MEAN-TO-SCORE DISCRETE DIFFUSION: POSTERIOR-MEAN DENOISERS FOR SCORE ENTROPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Score Entropy Discrete Diffusion (SEDD) parameterizes discrete reverse processes with unconstrained positive score ratios. Although positivity ensures nonnegative reverse jump rates, it does not ensure Bayes realizability: at a fixed noisy state, all candidate score ratios must arise jointly from a single clean-token posterior under the forward kernel. We show that the score-entropy loss has the correct population optimum but does not enforce Bayes realizability away from it. In a trained pure-uniform SEDD checkpoint, roughly one quarter of complete score vectors violate the coordinate box, while more than half satisfy every coordinate bound but remain materially incompatible with any valid clean-token posterior. Although the corresponding continuous-time reverse jump rates remain nonnegative, these violations can induce negative pre-normalization weights in the finite-step sampler update. Projecting the checkpoint's raw scores onto the bridge polytope removes all observed negative weights and lowers external generative PPL from 203.6 to 175.1 without changing the sampler. To enforce Bayes realizability by construction rather than through post-hoc projection, we introduce *mean-to-score* (M2S): the network predicts a clean-token posterior mean and converts it to the score through an exact kernel-dependent linear map. The map applies to any known coordinate-wise continuous-time Markov chain (CTMC) satisfying a mild support condition. For uniform corruption, it maps the probability simplex onto the bridge polytope; for absorbing-mask corruption, the resulting objective recovers MD4 exactly. In a controlled 28.4M-parameter CIFAR-10 comparison, M2S lowers test BPD from 3.173 to 3.129 and FID-50k from 42.83 to 28.09. A 170M-parameter M2S model trained on approximately 131B OpenWebText token slots outperforms the evaluated pure-uniform SEDD, GIDD, and Neural CTMC checkpoints at every tested sampling budget, reaching mean generative PPL 146.5 at 128 steps compared with 189.2 for the strongest pure-uniform baseline.

1 INTRODUCTION

Discrete diffusion models generate finite-valued data by reversing a continuous-time Markov chain (CTMC) that progressively corrupts a sample (Austin et al., 2021; Campbell et al., 2022). We study pure-uniform corruption, which treats vocabulary states symmetrically without an absorbing mask token. For site i, let X_0^i denote the clean token and $\mu_i^\star(x,t) := \Pr(X_0^i = \cdot \mid X_t = x) \in \Delta^{K-1}$ its posterior given the noisy sequence $X_t = x$, where Δ^{K-1} is the probability simplex over a vocabulary of size K. Unlike absorbing-mask training, which reduces to a weighted clean-token cross-entropy on marked positions, uniform corruption does not reveal which positions changed and requires the reverse model to coordinate transitions across the full vocabulary (Sahoo et al., 2024; Shi et al., 2024).

MD4 (Shi et al., 2024) already highlights a closely related parameterization issue for absorbing-mask diffusion: a freely parameterized score need not be induced by the conditional clean-token mean under the forward process. MD4 therefore predicts this mean and constructs the masked reverse model from it. We ask a vector-level question: whether one valid clean-token posterior induces the complete concrete-score vector predicted by SEDD. We call this property *Bayes realizability*. It reduces to MD4's score–mean constraint for an absorbing mask, but under pure-uniform corruption it couples all $K-1$ scores.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

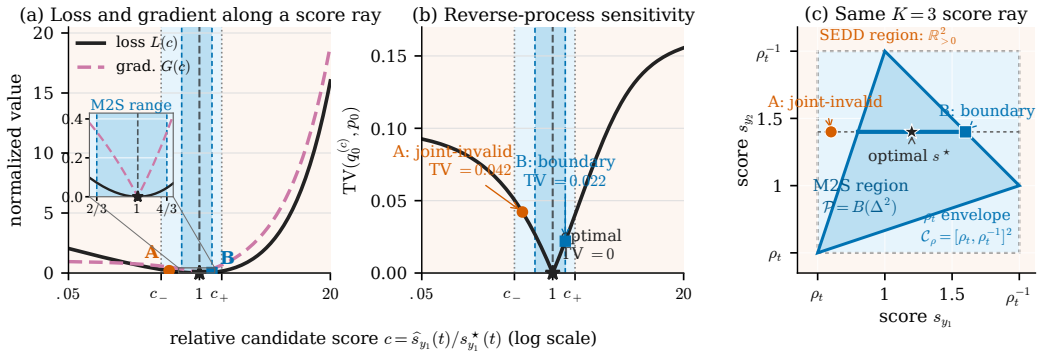


Figure 1: **M2S occupies the Bayes-realizable subset of the scalar envelope, whereas SEDD can output any positive score.** In a $K = 3$ uniform toy, panels (a)–(b) vary one score along $\widehat{s}(c) = (cs^\star_{y_1}, s^\star_{y_2})$ around the shared optimum $c = 1$. The labels c_- and c_+ mark the limits imposed by the ρ_t-based coordinate envelope, while the inner dark-blue band is the smaller Bayes-realizable subset enforced by M2S and enlarged in the inset. Panel (a) shows the score-entropy excess and log-score gradient; panel (b) shows the terminal reverse-process TV error; and panel (c) shows the joint score region. Point A passes every scalar bound but admits no valid joint posterior, while point B lies on the M2S boundary.

