z^i_0)$ and using homogeneity, the per-jump path cost in the score coordinate is

$$\Phi\big(a_i, a^\theta_{i,\mathrm{score}}\big) = R^i_t(y, z^i_t)\,\Phi(\sigma_\ell, s^\theta_i) = R^i_t(y, z^i_t)\big[s^\theta_i - \sigma_\ell \log s^\theta_i + \sigma_\ell(\log\sigma_\ell - 1)\big], \tag{E.7}$$

which summed and averaged is SEDD's denoising score entropy (Lou et al., 2024), with the forward rate as its weight. For masked diffusion the only active channel is $\mathfrak{m} \to e_\ell$, and the true score factors as $s^\star_i(\mathfrak{m}, e_\ell \mid z^{-i}_t, t) = (\alpha_t/\beta_t)\,\mu^\star_i(e_\ell \mid z^{-i}_t, t)$: the only learnable part is the cavity law, time-independent for fixed visible context, which is RADD's reparameterization (Ou et al., 2025). Our implementation folds the uniform-cavity score into the head: at a masked position the head outputs logits $f^\theta$ and represents

$$s^\theta_i(\mathfrak{m}, e_\ell \mid z^{-i}_t, t) = \frac{\alpha_t}{V\beta_t}\,e^{f^\theta_\ell}, \tag{E.8}$$

so $f^\theta \equiv 0$ is the uniform-cavity model and the optimum is $e^{f^\theta_\ell} = V\mu^\star_i(e_\ell)$. Substituting the true score and (E.8) into (E.7), homogeneity pulls out $\alpha_t/\beta_t$ and the logarithms cancel, leaving

$$\mathcal{J}^{\text{mask,score}}_{i,t} = -\frac{\alpha'_t}{\beta_t}\,\mathbf{1}\{z^i_t = \mathfrak{m}\}\Big[\frac{1}{V}\sum_{\ell=1}^{V} e^{f^\theta_\ell} - 1 - f^\theta_{z^i_0} + \log V\Big]. \tag{E.9}$$

With the induced cavity mass $\zeta := \frac{1}{V}\sum_\ell e^{f^\theta_\ell}$,

$$\frac{1}{V}\sum_{\ell} e^{f^\theta_\ell} - 1 - f^\theta_{z^i_0} + \log V = -\log\frac{e^{f^\theta_{z^i_0}}}{\sum_\ell e^{f^\theta_\ell}} + \big(\zeta - 1 - \log\zeta\big), \tag{E.10}$$

so the same logits read as a cavity head give the cross-entropy integrand (E.6), and the score loss exceeds it by the normalization slack $\zeta - 1 - \log\zeta \geq 0$, zero iff $\zeta = 1$: a masked score head is a cavity head whose normalization is learned rather than imposed.

**Prior term.** The natural masked prior is the deterministic mask law $\delta_{\mathfrak{m}}$, whose per-token KL against $q^i_{t_2|0}$ is 0 if $\alpha_{t_2} = 0$ and $+\infty$ otherwise. The uniform-forward prior $p^{\text{unif-fwd}}_{t_2} = \beta_{t_2}\delta_{\mathfrak{m}} + \frac{\alpha_{t_2}}{V}\sum_\ell \delta_{e_\ell}$ gives the finite truncation correction

$$\mathrm{KL}\big(q^i_{t_2|0}(\cdot \mid z^i_0)\,\big\|\,p^{\text{unif-fwd}}_{t_2}\big) = \alpha_{t_2}\log V, \tag{E.11}$$

vanishing as $t_2 \uparrow 1$.

**Sampling.** For a step $t \to s$, a visible token stays; a masked token unmasks with the plug-in step

$$p^\theta(Z^i_s = \mathfrak{m} \mid z_t) = \frac{\beta_s}{\beta_t}, \qquad p^\theta(Z^i_s = e_\ell \mid z_t) = \frac{\alpha_s - \alpha_t}{\beta_t}\,\mu^\theta_i(e_\ell \mid z^{-i}_t, t). \tag{E.12}$$

Once unmasked, a token cannot be revised. At the final step $t_1$, visible tokens are kept and remaining masks decode from $\pi^\theta_i(\cdot \mid z_{t_1}, t_1)$.

### E.3 Uniform diffusion

Uniform diffusion uses $\mathcal{Z} = \mathcal{Y}$ and mixes the clean token with the uniform law:

$$\begin{aligned} q^i_{t|0}(e_r \mid e_\ell) &= \alpha_t\mathbf{1}\{e_r = e_\ell\} + \frac{\beta_t}{V}, \qquad R^i_t(e_r, e_\ell) = \frac{\rho_t}{V}, \\ q^i_{t|s}(e_r \mid e_\ell) &= \frac{\alpha_t}{\alpha_s}\mathbf{1}\{e_r = e_\ell\} + \Big(1 - \frac{\alpha_t}{\alpha_s}\Big)\frac{1}{V} \end{aligned} \tag{E.13}$$

for $r \neq \ell$: the process proposes uniform replacements at rate $\rho_t$. The true reverse rate for $e_\ell \neq z^i_t$ is

$$a_i(t, z_0, z_t, e_\ell) = \frac{\rho_t}{V}\,\frac{\alpha_t\mathbf{1}\{e_\ell = z^i_0\} + \beta_t/V}{\alpha_t\mathbf{1}\{z^i_t = z^i_0\} + \beta_t/V}, \tag{E.14}$$

and the cavity head induces $m^\theta_i(e_\ell \mid z^{-i}_t, t) = \alpha_t\mu^\theta_i(e_\ell \mid z^{-i}_t, t) + \beta_t/V$. Proposition E.1 applies with $\gamma_i = \rho_t/V$ and

$$d_i(e_\ell) = q^i_{t|0}(e_\ell \mid z^i_0) = \alpha_t\mathbf{1}\{e_\ell = z^i_0\} + \frac{\beta_t}{V},$$

so that (E.14) is $\frac{\rho_t}{V}\,d_i(e_\ell)/d_i(z^i_t)$; unlike masked diffusion, every position can contribute. The three integrands are

$$\mathcal{J}^{\text{unif,cav}}_{i,t} = \frac{\rho_t/V}{d_i(z^i_t)}\left[\mathrm{KL}\big(d_i \,\big\|\, m^\theta_i\big) + D_{\text{IS}}\big(d_i(z^i_t) \,\big\|\, m^\theta_i(z^i_t)\big)\right],$$

$$\mathcal{J}^{\text{unif,den}}_{i,t} = \frac{\rho_t}{V}\sum_{e_\ell \neq z^i_t}\Phi\left(\frac{d_i(e_\ell)}{d_i(z^i_t)},\ \sum_r \pi^\theta_i(e_r \mid z_t, t)\,\frac{q^i_{t|0}(e_\ell \mid e_r)}{q^i_{t|0}(z^i_t \mid e_r)}\right),$$

$$\mathcal{J}^{\text{unif,score}}_{i,t} = \frac{\rho_t}{V}\sum_{e_\ell \neq z^i_t}\Phi\left(\frac{d_i(e_\ell)}{d_i(z^i_t)},\ s^\theta_i(z^i_t, e_\ell)\right),$$

the last being (E.7) with weight $\rho_t/V$. The second argument of $\mathcal{J}_{i,t}^{\mathrm{unif,den}}$, the denoiser's average of kernel ratios, is abbreviated $g_i^\theta(y) := \sum_\ell \pi_i^\theta(e_\ell \mid z_t, t)\, q_{t|0}^i(y \mid e_\ell)/q_{t|0}^i(z_t^i \mid e_\ell)$ below. The implementation represents the score by the head logits directly, $s_i^\theta(z_t^i, e_\ell) = e^{f_\ell^\theta}$, the uniform-cavity fold with factor one: the uniform cavity induces $m^u \equiv 1/V$ and score $s^u \equiv 1$, so at $f^\theta \equiv 0$ the model rate equals the forward rate, the reverse process at stationarity, as in SEDD (Lou et al., 2024).

**Prior term.** The natural prior, the stationary uniform law, coincides with the uniform-forward law, with per-token KL

$$\mathrm{KL}\big(q_{t_2|0}^i(\cdot \mid z_0^i) \,\big\|\, \mathrm{unif}(\mathcal{Y})\big) = \Big(\alpha_{t_2} + \frac{\beta_{t_2}}{V}\Big) \log\big(V\alpha_{t_2} + \beta_{t_2}\big) + \frac{(V-1)\beta_{t_2}}{V} \log \beta_{t_2} \xrightarrow[t_2 \uparrow 1]{} 0. \tag{E.15}$$

**Sampling.** Every position may be resampled. The plug-in step is

$$p^\theta(Z_s^i = e_\ell \mid z_t) = q_{t|s}^i(z_t^i \mid e_\ell)\, \frac{\alpha_s\, \mu_i^\theta(e_\ell \mid z_t^{-i}, t) + \beta_s/V}{\alpha_t\, \mu_i^\theta(z_t^i \mid z_t^{-i}, t) + \beta_t/V}, \tag{E.16}$$

a denoiser head entering after the inverse Bayes reweighting (12); the final step decodes every position from $\pi_i^\theta(\cdot \mid z_{t_1}, t_1)$.