In score-based CTMC models, each reverse jump rate uses a concrete score ratio $p_t(x^{i\to y})/p_t(x)$. SEDD (Lou et al., 2024) predicts each ratio with an exponentiated network output, guaranteeing positive scores and nonnegative continuous-time reverse jump rates. **Positivity alone, however, does not ensure that all candidate score ratios arise from one clean-token posterior.**

Formally, at noisy state x and site i with $k = x^i$, a candidate-score vector $s = (s_y)_{y\neq k}$ satisfies Bayes realizability if $s = B_{t,k}\mu$ for some posterior $\mu \in \Delta^{K-1}$. Under uniform corruption, any score induced by a clean-token posterior must lie in the coordinate box $\mathcal{C}_{t,k} := [\rho_t, \rho_t^{-1}]^{K-1}$. This restriction concerns Bayes realizability, not requirement for nonnegative continuous-time reverse rates, which only require $s_y > 0$. The coordinate box is necessary but not sufficient for Bayes realizability: the complete score vector must lie in the strictly smaller bridge polytope $\mathcal{P}_{t,k} = B_{t,k}(\Delta^{K-1}) \subsetneq \mathcal{C}_{t,k}$. Thus, constraining a SEDD head to positive outputs is insufficient to ensure Bayes realizability: its output may lie outside both the bridge polytope and even the coordinate box. Score entropy therefore has the correct population optimum but does not enforce Bayes realizability away from that optimum. Figure 1 visualizes the distinction for $K = 3$. The coordinate-box interval strictly contains its intersection with the bridge polytope, and panel (c) shows $\mathcal{P}_{t,k} \subsetneq \mathcal{C}_{t,k} \subsetneq \mathbb{R}^2_{>0}$. Point A passes every coordinate bound but has a negative signed inverse component.

Failure of Bayes realizability has measurable consequences at language scale. In a pure-uniform SEDD checkpoint, 6.69% of 3.29 billion audited candidate scores leave the coordinate box. To test Bayes realizability during generation, we run this checkpoint with SEDD's released sampler on 128 sequences and audit the unmodified score vector at every position and sampling step. Of these vectors, 25.02% contain a coordinate violation, while another 56.23% pass every scalar check but admit no valid joint posterior. During these runs, the sampler update encounters negative pre-normalization weights and operationally samples from their positive part. To isolate the effect, we project the scores onto $\mathcal{P}_{t,k}$ for 1,024 paired sequences while holding the time grid, initial states, random-number stream, final denoising, and sampler code fixed. Projection removes all observed negative weights and lowers external generative PPL from 203.60 to 175.07.

Rather than project scores at inference time, we enforce Bayes realizability in the model. For any known coordinate-wise forward kernel satisfying a mild support condition, the one-site clean-token posterior recovers every concrete score:

$$s^\star_i(x,t;y) = \mathbb{E}\left[\frac{P_t^{(i)}(y \mid X_0^i)}{P_t^{(i)}(x^i \mid X_0^i)} \,\middle|\, X_t = x\right], \tag{1}$$

where the integrand depends only on the clean token at site i. Mean-to-score (M2S) predicts $\mu^i_\theta(x_t,t) = \mathrm{softmax}(x^i_\theta(x_t,t))$ and applies the known linear map B_t to obtain the score. Because μ^i_θ