**Proposition E.2** (Uninformative denoiser ELBO diverges for uniform diffusion)**.** *For uniform diffusion with $\alpha_0 = 1$ and $V \geq 2$, the head $\pi_i^\theta \equiv 1/V$ has, as $t \downarrow 0$,*

$$\mathbb{E}_{q_{t|0}}\Big[\mathcal{J}_{i,t}^{\mathrm{unif,den}}\Big] = \frac{V-1}{V}\, \frac{-\alpha_t'}{\beta_t}\, (1 + o(1)),$$

*non-integrable at $t = 0$, so that $\mathrm{NELBO}_{[t_1,t_2]}(\theta) = \frac{(V-1)L}{V} \log \frac{1}{\beta_{t_1}} + O(1)$ as $t_1 \downarrow 0$.*

*Proof.* Fix one position with clean token $z_0^i$. The true rates (E.14) satisfy, as $\beta_t \to 0$: if $z_t^i = z_0^i$, every jump away decays like $O(\beta_t)$; if $z_t^i \neq z_0^i$, the correcting jump $e_\ell = z_0^i$ has rate $(-\alpha_t'/\beta_t)(1 + o(1))$ (using $\rho_t \alpha_t = -\alpha_t'$) and the others are $O(\rho_t)$. The uninformative model rate is, up to the factor $\rho_t/V$, the average $\frac{1}{V}\sum_r g$ with $e_r$ in place of $z_0^i$; since $e_\ell \neq z_t^i$, exactly one term ($e_r = e_\ell$) diverges like $\alpha_t/\beta_t$ while the rest stay bounded, so

$$a_{i,\mathrm{den}}^\theta(t, z_t, e_\ell) = \frac{-\alpha_t'}{V\beta_t} + O(-\alpha_t') \qquad \text{for every } e_\ell \neq z_t^i.$$

On the event $z_t^i = z_0^i$ (probability $1 - O(\beta_t)$), the ratio $a_i/a_i^\theta = O(\beta_t^2)$, so $\Phi(a_i, a_i^\theta) = a_i^\theta\,(1 + O(\beta_t^2|\log \beta_t|))$ and the $V-1$ destinations contribute

$$\mathbb{E}_{q_{t|0}}\Big[\mathcal{J}_{i,t}^{\mathrm{unif,den}} \mathbf{1}\{z_t^i = z_0^i\}\Big] = (V-1)\frac{-\alpha_t'}{V\beta_t} + O(-\alpha_t').$$

On the complementary event, of probability $(V-1)\beta_t/V$, both rates are $O(-\alpha_t'/\beta_t)$ with the model rate of the same order, so $\Phi(a_i, a_i^\theta) = O(-\alpha_t'/\beta_t)$ and the event probability cancels the singularity, contributing $O(-\alpha_t')$. Adding the two and integrating with $-\alpha_t'\,dt = d\beta_t$ gives the per-position path term $\frac{V-1}{V}\log(1/\beta_{t_1}) + O(1)$; the $L$ positions add. □

The failure mode is the jump support: the uninformative denoiser allows jumps away from clean tokens near $t = 0$, and those kernel ratios diverge like $1/\beta_t$, whereas the cavity's induced marginal carries a matching $O(\beta_t)$ Bayes factor that downweights them and keeps its ELBO finite.

### E.4 Generalized interpolating discrete diffusion

GIDD (von Rütte et al., 2025) chooses a differentiable mixing law $\omega_t$ on $\mathcal{Z}$ (a row vector) and sets

$$q_{t|0}^i(y \mid e_\ell) = \alpha_t \mathbf{1}\{y = e_\ell\} + \beta_t \omega_t(y), \qquad q_{t|0}^i = \alpha_t I + \beta_t \mathbf{1}\omega_t. \tag{E.17}$$

Solving $q_{t|0}^i = q_{s|0}^i\, q_{t|s}^i$ within the same rank-one family gives

$$q_{t|s}^i = \alpha_{t|s} I + \beta_{t|s} \mathbf{1}\omega_{t|s}, \qquad \alpha_{t|s} := \frac{\alpha_t}{\alpha_s}, \qquad \beta_{t|s}\omega_{t|s} := \beta_t\omega_t - \frac{\alpha_t}{\alpha_s}\beta_s\omega_s, \tag{E.18}$$

and expanding at $t = s + \Delta$ gives the destination-rate vector and off-diagonal generator

$$h_t(y) := \beta_t \omega_t'(y) + \rho_t \omega_t(y), \qquad R_t^i(x, y) = h_t(y), \qquad y \neq x, \tag{E.19}$$

with $\sum_y h_t(y) = \rho_t$; valid schedules require $h_t \geq 0$ off the diagonal. Consistent with the interior-window assumptions, we take the supports of $\omega_t$ and $h_t$ fixed on $(0, 1)$; strictly positive entries are then bounded away from zero on compact interior windows, which is how the clipped window fixes the supports of the family. Our experiments use the static mixture $\omega_t \equiv \omega = (1 - \lambda)\delta_{\mathtt{m}} + \lambda \,\mathrm{unif}(\mathcal{Y})$ (§5), for which $h_t = \rho_t \omega$; masked and uniform diffusion are $\omega = \delta_{\mathtt{m}}$ and $\omega = \mathrm{unif}(\mathcal{Y})$.

The true and cavity reverse rates take the form

$$a_i(t, z_0, z_t, y) = h_t(z_t^i) \, \frac{q^i_{t|0}(y \mid z_0^i)}{q^i_{t|0}(z_t^i \mid z_0^i)}, \qquad a^\theta_{i,\mathrm{cav}}(t, z_t, y) = h_t(z_t^i) \, \frac{m_i^\theta(y \mid z_t^{-i}, t)}{m_i^\theta(z_t^i \mid z_t^{-i}, t)}, \tag{E.20}$$

with $m_i^\theta(y \mid z_t^{-i}, t) = \alpha_t \mu_i^\theta(y \mid z_t^{-i}, t)\mathbf{1}\{y \in \mathcal{Y}\} + \beta_t \omega_t(y)$, and Proposition E.1 applies with destination-dependent $\gamma_i(t, y) = h_t(y)$ and $d_i(y) = \alpha_t \mathbf{1}\{y = z_0^i\} + \beta_t \omega_t(y)$:

$$\mathcal{J}_{i,t}^{\mathrm{GIDD,cav}} = \frac{h_t(z_t^i)}{d_i(z_t^i)} \left[ \mathrm{KL}\big(d_i \,\big\|\, m_i^\theta\big) + D_{\mathrm{IS}}\big(d_i(z_t^i) \,\big\|\, m_i^\theta(z_t^i)\big) \right]. \tag{E.21}$$

A denoiser head gives, for $y \neq z_t^i$, the rate

$$a^\theta_{i,\mathrm{den}}(t, z_t, y) = h_t(z_t^i) \sum_\ell \pi_i^\theta(e_\ell \mid z_t, t) \, \frac{\alpha_t \mathbf{1}\{y = e_\ell\} + \beta_t \omega_t(y)}{\alpha_t \mathbf{1}\{z_t^i = e_\ell\} + \beta_t \omega_t(z_t^i)}. \tag{E.22}$$

Since $\Phi$ is 1-homogeneous and the true rate is $h_t(z_t^i)\, d_i(y)/d_i(z_t^i)$, the denoiser and score integrands are

$$\mathcal{J}_{i,t}^{\mathrm{GIDD,den}} = h_t(z_t^i) \sum_{y \neq z_t^i} \Phi\left( \frac{d_i(y)}{d_i(z_t^i)}, \, \sum_\ell \pi_i^\theta(e_\ell \mid z_t, t) \, \frac{q^i_{t|0}(y \mid e_\ell)}{q^i_{t|0}(z_t^i \mid e_\ell)} \right), \tag{E.23}$$

$$\mathcal{J}_{i,t}^{\mathrm{GIDD,score}} = h_t(z_t^i) \sum_{y \neq z_t^i} \Phi\left( \frac{d_i(y)}{d_i(z_t^i)}, \, s_i^\theta(z_t^i, y \mid z_t^{-i}, t) \right), \tag{E.24}$$

with App. E.3 the case $\omega_t \equiv 1/V$, $h_t \equiv \rho_t / V$. The score head folds the uniform-cavity score: with $m^u(y) := \frac{\alpha_t}{V}\mathbf{1}\{y \in \mathcal{Y}\} + \beta_t \omega_t(y)$ the noisy marginal induced by a uniform cavity,

$$s_i^\theta(z_t^i, y \mid z_t^{-i}, t) = \frac{m^u(y)}{m^u(z_t^i)} \, e^{f_y^\theta}, \tag{E.25}$$

so $f^\theta \equiv 0$ represents the uniform-cavity model on every channel; the factor is constant on each reverse channel (content destinations share $m^u(y) = (\alpha_t + \beta_t \lambda)/V$ under the static mixture, and the mask destination carries $m^u(\mathtt{m}) = \beta_t \omega(\mathtt{m})$). Uniform diffusion is recovered at $\omega(\mathtt{m}) = 0$, where the factor is one, and masked diffusion at $\omega = \delta_{\mathtt{m}}$, where the only channel is $\mathtt{m} \to e_\ell$ and the factor is $\alpha_t/(V\beta_t)$, as in (E.8).