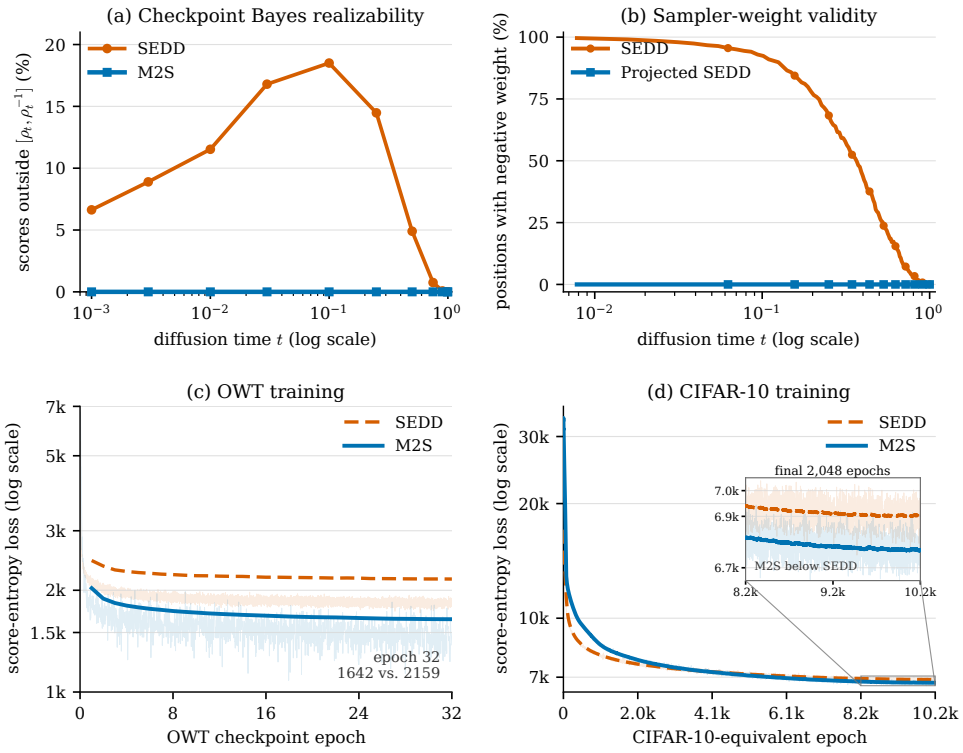


Figure 2: **M2S enforces Bayes realizability and achieves lower final loss on both text and image generation.** Panels (a)–(b) audit score realizability and sampler-weight validity. Panels (c)–(d) compare the score-entropy loss of M2S and SEDD on OpenWebText and CIFAR-10, respectively.

is a distribution, M2S enforces Bayes realizability. It changes the admissible off-optimum geometry, not the population target.

The bridge uses only a one-site posterior, not a posterior over the full sequence. For uniform corruption it is injective on the simplex, has a closed form, and evaluates all scores in $O(K)$ time per site. For absorbing-mask corruption, the same construction recovers the weighted clean-token cross-entropy of MD4 exactly. We also show that the conditional score-entropy risk is uniquely minimized at the true score (Proposition 4.3) and derive the rank condition $\mathrm{rank}([B_+^\top, \mathbf{1}]^\top) = K$ under which this optimum uniquely recovers the clean-token posterior (Theorem 4.4); pure-uniform corruption satisfies this condition.

Our experiments ask whether enforcing Bayes realizability matters for image and language generation. On 256-state MNIST, we compare M2S with SEDD: the two models share the architecture, forward process, objective, training budget, and sampler, and M2S improves FID from 126.1 ± 0.4 to 71.1 ± 4.3. On CIFAR-10, an augmentation-free 28.4M-parameter M2S checkpoint reaches a test BPD upper bound of 3.129, compared with 3.173 for SEDD under the same evaluation protocol; under the same 256-step Euler sampler, M2S also lowers FID-50k from 42.83 to 28.09. On OpenWebText (Gokaslan & Cohen, 2019), our 169.9M-parameter M2S model reaches mean generative PPL 146.5 at 128 steps and outperforms the evaluated pure-uniform SEDD (Lou et al., 2024), GIDD (von Rütte et al., 2025), and Neural CTMC (Li et al., 2026) baselines at all tested budgets. Figure 2 further compares M2S with pure-uniform SEDD through Bayes realizability and score-entropy audits, and adds a CIFAR-10 optimization trace under identical compute. Keeping the SEDD checkpoint fixed, we also project its scores onto $\mathcal{P}_{t,k}$ during sampling using the simplex-constrained Euclidean projection in Eq. (11), lowering external GenPPL from 203.60 to 175.07 (Section 5.1).

Contributions.

1. **We introduce joint Bayes realizability for discrete score vectors.** A complete score vector satisfies Bayes realizability exactly when it is induced by one valid clean-token posterior. Positivity and even coordinate-wise feasibility do not suffice. Under uniform corruption, such vectors form a strict bridge polytope $\mathcal{P}_{t,k} \subsetneq \mathcal{C}_{t,k}$, and score entropy does not enforce this constraint away from its population optimum.
2. **We derive M2S to enforce Bayes realizability.** M2S maps a one-site clean-token posterior to all concrete scores through an exact kernel-dependent linear bridge. We establish score consistency and posterior recovery, derive the uniform-kernel form, and recover the MD4 objective under absorbing-mask corruption.
3. **We show empirically that Bayes realizability matters.** For a fixed pure-uniform SEDD checkpoint, simplex-constrained Euclidean projection removes negative sampler weights and lowers generative PPL from 203.6 to 175.1. On MNIST and CIFAR-10, M2S improves FID over SEDD under identical settings, while a 170M-parameter OpenWebText model outperforms the evaluated pure-uniform SEDD, GIDD, and Neural CTMC baselines at every tested sampling budget.