**Prior term.** The natural prior is $\omega_{t_2}$; for $\omega_{t_2}(z_0^i) > 0$ the per-token KL is

$$\mathrm{KL}\big(q^i_{t_2|0}(\cdot \mid z_0^i) \,\big\|\, \omega_{t_2}\big) = \big(\alpha_{t_2} + \beta_{t_2}\omega_{t_2}(z_0^i)\big) \log \frac{\alpha_{t_2} + \beta_{t_2}\omega_{t_2}(z_0^i)}{\omega_{t_2}(z_0^i)} + \beta_{t_2}(1 - \omega_{t_2}(z_0^i)) \log \beta_{t_2}, \tag{E.26}$$

infinite by support mismatch if $\omega_{t_2}(z_0^i) = 0 < \alpha_{t_2}$. The uniform-forward prior $p_{t_2}^{\text{unif-fwd}}(y) = \frac{\alpha_{t_2}}{V}\mathbf{1}\{y \in \mathcal{Y}\} + \beta_{t_2}\omega_{t_2}(y)$ is finite whenever $\alpha_{t_2} > 0$:

$$\mathrm{KL}\big(q^i_{t_2|0}(\cdot \mid z_0^i) \,\big\|\, p_{t_2}^{\text{unif-fwd}}\big) = \sum_{y \in \mathcal{Y}} \big(\alpha_{t_2}\mathbf{1}\{y = z_0^i\} + \beta_{t_2}\omega_{t_2}(y)\big) \log \frac{\alpha_{t_2}\mathbf{1}\{y = z_0^i\} + \beta_{t_2}\omega_{t_2}(y)}{\alpha_{t_2}/V + \beta_{t_2}\omega_{t_2}(y)}, \tag{E.27}$$

recovering $\alpha_{t_2} \log V$ for masked and the stationary case for uniform.

**Corollary E.1** (Conversions for GIDD). *For the kernel* (E.17) *with* $\alpha_t > 0$*, the channel* $q^i_{t|0}$ *has full row rank and the score-to-cavity map is affine:*

$$\mu^\theta_i(e_\ell) = \frac{m^\theta_i(e_\ell) - \beta_t\,\omega_t(e_\ell)}{\alpha_t}, \qquad m^\theta_i(y) = \frac{s^\theta_i(z^i_t, y)}{\sum_{y'} s^\theta_i(z^i_t, y')}, \tag{E.28}$$

*a probability vector iff* $m^\theta_i(e_\ell) \geq \beta_t\omega_t(e_\ell)$ *for every* $\ell$ *and* $m^\theta_i(\mathfrak{m}) = \beta_t\omega_t(\mathfrak{m})$.

*Proof.* The noise floor $\beta_t\omega_t(y)$ is source-independent, so $m^\theta_i(e_\ell) = \alpha_t\mu^\theta_i(e_\ell) + \beta_t\omega_t(e_\ell)$ decouples across $\ell$; solve for $\mu^\theta_i$, with $m^\theta_i$ recovered from a score as in Proposition 3. The mask carries no cavity mass, so $\mu^\theta_i$ sums to one iff the mask entry sits at its noise floor. □

Table 6: Conversions for GIDD (Corollary E.1), with current state $z^i_t$, Bayes factor $q^i_{t|0}(z^i_t \mid e_\ell) = \alpha_t\mathbf{1}\{e_\ell = z^i_t\} + \beta_t\omega_t(z^i_t)$, and cavity-induced marginal $m^\theta_i(y) = \alpha_t\mu^\theta_i(y)\mathbf{1}\{y \in \mathcal{Y}\} + \beta_t\omega_t(y)$ (taken $\propto s^\theta_i(z^i_t, y)$ when starting from a score). Masked ($\omega_t = \delta_\mathfrak{m}$) gives $\mu^\theta_i(e_\ell) = (\beta_t/\alpha_t)\,s^\theta_i(\mathfrak{m}, e_\ell)$, which is RADD; uniform gives $\mu^\theta_i(e_\ell) = (m^\theta_i(e_\ell) - \beta_t/V)/\alpha_t$.

| From \ to | denoiser $\pi^\theta_i(e_\ell)$ | cavity $\mu^\theta_i(e_\ell)$ | score $s^\theta_i(z^i_t, y)$ |
|---|---|---|---|
| denoiser | — | $\propto \pi^\theta_i(e_\ell)/q^i_{t\|0}(z^i_t \mid e_\ell)$ | $m^\theta_i(y)/m^\theta_i(z^i_t)$ |
| cavity | $\propto \mu^\theta_i(e_\ell)\,q^i_{t\|0}(z^i_t \mid e_\ell)$ | — | $m^\theta_i(y)/m^\theta_i(z^i_t)$ |
| score | $\propto \mu^\theta_i(e_\ell)\,q^i_{t\|0}(z^i_t \mid e_\ell)$ (via $s \to \mu$) | $\dfrac{m^\theta_i(e_\ell) - \beta_t\omega_t(e_\ell)}{\alpha_t}$ | — |

*Remark* E.1 (The divergence for GIDD). The divergence is not special to uniform diffusion. For a kernel $q^i_{t|0}(y \mid e_\ell) = \alpha_t\mathbf{1}\{y = e_\ell\} + \beta_t\omega_t(y)$, fix a clean token $x$ with $\omega_t(x) > 0$ near $t = 0$ and consider the revealed event $z^i_t = z^i_0 = x$, of probability tending to one. For any clean destination $y \neq x$, the $e_\ell = y$ term of the uninformative denoiser gives $g^\theta_i(y) \geq \frac{1}{V}\,\frac{\alpha_t + \beta_t\omega_t(y)}{\beta_t\omega_t(x)}$ while $d_i(y)/d_i(z^i_t) = O(\beta_t)$, so $\Phi\big(d_i(y)/d_i(z^i_t), g^\theta_i(y)\big) = g^\theta_i(y)\,(1 + o(1))$, and this one destination contributes at order $h_t(x)/(\beta_t\omega_t(x))$, where $h_t$ is the forward destination rate of (E.19). This has the exact primitive

$$\frac{h_t(x)}{\beta_t\omega_t(x)} = \frac{d}{dt}\log\frac{\beta_t\omega_t(x)}{\alpha_t},$$

whose integral diverges as the lower endpoint tends to zero, since $\beta_t\omega_t(x)/\alpha_t \to 0$. Any clean token carrying positive $\omega_t$-mass and positive data probability produces the divergence, including uniform and every full-support hybrid.

## F Exact verification on a solvable model

Every identity of the paper has been verified on an exactly enumerable model, whose oracle laws, entropies, and path KLs are available in closed form. The data law is a homogeneous bigram chain on $V = 8$ tokens of length $L = 8$ (transition matrix drawn from a row-softmax of $\mathcal{N}(0, 1)$ entries at temperature 0.3), with $H(q_0) = 10.00$ nats, that is 1.25 per token. We instantiate masked, uniform, and GIDD noising with the linear schedule $\alpha_t = 1 - t$ on the window $[t_1, t_2] = [0.02, 0.98]$, the GIDD member mixing through $\omega = (1 - \lambda)\delta_\mathfrak{m} + \lambda\,\mathrm{unif}(\mathcal{Y})$ with $\lambda = \frac{1}{2}$ except where swept. Every floor, calibration, and conversion reported is oracle-exact and assembled by the implementation of §5 evaluated at the exact head; only Figure 5 involves training (a small DiT: hidden 128, 4 heads, 3 blocks; AdamW (Loshchilov & Hutter, 2019) at $10^{-3}$ under a cosine decay, batch 256, 32,000 steps, each head trained on the ELBO of its own coordinate). On a finite window the oracle's own NELBO sits 0.001–0.003 nats/token above $H(q_0)/L$ and equals the residual boundary terms (namely, the factorization error from the final decoding step at $t_1$ plus the excess of the fixed prior $I(Z_0; Z_{t_2})$). The three figures are placed with the results they verify: Figure 4 verifies the shared floor and its rate (App. C.2), Figure 5 the Oracle Distance decay and the Pythagorean split (App. C.3), and Figure 6 the calibration identities (App. D.2).