2 Related Work

Discrete diffusion. Diffusion probabilistic models originate from progressively corrupting Markov chains (Sohl-Dickstein et al., 2015). Discrete variants use multinomial, structured, or absorbing transition kernels (Hoogeboom et al., 2021; Austin et al., 2021; Bond-Taylor et al., 2022; He et al., 2023; Zheng et al., 2024). Continuous-time formulations learn reverse CTMC rates or more general denoising Markov dynamics (Campbell et al., 2022; Sun et al., 2023; Benton et al., 2024). Discrete score learning includes concrete and target-concrete score matching (Meng et al., 2022; Zhang et al., 2025), while SEDD learns marginal probability ratios with score entropy (Lou et al., 2024). Recent scalable systems use score, clean-token denoiser, interpolating-kernel, or factorized-rate parameterizations (Sahoo et al., 2024; Shi et al., 2024; Ou et al., 2025; von Rütte et al., 2025; Li et al., 2026). Concurrent work studies what discrete diffusion models learn under different parameterizations (Casado Noguerales et al., 2026). M2S retains SEDD's score-entropy CTMC and changes how the complete score vector is parameterized.

Posterior and score parameterizations. Posterior parameterizations construct reverse dynamics from clean-token predictions, as in D3PM and subsequent reparameterized or absorbing models (Austin et al., 2021; He et al., 2023; Zheng et al., 2024). For absorbing corruption, RADD factors the concrete score through conditional clean-data probabilities (Ou et al., 2025), while MDLM and MD4 reduce their objectives to weighted clean-token cross-entropy and MD4 highlights score–mean consistency (Sahoo et al., 2024; Shi et al., 2024). GIDD interpolates between masked and uniform corruption (von Rütte et al., 2025), and concurrent work derives exact coordinate-level score–denoiser conversions for uniform diffusion (Gourevitch et al., 2026). These works characterize individual score coordinates or population objectives. M2S instead asks whether the complete SEDD score vector is induced by one clean-token posterior, audits this joint condition in trained checkpoints, and enforces it by construction.

3 Preliminaries

Let $\mathcal{V} = \{1, \dots, K\}$ be a finite state space. A time-inhomogeneous CTMC on $\mathcal{V}$ over $[0, T]$ is specified by a rate matrix Q_t satisfying $Q_t(a, b) \geq 0$ for $a \neq b$ and $\sum_b Q_t(a, b) = 0$.

Definition 3.1. For some start time s and end time $t = s + \Delta$ ($t > s$) as $\Delta \to 0$, we have

$$q_{t|s}(b \mid a) = \delta_{a,b} + Q_t(a, b)\Delta + o(\Delta), \tag{2}$$

where Q_t is called the forward transition rate.

The Kolmogorov forward equation $\frac{\mathrm{d}}{\mathrm{d}t} q_t = q_t Q_t$ determines the finite-time transition kernel $P_{t|s}$ (Campbell et al., 2022). For the corruption processes considered here, we use the kernel in Definition 3.2, which covers both uniform and absorbing corruption.

Definition 3.2. The cumulative transition probabilities of the CTMC are given by

$$q_{t|0}(a \mid z) = \mathrm{Cat}(a; P_t(z, \cdot)), \qquad P_t = \alpha_t I + \beta_t \mathbf{1}\pi^\top, \tag{3}$$

where $\alpha_t + \beta_t = 1$, with $\alpha_0 = 1$ and $\alpha_T = 0$, and π is a fixed distribution on $\mathcal{V}$. For uniform corruption, $\pi = \mathbf{1}/K$; for absorbing corruption, $\pi = e_m$, where e_m is the one-hot vector for the mask state m. Here $\mathbf{1}$ is the all-ones vector. As $t \to T$, $q_{t|0}(\cdot \mid z) \to \pi$ for every $z \in \mathcal{V}$, so the reference distribution is $p_{\mathrm{ref}} = \pi$.

For sequences, the forward process acts independently across sites. With site-wise kernels $P_t^{(i)}$ and rates $Q_t^{(i)}$, $q_t(x_t \mid x_0) = \prod_{i=1}^{L} P_t^{(i)}(x_t^i \mid x_0^i)$ and $p_t(x) = \sum_{x_0} p_{\mathrm{data}}(x_0) \prod_{j=1}^{L} P_t^{(j)}(x^j \mid x_0^j)$.