Table 7: The three processes as instances of the common form, with $d_i(\cdot) = q^i_{t|0}(\cdot \mid z^i_0)$, $m^\theta_i = m^\theta_i(\cdot \mid z_t^{-i}, t)$ (Proposition E.1), $g^\theta_i$ the denoiser's average of kernel ratios (App. E.3), and $h_t(y) = \beta_t\omega'_t(y) + \rho_t\omega_t(y)$ the GIDD destination rate of (E.19); heads are written without their conditioning arguments. At masked positions the denoiser and cavity rows coincide (Corollary 3). The score rows use the folded parameterizations (E.8) and (E.25). Both terminal priors are recorded; the uniform-forward one is the support-matched choice of Proposition 4.

| | Masked | Uniform | GIDD |
|---|---|---|---|
| Forward rate | $R^i_t(e_y, \mathtt{m}) = \rho_t$ | $R^i_t(e_y, e_x) = \dfrac{\rho_t}{V}$ | $R^i_t(x, y) = h_t(y)$ |
| Reverse rate $(\widehat{Q}^\theta_t)_i(z_t, y)$ | $-\dfrac{\alpha'_t}{\beta_t}\mathbf{1}\{z^i_t = \mathtt{m}\}\mu^\theta_i(y^i)$ | $h_t(z^i_t)\,\dfrac{\alpha_t\mu^\theta_i(y^i) + \beta_t\omega_t(y^i)}{\alpha_t\mu^\theta_i(z^i_t) + \beta_t\omega_t(z^i_t)}$ (uniform: $h_t \equiv \rho_t/V, \omega_t \equiv 1/V$) | |
| Sampling $p^\theta_{s\|t}(z^i_s \mid z_t)$ | $\begin{cases} \dfrac{\alpha_s - \alpha_t}{\beta_t}\,\mu^\theta_i(z^i_s), & z^i_s \in \mathcal{Y} \\ \dfrac{\beta_s}{\beta_t}, & z^i_s = \mathtt{m} \end{cases}$ | $q^i_{t\|s}(z^i_t \mid z^i_s)\,\dfrac{\alpha_s\,\mu^\theta_i(z^i_s) + \beta_s\omega_s(z^i_s)}{\alpha_t\,\mu^\theta_i(z^i_t) + \beta_t\omega_t(z^i_t)}$ | |
| $\mathcal{J}^{\mathrm{den}}_{i,t}$ | $-\dfrac{\alpha'_t}{\beta_t}\,\mathbf{1}\{z^i_t = \mathtt{m}\}\big[-\log\pi^\theta_i(z^i_0)\big]$ | $h_t(z^i_t)\sum_{y \neq z^i_t}\Phi\big(d_i(y)/d_i(z^i_t),\, g^\theta_i(y)\big)$ | |
| $\mathcal{J}^{\mathrm{cav}}_{i,t}$ | $-\dfrac{\alpha'_t}{\beta_t}\,\mathbf{1}\{z^i_t = \mathtt{m}\}\big[-\log\mu^\theta_i(z^i_0)\big]$ | $\dfrac{h_t(z^i_t)}{d_i(z^i_t)}\Big[\mathrm{KL}\big(d_i \,\big\|\, m^\theta_i\big) + D_{\mathrm{IS}}\big(d_i(z^i_t) \,\big\|\, m^\theta_i(z^i_t)\big)\Big]$ | |
| $\mathcal{J}^{\mathrm{score}}_{i,t}$ | $-\dfrac{\alpha'_t}{\beta_t}\mathbf{1}\{z^i_t = \mathtt{m}\}$ $\cdot\big[\frac{1}{V}\sum_\ell e^{f^\theta_\ell} - 1 - f^\theta_{z^i_0} + \log V\big]$ | $h_t(z^i_t)\sum_{y \neq z^i_t}\Phi\big(d_i(y)/d_i(z^i_t),\, s^\theta_i(z^i_t, y)\big)$ | |
| $1/V$ denoiser NELBO | $L\log V$ | $\frac{(V-1)L}{V}\log\frac{1}{\beta_{t_1}}$ $+ O(1)$ | diverges (Rmk. E.1) |
| $1/V$ cavity NELBO | | $L\log V$ | |
| Natural prior per-token KL | 0 if $\alpha_{t_2} = 0$, else $+\infty$ | (E.15) | (E.26) |
| Unif.-fwd. per-token KL | $\alpha_{t_2}\log V$ | (E.15) | (E.27) |

# G EXPERIMENTAL DETAILS

## G.1 IMPLEMENTATION DETAILS

All 11 runs share the configuration of Table 8, differing only in noise kernel and coordinate. The architecture is the diffusion transformer of Sahoo et al. (2024), the same one used by Lou et al. (2024), with time conditioning by adaLN in every run; of the 169.6M total parameters, 92.2M are non-embedding, 7.4M of those the time-conditioning path, and the remainder the token embedding and the output head over the GPT-2 vocabulary ($V = 50{,}257$, plus the mask state). Muon (Jordan et al., 2024) is applied to the hidden matrices under the RMS-matched update scaling of Liu et al. (2025) (token embedding, output head, and sub-matrix parameters routed to AdamW; Loshchilov & Hutter, 2019).

Table 8: Shared configuration of the 11 OpenWebText training runs and of every reported evaluation and sampling procedure. Kernel and coordinate are the only differences between runs.

| Architecture | | Optimization | |
|---|---|---|---|
| Layers | 12 | Objective | ELBO, own coordinate |
| Hidden width | 768 | Optimizer | Muon, RMS-matched |
| Attention heads | 12 | Learning rate | $10^{-3}$, constant |
| MLP width | 3072 | Warmup | linear from 0, 1k steps |
| Position embedding | rotary | Steps | 500k (65B tokens) |
| Time conditioning | adaLN, width 128 | Batch size | 512 |
| Non-embedding params | 92.2M | Muon momentum | 0.95, Nesterov |
| Total params | 169.6M | Weight decay | 0.01, excl. norms/emb. |
| Sequence length | 256 | Gradient clip | 8 |
| Vocabulary | 50,257 + mask | EMA decay | 0.9999 |
| Dropout | 0 | Time sampling | low-discrepancy uniform |
| | | Window $[t_1, t_2]$ | $[0.001, 0.9999]$ |
| | | Adam $(\beta_1, \beta_2)$ | $(0.9, 0.999)$ |
| **Likelihood evaluation** | | **Sampling** | |
| Weights | EMA | Sampler | ancestral, denoiser coord. |
| Split | test | Time grid | cosine |
| Batches | 100 (51,200 sequences) | NFE | 256, swept $\{16, \dots, 1024\}$ |
| Time bins | 50, equal mass | Temperature, top-$p$ | 1, 1 |
| Units | bits/token, full bound | Samples per point | 2,048 |

Every run trains the path term of the ELBO with the estimator of (B.9), under the linear schedule $\alpha_t = 1 - t$ and uniform time sampling, and the boundary terms enter only the reported NELBOs. Score heads use the folded parameterizations (E.8) and (E.25).

All heads are stored as logits and converted in log space following Table 6. A distribution head maps to the score through its noisy marginal; a score head maps back by normalizing $s_i^\theta(z_t^i, \cdot)$ to $m_i^\theta$ and applying the affine inverse (E.28), where entries violating $m_i^\theta \geq \beta_t \omega_t$ lie off the image of the forward channel and are set to zero, and the rest renormalized.

All coordinates share the reconstruction term,

$$-\mathbb{E}_{q_{t_1|0}(z_{t_1}|z_0)}\left[\sum_i \log \pi_i^\theta(z_0^i \mid z_{t_1}, t_1)\right], \tag{G.1}$$

with the decoder obtained from the trained head by these conversions (for a score head, score $\to$ cavity $\to$ denoiser); the sampler's final decode uses the same decoder, so the reconstruction term scores the decoder actually sampled from. Because the affine inverse can place mass exactly zero on the true token, the implementation smooths the decoder, $\pi_i^\theta \mapsto (1-\epsilon)\pi_i^\theta + \epsilon/V$ with $\epsilon = 10^{-8}$, bounding the reconstruction term by $\log(V/\epsilon)$ per token, and the ELBO is exact for the smoothed decoder.

**Samplers.** Two samplers appear in the paper, both applied in the denoiser coordinate, so a head trained in another coordinate is converted first through Proposition 3. The *ancestral* sampler is the factorized backward jump (14): on a grid $t_2 = t_0 > \cdots > t_N = t_1$ it draws every position of $z_{t_{k+1}}$

from its exact conditional given $z_{t_k}$, so each step multiplies the head's law by the analytic factor $q^i_{t|s}(z^i_t \mid y)$ and keeps the state on the noising path. The *re-corrupt* sampler instead decodes a clean sequence $\hat{z}_0$ from the head's law at $z_{t_k}$ and forward-corrupts it to the next level, $z_{t_{k+1}} \sim q_{t_{k+1}|0}(\cdot \mid \hat{z}_0)$. It never consults $q^i_{t|s}$, so the whole step is carried by the decoded law; in exchange every state it produces is a forward corruption of a complete, self-consistent sequence, which is what lets a solver revise a committed answer instead of only refining an unfinished one. Reading a head in the wrong coordinate, the misread of §5.1, changes only the head-to-law map and leaves both samplers otherwise untouched.

**Data.** OpenWebText documents are tokenized with the GPT-2 tokenizer ($V = 50{,}257$ content tokens plus a mask state) and concatenated, then cut into records of exactly 256 tokens, so no record carries padding and the token stream and its order are those of the corpus. OpenWebText has no released validation split, so we hold out one of the 73 shards the corpus is cut into, about $1.4\%$ of the documents, and train on the rest. Splitting the held-out shard's records into even and odd halves gives the validation and test sets. All paper numbers are computed on the test half, on a fixed set of 100 batches (51,200 sequences, SE $\approx 0.004$ bits/token). Only the training curves of Figure 3 are validation (32 batches per point), labeled as such.

### G.2 Logic tasks

Each instance is a token sequence whose problem part is a clean anchor, never corrupted and excluded from the loss, and whose solution part is diffused and checked by a verifier after decoding.

**Sudoku.** The Sudoku-Extreme boards of Wang et al. (2025), as 81 digit tokens with the given cells fixed: 3.8M training boards and a test pool of 423k, split into orbit-disjoint validation and test sets of about 200k and 223k. An embedding added to the input marks the given cells, without which a uniformly corrupted cell is indistinguishable from a given clue and uniform diffusion does not train. A board is solved when it satisfies every Sudoku constraint.

**Boxoban.** $10 \times 10$ Sokoban levels (Guez et al., 2018) with the A$^*$ solutions of Garriga-Alonso et al. (2024), encoded as the 100 board cells followed by 128 action tokens over {up, right, down, left, stay}, where *stay* pads the solution and is scored like a move. The 92k solved levels of the official validation pool give validation and test sets of 25k each, the rest joining the 936k training levels. A level is solved when the simulated moves place every box on a target.

**Maze.** $40 \times 40$ mazes with their shortest path, encoded as 1,600 cells followed by 192 moves in the same alphabet: 500k generated training mazes and 25k held-out ones. The model encodes the board with a convolutional grid encoder, which the move tokens read by cross-attention. A maze is solved when the path reaches the goal.

**Graph coloring.** Planted 4-colorings of $n = 64$ vertices with cross-color edges present independently with probability $p = 0.2$, an average degree of 9.6 just above the Kesten–Stigum detectability threshold of 9. The sequence is one color token per vertex, the graph enters as the attention mask of half the heads, and instances are drawn fresh from seeded generators. Any proper coloring counts as solved.

**Training and evaluation.** Table 9 lists the configuration of every run and of its evaluation. Training augments Sudoku with digit relabeling, row and column permutations within and across bands and stacks, and transposition, and Boxoban and maze with the eight rotations and reflections of the board, the moves remapped to match. Solve rates (Figure 2) and NELBOs (Table 2) are averaged over the three training seeds. Reading the head as a denoiser skips the Bayes reweighting of Proposition 3 and leaves everything else unchanged.

### G.3 Metric estimation details

**Likelihood evaluation.** Reported NELBOs are the full finite-window bound (B.3): the path term plus the reconstruction term at $t_1$ and the prior term at $t_2$, taken against the support-matched uniform-forward prior of Proposition 4. Table 10 reports how much the two endpoint terms contribute. Units

Table 9: Configuration of the logic-task runs and of their evaluation: settings shared by the four tasks (top) and per task (bottom).

| Architecture | | Optimization | |
|---|---|---|---|
| Layers | 12 | Objective | ELBO, cavity coordinate |
| Hidden width | 512 | Kernel | uniform |
| Attention heads | 8 | Optimizer | AdamW |
| MLP width | 2048 | Adam $(\beta_1, \beta_2)$ | $(0.9, 0.999)$ |
| Time conditioning | adaLN, width 128 | Warmup | linear from 0, 1k steps |
| Dropout | 0 | Batch size | 512 |
| | | Weight decay | 0.05, excl. norms/emb. |
| | | Gradient clip | 1 |
| | | EMA decay | 0.9999 |
| | | Window $[t_1, t_2]$ | $[0.01, 0.9999]$ |
| | | Training seeds | 3 |
| **Likelihood evaluation** | | **Sampling** | |
| Weights | EMA | Samplers | ancestral, re-corrupt |
| Split | test | Instances | 4,096 test, one attempt each |
| Batches | 250 of 256 sequences | Time grid | uniform |
| Units | bits/token, full bound | NFE | $\{8, 16, \ldots, 512\}$ |
| | | Temperature, top-$p$ | 1, 1 |

| | Sudoku | Boxoban | Maze | Graph coloring |
|---|---|---|---|---|
| Network | DiT | DiT | DiT, grid encoder | graph DiT |
| Parameters | 42.7M | 42.7M | 61.6M | 42.8M |
| Sequence length | 81 | $100 + 128$ | $1{,}600 + 192$ | 64 |
| Solution vocabulary | 9 | 5 | 5 | 4 |
| Training instances | 3.8M | 936k | 500k | fresh draws |
| Learning rate, constant | $10^{-4}$ | $3 \cdot 10^{-4}$ | $10^{-4}$ | $10^{-4}$ |
| Steps | 1M | 2.5M | 400k | 300k |

are bits/token throughout; $\log V = 10.82$ nats $= 15.62$ bits at $V = 50{,}257$. All reported NELBOs use the EMA weights, except Figure 3 (top), which plots the same quantity for the online weights at every validation step, including step 0.

Table 10: Size of the two boundary terms of the finite-window NELBO at the training window $[0.001, 0.9999]$, in units of $10^{-3}$ bits/token. The prior term is a property of the noising process and does not depend on the head; the reconstruction term is the endpoint decode and does.

| | Prior KL | Reconstruction | | |
|---|---|---|---|---|
| Kernel | | den. | cav. | score |
| Masked | 1.562 | =cav. | 1.66 | 1.73 |
| GIDD $\lambda$=0.1 | 0.435 | 2.63 | 1.96 | 2.13 |
| GIDD $\lambda$=0.2 | 0.345 | 3.12 | 2.09 | 2.33 |
| Uniform | 0.166 | 5.19 | 2.67 | 3.96 |

**Sample quality metrics.** Generative perplexity (GenPPL) is the exponentiated mean per-token NLL of generated samples under GPT-2 Large (Radford et al., 2019). Word entropy is the histogram entropy of the whitespace-split words of a sample, averaged over samples, the word-level form of the token-histogram entropy of Zheng et al. (2025); Sahoo et al. (2025); the test set's value at $L = 256$ is 4.625 nats. MAUVE (Pillutla et al., 2021) compares the sample and test distributions in a quantized GPT-2 Large feature space, with 2,048 samples against 2,048 held-out test sequences and the default bucket count $n/10 \approx 205$, averaged over five quantization seeds. Text samples are drawn at temperature 1 and top-$p$ 1 unless a sweep varies them, on a cosine grid in $t$. Every cell at the reference budget, NFE 256 at any temperature, natively or misread, is drawn at three sampling seeds of 2,048 samples and reported as the seed mean; the NFE grid uses one seed. At the reference operating point the spread over the three sampling seeds is about 0.01 in MAUVE and 1% in GenPPL. The temperature sweep of Figure 3 (bottom) covers temperatures from 0.5 to 1.3 at top-$p$ 1.

**Head agreement comparison.** Two heads are compared on shared corrupted batches $(z_0, z_t)$ from the test split: both are converted to the denoiser coordinate and scored by the per-token Jensen–Shannon divergence $\frac{1}{2}\big[\mathrm{KL}\big(\pi_i^A \,\big\|\, \bar{\pi}\big) + \mathrm{KL}\big(\pi_i^B \,\big\|\, \bar{\pi}\big)\big]$ with $\bar{\pi} = \frac{1}{2}(\pi_i^A + \pi_i^B)$, over the content vocabulary, in nats/token and bounded by $\log 2$. The statistic is averaged over the positions whose denoiser is not analytically a point mass (under masked noise, the masked ones) and either over $t$ drawn per sequence from the training time proposal over the full $51{,}200$ test sequences (Table 3) or at fixed $t$ over $6{,}400$ sequences per point (Figure 7). The misread entries are the same statistic between a run's converted $\pi_i^\theta$ and its own raw head read in the other coordinate.