Definition 3.3. For a sequence state $x \in \mathcal{V}^L$ and a candidate token $y \neq x^i$, let $x^{i \to y}$ denote the sequence obtained by replacing site i of x with y. Whenever $p_t(x) > 0$, the concrete score and its corresponding reverse transition rate are

$$s_i^\star(x, t; y) = \frac{p_t(x^{i \to y})}{p_t(x)}, \qquad \overline{Q}_t^{(i)}(x^i, y \mid x) = Q_t^{(i)}(y, x^i)\, s_i^\star(x, t; y), \tag{4}$$

respectively (Campbell et al., 2022; Lou et al., 2024).

MDLM and GIDD use an x_0 parameterization, whereas SEDD directly predicts the concrete score in Definition 3.3. M2S instead predicts the site-wise clean posterior $\mu^i(x_t, t) \in \Delta^{K-1}$, with

$$[\mu^i(x_t, t)]_z = \Pr(X_0^i = z \mid X_t = x_t),$$

and maps it to the concrete score in Section 4.

4 METHODOLOGY

This section presents the M2S bridge, its training loss, and its connection to MD4. All proofs are provided in Appendix A.

Theorem 4.1. *Under Assumption A.1, for any site i, candidate token y, and noisy state x with $p_t(x) > 0$, let $\pi^\star(z) = \Pr(X_0^i = z \mid X_t = x)$. For every $z \in \mathrm{supp}(\pi^\star)$, we have $P_t^{(i)}(x^i \mid z) > 0$. Then*

$$s_i^\star(x, t; y) = \mathbb{E}\left[\frac{P_t^{(i)}(y \mid X_0^i)}{P_t^{(i)}(x^i \mid X_0^i)} \,\middle|\, X_t = x \right] = \sum_{z \in \mathrm{supp}(\pi^\star)} \pi^\star(z)\, \frac{P_t^{(i)}(y \mid z)}{P_t^{(i)}(x^i \mid z)}. \tag{5}$$

Replacing the exact posterior in Theorem 4.1 with the neural network prediction $\mu_\theta^i(x, t)$ defines the M2S score as

$$s_{\theta,i}(x, t; y) = \sum_{z=1}^{K} [\mu_\theta^i(x, t)]_z\, \frac{P_t^{(i)}(y \mid z)}{P_t^{(i)}(x^i \mid z)} =: (B\mu_\theta^i)_y. \tag{6}$$

For the uniform process, define the off-diagonal-to-diagonal kernel ratio as $\rho_t = (\beta_t/K)/(\alpha_t + \beta_t/K) \in (0, 1)$. At a noisy input x_t, Equation (6) then simplifies, for each candidate $y \neq x_t^i$, to

$$s_{\theta,i}(x_t, t; y) = 1 + (\rho_t - 1)\big[\mu_\theta^i(x_t, t)\big]_{x_t^i} + (\rho_t^{-1} - 1)\big[\mu_\theta^i(x_t, t)\big]_y. \tag{7}$$

All candidate scores in Equation (7) are computed in $O(K)$ time. We next introduce the training objective.

Theorem 4.2. *For the uniform CTMC with $\alpha_t = e^{-\sigma(t)}$ and $\dot\sigma(t) \geq 0$, let $x_0 \sim p_{\mathrm{data}}$, $t \sim \mathrm{Unif}[0, 1]$, and $x_t \sim q_t(\cdot \mid x_0)$. For $y \neq x_t^i$, define $r_i(x_0, x_t, t; y) = P_t^{(i)}(y \mid x_0^i)/P_t^{(i)}(x_t^i \mid x_0^i)$. For $s, r > 0$, define $h(s, r) = s - r \log s + r \log r - r = r\big(s/r - 1 - \log(s/r)\big) \geq 0$, with equality if and only if $s = r$, and extend this definition continuously to $r = 0$ by $h(s, 0) = s$. Define the M2S objective*

$$\mathcal{L}_{\mathrm{M2S}}(\theta) = \mathbb{E}_{t, x_0, x_t}\left[\sum_{i=1}^{L} \sum_{y \neq x_t^i} w_{t,i}(x_t^i, y)\, h\big((B\mu_\theta^i)_y, r_i(x_0, x_t, t; y)\big) \right], \tag{8}$$

$$w_{t,i}(x_t^i, y) = Q_t^{(i)}(y, x_t^i) = \frac{\dot\sigma(t)}{K}\ \big(y \neq x_t^i\big).$$

Let $p_{\theta,0}$ be the time-zero marginal obtained by initializing the reverse process from p_{ref} and replacing $s_i^\star$ with $s_{\theta,i}$ in its rates. Then

$$\mathbb{E}_{x_0 \sim p_{\mathrm{data}}}[-\log p_{\theta,0}(x_0)] \leq \mathcal{L}_{\mathrm{M2S}}(\theta) + \mathbb{E}_{x_0 \sim p_{\mathrm{data}}} D_{\mathrm{KL}}(q_T(\cdot \mid x_0) \,\|\, p_{\mathrm{ref}})\,. \tag{9}$$

The following results show that minimizing this objective recovers the true score and the clean posterior.

Proposition 4.3. *For fixed t, x, i, y with $s_i^\star(x,t;y) > 0$, the conditional risk $R_x(s) := \mathbb{E}[h(s, r_i) \mid X_t = x]$ satisfies $R_x(s) - R_x(s_i^\star) = h(s, s_i^\star) \geq 0$, with equality if and only if $s = s_i^\star$.*

Theorem 4.4. *For fixed t, x, i, let $\pi^\star$ be the clean posterior from Theorem 4.1, let $\mathcal{Y}_+ = \{y \neq x^i : w_{t,i}(x^i, y) > 0\}$, and let B_+ contain the rows of B indexed by $\mathcal{Y}_+$. Assume $s_i^\star(x,t;y) > 0$ for every $y \in \mathcal{Y}_+$. Then μ minimizes the conditional M2S risk if and only if $B_+\mu = B_+\pi^\star$. Moreover, B_+ is injective on Δ^{K-1} if and only if*

$$\ker(B_+) \cap \{v \in \mathbb{R}^K : \mathbf{1}^\top v = 0\} = \{0\}, \qquad \textit{equivalently} \qquad \operatorname{rank}\begin{bmatrix} B_+ \\ \mathbf{1}^\top \end{bmatrix} = K. \tag{10}$$

When this rank condition holds, the unique minimizer is $\mu = \pi^\star$. For the uniform kernel, $\dot{\sigma}(t) > 0$ makes every candidate positively weighted, and $\alpha_t \in (0,1)$ makes the augmented matrix in Eq. (10) full rank. Therefore, $\mu^\star = \pi^\star = \mathbb{E}[e_{X_0^i} \mid X_t = x]$.

Beyond the realizable case, this exact-recovery result suggests a canonical repair for an arbitrary predicted score: project it onto the bridge image and then invert the bridge. The uniform bridge is injective on the probability simplex, so each realizable score vector corresponds to a unique posterior. Fix t, x, i, let $k = x^i$, and let $B_{t,k} \in \mathbb{R}^{(K-1)\times K}$ map a posterior to its scores for candidates $y \neq k$. Given any score vector $s \in \mathbb{R}^{K-1}$, define the posterior whose induced score is closest to s by

$$\mu^{\mathrm{proj}}(s) := \operatorname*{arg\,min}_{\mu \in \Delta^{K-1}} \|B_{t,k}\mu - s\|_2^2. \tag{11}$$

This projection distinguishes coordinate-wise feasibility from joint Bayes realizability. Define the coordinate box and the bridge polytope by

$$\mathcal{C}_{t,k} := [\rho_t, \rho_t^{-1}]^{K-1}, \qquad \mathcal{P}_{t,k} := B_{t,k}(\Delta^{K-1}). \tag{12}$$

Equation (7) gives $\mathcal{P}_{t,k} \subseteq \mathcal{C}_{t,k}$, with strict inclusion for $K > 2$. Hence the score space has the following exact disjoint decomposition:

$$\mathbb{R}^{K-1} = \underbrace{\mathbb{R}^{K-1} \setminus \mathcal{C}_{t,k}}_{\text{coordinate-outside}} \;\dot\cup\; \underbrace{\mathcal{C}_{t,k} \setminus \mathcal{P}_{t,k}}_{\substack{\text{joint-only} \\ \text{non-realizable}}} \;\dot\cup\; \underbrace{\mathcal{P}_{t,k}}_{\text{Bayes-realizable}}. \tag{13}$$

The coordinate envelope is therefore necessary but not sufficient: a vector can satisfy every scalar bound while its unique affine inverse has a negative posterior component. Appendix D evaluates this same decomposition on raw SEDD scores. To separate material violations from sign-level floating-point effects, it divides the middle class at $\varepsilon_\mu = 10^{-6}$ into material and numerical-boundary subclasses; their union corresponds to $\mathcal{C}_{t,k} \setminus \mathcal{P}_{t,k}$ under the computed-sign convention.

Proposition 4.5. *Let $\pi^\star = \Pr(X_0^i = \cdot \mid X_t = x)$ and let $s^\star = (s_i^\star(x,t;y))_{y\neq k} = B_{t,k}\pi^\star$. For the uniform kernel with $\alpha_t \in (0,1)$, the projection in Eq. (11) uniquely recovers the clean posterior:*

$$\mu^{\mathrm{proj}}(s^\star) = \pi^\star. \tag{14}$$

Proposition 4.5 guarantees exact recovery only at the population optimum. For an arbitrary learned score s, $\mu^{\mathrm{proj}}(s)$ is a valid posterior but need not be the true one. We next show that, for the absorbing-mask process, the M2S objective in Eq. (8) reduces exactly to the MD4 loss (Shi et al., 2024).