**Own-token sensitivity.** The sensitivity columns of Table 3 and Figure 8 use the batches, seeds, time proposal, and grid of the head agreement comparison. After corrupting a batch to $z_t$, one position $i$ per sequence is redrawn from the forward kernel $q^i_{t|0}(\cdot \mid z_0^i)$ at the same $t$, giving $z'_t$ that differs from $z_t$ at most at $i$; the statistic is the Jensen–Shannon divergence at $i$ between the head's law at $z_t$ and at $z'_t$, read in either coordinate through Proposition 3, averaged over the positions where the redraw changed the token. Masked diffusion is excluded: its denoiser equals the cavity at masked positions (Corollary 3) and is a point mass on the observed token at unmasked ones, so the statistic is fixed analytically.

### G.4 Further experiments

Table 11 lists the endpoint numbers of every run (§5.3): denoiser and cavity are close to each other in NELBO and overlap in GenPPL and MAUVE, while the score trails behind and has the highest GenPPL on every kernel, all of them at a similar word entropy (within $0.03$ nats per kernel).

Table 11: Endpoint numbers of the 11 runs at 500k steps: test NELBO, with GenPPL, word entropy, and MAUVE at the reference operating point.

| | NELBO (bits/token) | | | GenPPL | | | Word entropy | | | MAUVE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Kernel | den. | cav. | score | den. | cav. | score | den. | cav. | score | den. | cav. | score |
| Masked | =cav. | **4.900** | 4.999 | =cav. | **108.4** | 115.8 | =cav. | 4.720 | 4.713 | =cav. | **0.869** | 0.866 |
| GIDD $\lambda$=0.1 | **4.976** | 4.990 | 5.115 | **78.2** | 78.8 | 91.0 | 4.696 | 4.688 | 4.704 | **0.891** | 0.884 | 0.870 |
| GIDD $\lambda$=0.2 | **4.999** | 5.017 | 5.132 | **84.3** | 86.7 | 107.2 | 4.700 | 4.697 | 4.717 | 0.883 | **0.896** | 0.860 |
| Uniform | 5.081 | **5.070** | 5.257 | **101.7** | 105.0 | 152.8 | 4.712 | 4.714 | 4.738 | **0.887** | 0.873 | 0.814 |

**Misreading in both directions.** Table 12 extends Table 1 with GenPPL and the reverse misread, a denoiser head read as a cavity, which receives the Bayes reweighting a second time during sampling.

Table 12: Each run evaluated natively and with its head read in the wrong coordinate.

| | Trained | Read | NELBO (bits/token) | | GenPPL | | MAUVE | |
|---|---|---|---|---|---|---|---|---|
| Kernel | as | as | Native | Misread | Native | Misread | Native | Misread |
| Masked | cav. | den. | 4.900 | =native | 108.4 | =native | 0.869 | =native |
| GIDD $\lambda$=0.1 | cav. | den. | 4.990 | 12.416 | 78.8 | 3573 | 0.884 | 0.008 |
| GIDD $\lambda$=0.2 | cav. | den. | 5.017 | 12.289 | 86.7 | 3987 | 0.896 | 0.007 |
| Uniform | cav. | den. | 5.070 | 11.788 | 105.0 | 2797 | 0.873 | 0.011 |
| GIDD $\lambda$=0.1 | den. | cav. | 4.976 | 6.568 | 78.2 | 218 | 0.891 | 0.777 |
| GIDD $\lambda$=0.2 | den. | cav. | 4.999 | 6.905 | 84.3 | 290 | 0.883 | 0.679 |
| Uniform | den. | cav. | 5.081 | 8.155 | 101.7 | 677 | 0.887 | 0.266 |

**Agreement in time.** Figure 7 resolves the agreement columns of Table 3 in diffusion time.

**Own-token sensitivity in time.** Figure 8 resolves the sensitivity columns of Table 3 in diffusion time. In the denoiser coordinate the two heads are indistinguishable at every $t$: the denoiser's own-token dependence is the Bayes reweighting of the cavity's throughout the window. In the cavity coordinate the cavity head sits below the reweighting-corrected denoiser at every $t$, both roughly flat over the interior of the window. All four curves peak in the interior and vanish as $t \uparrow 1$, where the context is noise and a redrawn token changes nothing; at the clean end a single redrawn token is outweighed by an almost clean context. What remains of a denoiser once the reweighting is divided

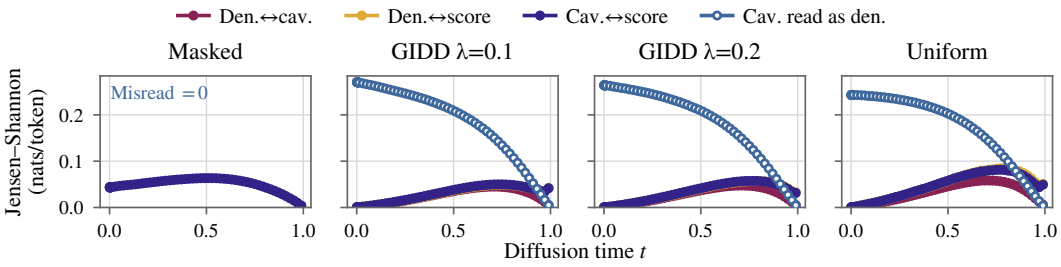


Figure 7: Pairwise Jensen–Shannon divergence between the independently trained heads of each kernel, converted to the denoiser coordinate, against diffusion time (filled markers). The open markers compare each cavity head against itself read as a denoiser without the conversion, a disagreement that grows as $t \to 0$. Masked has no denoiser run.

out is about twice the cavity's own residual, so the denoiser learns the Bayes factor somewhat less exactly than the cavity learns to discard its token.

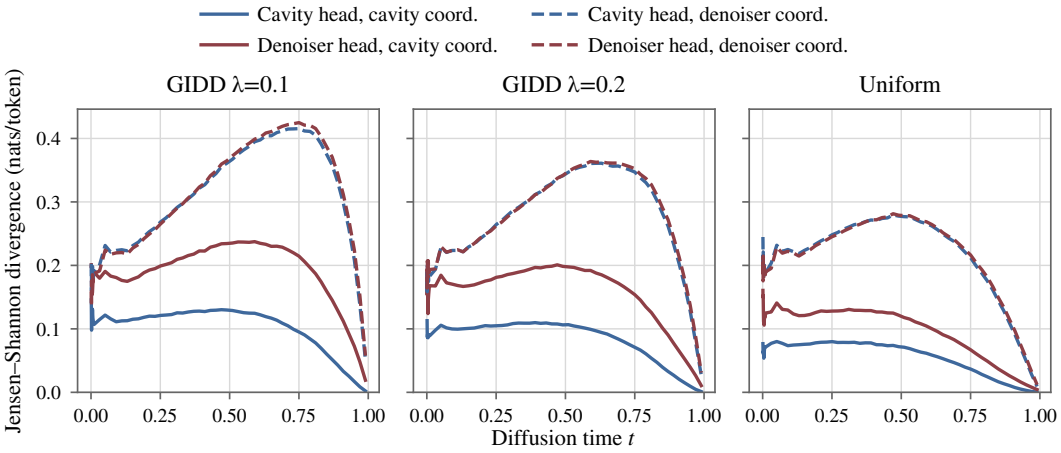


Figure 8: Own-token sensitivity against diffusion time: Jensen–Shannon divergence between a head's law before and after its noisy token is redrawn, for the cavity and denoiser heads read in the cavity coordinate (solid) and the denoiser coordinate (dashed).

**Sampling sweeps.** Figure 9 traces the NFE sweep $\{16, \ldots, 1024\}$ at temperature 1 and top-$p$ 1 for every head, natively and read in the wrong coordinate: every misread series is dominated by its native counterpart at every budget. Figure 10 shows the temperature frontiers of Figure 3 in GenPPL.

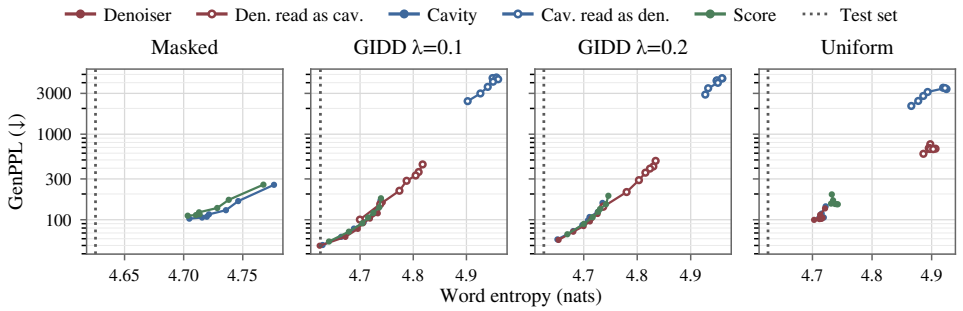


Figure 9: Quality–entropy frontiers traced by the NFE sweep $\{16, \ldots, 1024\}$: GenPPL under GPT-2 Large against word entropy, per kernel, for every head in its native coordinate (filled markers) and read in the wrong one (open markers); the dotted line is the test-set word entropy.

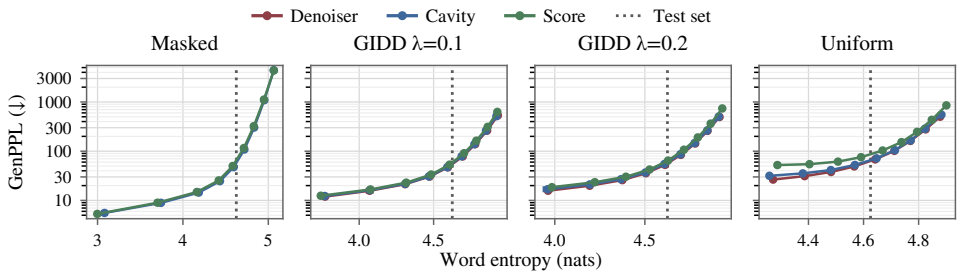


Figure 10: GenPPL counterpart of Figure 3 (bottom): GenPPL under GPT-2 Large against word entropy along the temperature sweep at NFE 256, per kernel. The dotted line is the test-set word entropy.

## H Extended related work

Discrete diffusion originates with structured categorical corruption (Hoogeboom et al., 2021; Austin et al., 2021), the discrete analogue of Gaussian diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020), and now scales to large language models (Nie et al., 2025).

**Path-KL identities.** For SDE diffusions, Song et al. (2021) bound the likelihood by a weighted score-matching objective whose gap is a Girsanov path KL, and Huang et al. (2021) derive the continuous-time ELBO variationally. Benton et al. (2024) give a denoising framework and ELBO for general Markov processes; Ren et al. (2025) develop a stochastic-integral calculus for discrete diffusion with CTMC change-of-measure theorems and error bounds for $\tau$-leaping; Chen & Ying (2024) prove sampling guarantees for uniformization with the score-estimation error entering as an assumed loss. Theorem 1 uses the same $\Phi$ integrand as an exact decomposition of the ELBO, with the oracle reverse process as reference, for arbitrary noising processes, with the finite-window boundary form in Theorem C.1. A tutorial treatment of diffusion in general state spaces that considers similar path-KL identities is Pauline et al. (2025).

**Weights and clocks.** Kingma & Gao (2023) show that common continuous-diffusion losses are ELBOs under noise-level distributions when the weighting is monotone; the clock equivalence of §3 is the CTMC counterpart, and separates unbiased importance densities from bound-changing reweightings (Nichol & Dhariwal, 2021; Shi & Titsias, 2025; Raya et al., 2026). The masked cross-entropy at a uniform mask rate is ARDM's any-order objective (Hoogeboom et al., 2022) and LLaDA's estimator (Nie et al., 2025).

**Masked diffusion.** Simplified masked objectives are due to Sahoo et al. (2024); Shi et al. (2024); Zheng et al. (2024). Zheng et al. (2025) show that masked training and sampling do not depend on the time variable. This is the flat-Bayes-factor collapse of Corollary 3 and (E.6): for a fixed context the masked cavity does not depend on time. RADD (Ou et al., 2025) names the law of the clean token given the unmasked context, which does not depend on time, and derives the exact per-position reverse step from it.

**Score-based discrete diffusion.** Concrete scores originate with Meng et al. (2022); Sun et al. (2023), and SEDD (Lou et al., 2024) trains them by denoising score entropy, which is (E.7) with the forward rate as its weight. MD4's score parameterization enforces the constraint surface that a free score head can leave (Shi et al., 2024), which our conversions parameterize. Target Concrete Score Matching (Zhang et al., 2025) unifies objectives through the target score. For uniform diffusion, Sun et al. (2023) parameterize the conditional of each noisy token given the others as the cavity pushed through the forward kernel, $m_i^\theta = \mu_i^\theta \, q_{t|0}^i$ in the notation of Proposition 3, read their ratios off it, and sample with the corresponding plug-in step. They train it by ratio matching, a cross-entropy on each noisy token given the others, which a network that sees the token would minimize by copying it, so they have to exclude each position's own token with a hollow transformer. The ELBO needs no such restriction: its minimizer is the cavity for every token-factorizing kernel, so the plug-in models, built on standard transformers, learn the cavity directly.

**Plug-in lineage.** D3PM (Austin et al., 2021) introduced the $x_0$-parameterization and the posterior-sum sampler, algebraically the plug-in step (§4); UDLM (Schiff et al., 2025), GIDD (von Rütte et al., 2025), Duo (Sahoo et al., 2025), and the scaled non-masked models of Sahoo et al. (2026) train it on uniform and GIDD kernels, where its optimum is the cavity law. Both steps appear in D3PM's own formulation. Writing $\tilde{p}^\theta$ for its $x_0$-head, its reverse step is $p^\theta(Z_s^i{=}y \mid z_t) \propto q^i_{t|s}(z_t^i \mid y) \sum_{z_0^i} \tilde{p}^\theta(z_0^i \mid z_t)\, q^i_{s|0}(y \mid z_0^i)$, an average of joints: its summand is $q^i_{s|0}(y \mid z_0^i)\, q^i_{t|s}(z_t^i \mid y) = q^i_{t|0}(z_t^i \mid z_0^i)\, q^i_{s|t,0}(y \mid z_t^i, z_0^i)$, the bridge weighted by the Bayes factor, so normalizing by $m_i^\theta(z_t^i \mid z_t^{-i}, t)$ recovers the cavity step (D.9); averaging the bridge alone, as their text describes, gives the denoiser step (D.7). The two are introduced as one parameterization, yet they optimize different laws, equal only where the Bayes factor is flat (Corollary 3), that is, on their absorbing kernels. Their auxiliary cross-entropy $\mathbb{E}\left[-\log \tilde{p}^\theta(z_0^i \mid z_t)\right]$ therefore trains the denoiser while the plug-in loss and sampler assume the cavity, which would explain why it helps their absorbing models and reduces performance on their uniform and nearest-neighbor ones, the double application measured in §5.1. Gourevitch et al. (2026) are discussed in §6.

**Flows, generators, and other frameworks.** Discrete flow matching (Gat et al., 2024; Campbell et al., 2024) regresses probability velocities along prescribed corruption paths; Shaul et al. (2025) decouple the path from the generating velocity and characterize kinetic-optimal choices, a freedom that in the likelihood picture appears as the coordinate and clock invariances of one optimal rate. Generator Matching (Holderrieth et al., 2025) regresses the posterior-averaged marginal generator with Bregman losses; §3 derives that average as the unique optimizer of the likelihood itself. Amin et al. (2025) explain the empirical differences between kernels through conditioning on the jump schedule, complementary to the coordinate mechanism here and to the shared floor of Corollary 2.

**Information-theoretic identities.** I-MMSE ties information to estimation in the Gaussian channel (Guo et al., 2005); its Poisson analogue has $\Phi$ as the natural loss (Atar & Weissman, 2012); Kong et al. (2023) write the continuous diffusion likelihood as an exact integral of MMSE, decomposed per variable by Kong et al. (2024). The discrete I-MDSE/I-MDCE identities of Jeon et al. (2025) state the oracle case with $T \to \infty$ in the score and masked cross-entropy coordinates; Proposition 2 with (7) extends them to arbitrary model rates, finite windows with boundary terms, and every coordinate. Conditional-mean optimality and the Pythagorean identity for Bregman divergences are due to Banerjee et al. (2005a), with the exponential-family geometry in Banerjee et al. (2005b); Vincent (2011) is the continuous denoiser/score identity our conversions mirror.

# I GENERATED SAMPLES

One sample from each run, drawn uniformly at random from the 2,048 generated, from the 500k-step EMA checkpoints, sampled with 256 NFEs at temperature 0.9.