Theorem 4.6. *Let m be an absorbing mask. For every clean token $z \neq m$, let $P_t(z \mid z) = \alpha_t$ and $P_t(m \mid z) = 1 - \alpha_t$, and let $P_t(m \mid m) = 1$, with no transitions between distinct clean tokens. Suppose that each μ_θ^i is a distribution over the clean vocabulary, so $[\mu_\theta^i]_m = 0$. Then*

$$\mathcal{L}_{\mathrm{M2S}}^{\mathrm{abs}}(\theta) = \mathcal{L}_{\mathrm{MD4}}(\theta) := \mathbb{E}_{t,x_0,x_t}\left[\sum_{i:\, x_t^i = m} \frac{-\dot\alpha_t}{1-\alpha_t}\left(-\log[\mu_\theta^i(x_t,t)]_{x_0^i}\right)\right]. \tag{15}$$

5 EXPERIMENTS

5.1 SAME-CHECKPOINT SCORE REPAIR

We first audit the raw score vectors along unmodified trajectories from the epoch-32 pure-uniform SEDD checkpoint. Across $8{,}388{,}608$ position-step vectors, 25.0177% fall outside the coordinate box $\mathcal{C}_{t,k}$, while 56.2310% satisfy every coordinate bound but materially fail membership in the bridge polytope $\mathcal{P}_{t,k}$. Only 0.0094% lie in $\mathcal{P}_{t,k}$; the remaining vectors form a narrow numerical boundary band. In a separate four-trajectory sampler diagnostic, 37.632% of positions produce at least one negative pre-normalization weight. These measurements show that coordinate-wise admissibility alone does not ensure joint compatibility with a clean-token posterior.

To isolate the effect of Bayes realizability at inference time from training differences, we keep the checkpoint, 128-step sampler, initial noise, categorical random-number stream, and final denoising update fixed. For each raw score s^{SEDD}, the repair finds the nearest simplex-constrained posterior and maps it back through the same bridge, yielding $s^{\mathrm{proj}} \in \mathcal{P}_{t,k}$. We then generate 1,024 paired length-512 sequences using either s^{SEDD} or s^{proj}. Projection removes all observed negative sampler weights and lowers external GenPPL from 203.60 to 175.07, a 14.0% reduction. The paired average-NLL difference is -0.1510 with a 95% bootstrap interval of $[-0.1774, -0.1251]$, and all generated sequences remain distinct.

Because the checkpoint and sampler are unchanged, this intervention isolates the effect of score repair for this checkpoint. It does not claim that post-hoc projection is equivalent to training M2S, nor that external GenPPL is the model likelihood. The complete protocol, score-space partition, and the distinction between direct, reconstructed, and projected scores are given in Appendix D.

5.2 IMAGE AND LANGUAGE GENERATION

We evaluate M2S on discrete image and language generation tasks. All models use the uniform forward process (Appendix B) with $\alpha_t = 1 - t$ and $\beta_t = t$. For a fair comparison, M2S adopts the same DiT (Peebles & Xie, 2023) backbone and parameter count as the mainstream baselines; full hyperparameters are given in Appendix C.

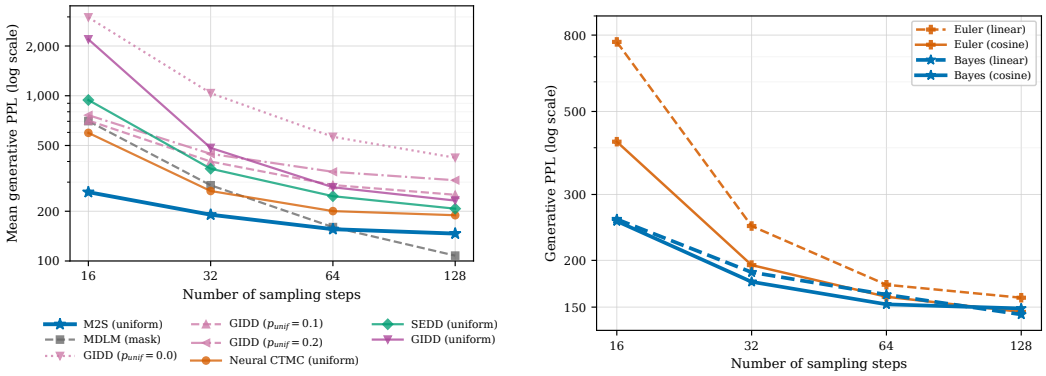
Image Generation: We train M2S on MNIST, representing each 28×28 image as a sequence over $\mathcal{S} = \{0, 1, \ldots, 255\}$. Using identical architectures, parameter counts, optimization hyperparameters, training budgets, and samplers, M2S reduces FID by more than 52 points relative to SEDD at every tested sampling budget, with full details provided in Appendix C.1.