**Masked, cavity.**

> told his driver to file a report accordingly. An arrest and from a prosecutor's office in court heard Mr Mitchell could go to trial next month for commingling relating to the raid. In the programme, one revelation is that Mr Begg known and realised his wrongdoing after he was told he could not get out of his positions. "He's the one covering it up with the information," he told the court. But on 7 October a 21-year-old Begg man was arrested. The case was closed when for a time there was "shame and sorrow". There are rumours Begg lied to investigators and witnesses. "He has no influence on the case from a rented home in Cambridgeshire," said Assistant Information Commissioner Sebastian Aldatts. The prosecutors were involved in a recent High Court-flung case hearing for DIA as it investigated a British-made helicopter in September 2012. It was the first military helicopter crashed on UK-controlled territory and since last year the British air force was operating in Afghanistan. It was the first time a helicopter had been used in a military attack in New Zealand. After the 2012 raid, the British Air Force launched its

**Masked, score.**

> been linked to a dramatic penalty goal near QPR Park on Sunday during the campaign in which they, who miss the players in the second round, were impressed by their side When

Rodgers aired his bad play on Sports 4 TV when the game was in July, he said a ban kept him from the club. But it is understood QPR are exceeding the £48 million budget while City are not. The club are concerned and Rodgers is under the orders of Chris Wilson. Finishing with his role, Rodgers responded to the media: "I have 30 days to be there in training, I have two days to train and you can't second what I have done. "What's the problem? We are not getting Tom's cut. It's his own fault. We have a terrible line that gets him goal but that is not all it is. We need to tell you that you need to do a lot of things, you've got some cap space and we learn. "I move like this so it is very important for me as a player, we learn. "If Tom is training then he will be doing well ahead of us. "The most we ever spend

**GIDD $\lambda$=0.1, denoiser.**

or an incurable infection. The serotonin role is what sticks around to wake you and won't get to you. When you feel sick, you have limited information. That mechanism requires the release of the strong neurotransmitter serotonin stored in your prefrontal cortex to help you survive medical emergencies. You actually want it to go down on top of you, to talk to your brain. "It doesn't necessarily cause things like allergies to deteriorate and fall out," Dodging said. But compared with the alcohol or crack cocaine you take and makes your brain nervous mate, "this doesn't jeopardize the science or anything. You don't even have a scare by going after serotonin while you are suffering. The studies that have linked it would have said they risk it. It does lock up some of the protein, too, to process this. And some of what was supposed to happen to serotonin, did not happen." Helmet tablets, along with the new drugs, have a particular place for the internal nervous system. Things don't already work, he said. Going to use it, however, has a few limitations. First, several treatments are likely to be done before

**GIDD $\lambda$=0.1, cavity.**

, so I worked my way around, work my way back around" — Architectural Abstract Architecture What did you think of Airdropels and Airdropism? The concept of Airdropels is simple. In fact, they are the form of New Technologies technologies and, to a certain extent, the development of the form of government, as it is a new tool for manipulation and deception. They will rebuild and rebuild to strengthen their position, ultimately, they will change the way they vanquished causes of conflict, but they also will try to create rules, and if the rules create perpetual changes, then the they will. They say, "There is any history of time spent change and you've got the government spending trillions of dollars trying to avoid it. It happens because big corporations are creating laws. Laws that the U.S. Congress is coming up, which have nothing to do with being a government, because I believe the values encoded in the Constitution that we must do things properly, and that we are subject to principle, and history is a part of it. That makes me happy that as a civilization, Philosophy of Freedom is at the heart and is trying to lead us in the direction of this direction.

**GIDD $\lambda$=0.1, score.**

and the more liberal crowd would be too concerned about fair/fast, clean/money-action/full investment in criminal justice reform. When Sanders made it on stage, it was clear that the tactics of his campaign weren't going to work, and he was just sitting in the bus by the state line. And rather than accusing Clinton and Kevin Spacey, Sanders's Democratic friend, which had raised complaints about him, Clinton "could have let it all down like a bird" so she would stand by Clinton's "less policies " and that "don't have the money to keep our words off the air." Still, Sanders was prepared to engage with Clinton and articulate her message. And he feels he can define it better in his speeches Saturday night. "Yes, I'm on the side of PIPA," he said, "but I represent America. " " He will have much more to offer from Clinton's campaign money and celebrity money, with another governor, donors morale and opportunity to continue the business with his political base. Look, she is a passionate charity fighter who's got the kind of support Obama had in the campaign process, working quietly just a few months a year, and she certainly has accrued to the

**GIDD $\lambda$=0.2, denoiser.**

City. He'd secured himself a contract with LeBron and as we've seen before had dramatically increased his numbers for the Thunder. In at 8-1 against Detroit though he had missed just three straight games to keep his teammates out absent injury. Neither Westbrook nor

Kyle Orton even had a factor that game provided he was not on the floor, he would have been playing better than Oklahoma City, with 18 of 25 teams on their schedule beating Westbrook's suffocating game by 18 points. I can't pick these years, every one was that or even a half the work of LeBron. What didn't really care or much really care was the Thunder team went to 2-0 and did not disappoint in 2-0, earning two wins and winning a total of 3 games. What was there? One evening after the performance, leading to the 40-point TD/flash in Houston it's 7-0 opportunity to beat the team. All in all that they got the win and Dris beat them, so much like boo-ha and then "ha" beats. When LeBron, being rubbed out at the media, told me unprompted the media he paid attention to the rest

**GIDD $\lambda$=0.2, cavity.**

fully inspired, and a large audience like Long has more of that attention. "If it wasn't for Gears of War I can't imagine how many people say that those games will change gaming," he says. "We all watch great videogames, and one person goes, 'Well, then isn't the time.' But I think we'll never see all the heck of it." Sunday's video's rebrings on spoilers from the film, but you'll learn more about Wes' relationship from the wiki that reads the comic for him. It's just a fun video with just a pop-up in tone and lots of detail in how things are done, and how the art can be turned into a standalone MMO campaign for whatever purpose. I initially wasn't exactly excited with the idea of Wes as a JRPG...to-the-day. They were quite right, though I wasn't sure the visuals would be perfect (or at the very least, I still recall using a coworker's sports figure), but, what did you build on Wes' return as a nerd? The mix of playable animations with dialogue (I have not even played a dialogue for the sidequest game for that matter. There needs to be more and

**GIDD $\lambda$=0.2, score.**

. Times on Friday, with concern about "the guidance put in place to it easier for regulators to apply what is benefit gains to investors in the region." It said, for instance, the ICTG is willing to offer at least $30,000 shares to low-cost investors going back two years, and that the federal government would have an eager hand in exchange for ICTG's assurance that it could buy a franchise in a foreign country for $18. The newspaper also said: "We demanded that this happen. We can't blame them. It could end in bad views for us. We do not believe the regulators have given it their clear instructions about the rigging. Mr. Smith, who was contacted by Mr. Baker, told The Washington Post on Saturday that ICTG plans to open more coal mines in the Middle East than in the years to March 1, so people might try the black market, rather than retail. "I think it's such a complicated issue, we have to make a change, and... again, I'm very concerned about what it is," he said. He cited The Huffington magazine's Crispin Pelt, and his work with Justice Johnson, to be the source. Despite the concerns, he told Fox News,

**Uniform, denoiser.**

of every year Johnny Manziel's coaching job along with several other 49ers, Saints, and Broncos players, it's a very intricate thought process. In that one thing, Ryan and Murray even left it all too late in another direction as well: So, we are yapping today and are focusing on the topics these next ten days. Find your team on the field we're doing, and you can head back to practice field when Coach Murray posts an announcement via Facebook. Every now, as we take the schedule bashing at our best, please CONNECT! For the holiday, Being in a New Year, here is a special series on practice fields, focusing again. Without further ado, we evaluate the top three teams today. We chose the following preseason schedule: the first NFL Combine, held in Iowa, the first, but not the nation's great Combine, in May, and the initial Pacific Player of the Year in June. "Um, uh, okay? Really? Coach?!" Manziel nodded at me with a freaking grin and said, "Fuck that!" So freaking awesome! We picked the first NFL-firing Cedar Falls on July 1. The NFL

**Uniform, cavity.**

's movie studio() adopted baselina analog features in addition to converting to MPEG format as direct-to-feedback, but they also built heavily on the analysis of original time frames in the 4D digital format. This took off as a wise decision, as the "green" philosophy of the time was logical enough to suppress some original episodes from the making. In an example, the 3D disc featured four ;mins of broadcasts from the viewing ramp on morning 1, while someone listened to Marshall Brownock's Morning 2 (middle door). Just what

did this pattern mean? For the most part, I remember Scott: due to the escalating power of projector technology one at a time, Scott vividly took a CD, read dozens of photocopies, recorded by Joel Entrance, took the tape into his revisit cabinet, and then listened to phonos from before he passed away. The use of abeforms, to separated reads, gave these an ex post facto intellectual continuity over many clips from different shows in the movie. In 1987, TMI was adapted to a multimedia format In 2004, after discussions with Tim Nachows and Stephen North, the Edison Broadcasting Corporation (ICC) decided to adopt the MPEG/P/TTV format as a

**Uniform, score.**

after a year (AC4/1199, the APSS Design) and for select users, like the Sony Xperia A10 and the LG A township Shoot, you is the number two RD battery/hot. 15 THE R3 1900 The United States R3, which starts with a Galaxybird (US for the arc phone and desktop device) cheap and also is actually is compatible with the Ace as well well as a 1028 and an SDATA/STAT port. 16 THE SARM (8 years) In case you pond, I will be reading for you to ask to see this ride back. To my knowledge, my request increased (Monica Perry declined to provide this one) to S co (Earthshake, only to be issued) at the fee. 19 Thunderbacks — presented by Pat Colby at Sheffield University this week. Talk about therapies endorsed by the medical, scientific professions — and just that, go far back and time to present. This week I'll follow along some of the topics of one of these four experiments hosted by the Crossroads's Independent Submarine/FCC. Several of the parameters examined are 3 days during the course of