We also evaluate M2S on CIFAR-10 using a 28.4M-parameter U-Net with self-attention. Following the MD4 setup, we match its total training budget of 512M image presentations while using pure-uniform corruption and the M2S parameterization. Without data augmentation, the model reaches a test BPD upper bound of 3.129, compared with 3.173 for SEDD under the same evaluation protocol. Thus, M2S lowers test BPD by approximately 0.045. With the same 256-step Euler sampler and coupled random-number stream, M2S also lowers FID-50k from 42.83 to 28.09 (a 34.4% reduction); qualitative samples and the complete protocol are reported in Appendix C.2. It also improves over the D3PM and Campbell et al. results in Table 1. This result provides evidence that the Bayes-realizable parameterization remains effective beyond MNIST. Full training and evaluation details are given in Appendix C.2.

Language Modeling: We compare M2S against SEDD (Lou et al., 2024), MDLM (Sahoo et al., 2024), GIDD (von Rütte et al., 2025), and Neural CTMC (Li et al., 2026) on OpenWebText (Gokaslan & Cohen, 2019). To ensure a fair comparison, M2S and all baselines (SEDD, MDLM, GIDD, Neural CTMC) use an identical backbone architecture (12-layer DiT-style Transformer, 768 hidden dim, 12 heads, $\sim$169M parameters); the only difference across methods is the parameterization of the reverse process and its associated loss. For GIDD we report results with $p_{\mathrm{unif}} \in \{0.0, 0.1, 0.2\}$, where $p_{\mathrm{unif}} = 0.0$ corresponds to a pure mask process. For evaluation, we draw 1024 unconditional samples for each method and sampling seed and score them with a pretrained **Gemma 2 9B** model to obtain *generative perplexity* (PPL). We use five sampling seeds and report their arithmetic mean; Appendix C.3 gives the corresponding standard deviations and 256-step results.

Table 1: CIFAR-10 test BPD (↓). M2S and SEDD are evaluated under the same test protocol; the remaining baseline values are published results. $\leq$ denotes a variational upper bound. The M2S and SEDD estimates are rounded to three decimal places. M2S is trained without data augmentation. A dash denotes an unreported parameter count.

Method	# Params	BPD (↓)
Autoregressive		
PixelRNN (van den Oord et al., 2016b)	–	3.00
Gated PixelCNN (van den Oord et al., 2016a)	–	3.03
PixelCNN++ (Salimans et al., 2017)	53M	2.92
PixelSNAIL (Chen et al., 2018)	46M	2.85
Image Transformer (Parmar et al., 2018)	–	2.90
Sparse Transformer (Child et al., 2019)	59M	2.80
Pure-uniform discrete diffusion		
M2S (ours)	**28.4M**	$\leq$ **3.129**
SEDD (Lou et al., 2024)	28.4M	$\leq$ 3.173
Absorbing-mask discrete diffusion		
D3PM Absorb (Austin et al., 2021)	37M	$\leq$ 4.40
Campbell et al. Absorb (Campbell et al., 2022)	28M	$\leq$ 3.52
MD4 (Shi et al., 2024)	28M	$\leq$ **2.75**
Discrete-Gaussian diffusion		
D3PM Gauss + logistic (Austin et al., 2021)	36M	$\leq$ 3.44
Campbell et al. (τLDR) (Campbell et al., 2022)	36M	$\leq$ 3.59



(a) (b)

Figure 3: (a) Five-seed mean generative-PPL comparison between M2S and the evaluated baselines. (b) A separate 1,024-sample comparison of M2S Euler and Bayes samplers under linear and cosine time grids.

Table 2 shows that M2S with Bayes sampling on the cosine grid achieves the best overall mean PPL at 16–64 steps (261.0, 190.6, and 155.7) and the best mean PPL among 131B-token checkpoints at 128 steps (146.5). Figure 3(b) separately compares M2S sampling configurations; in that fixed 1,024-sample ablation, Bayes sampling on the linear grid reaches 143.3 at 128 steps. Cross-method claims use the five-seed means in Table 2. Overall, M2S outperforms every evaluated uniform and mixture model at all four budgets. Against the strongest uniform baseline, M2S reduces mean PPL by 56% at 16 steps (596.3 to 261.0) and by 23% at 128 steps (189.2 to 146.5). At 128 steps, M2S ranks second overall only to SEDD (mask), which uses 682B training tokens compared with 131B for M2S.

Table 2: Generative perplexity (↓) on OpenWebText for varying numbers of sampling steps. All checkpoints use DiT-style backbones at a comparable parameter scale. Entries are means over five sampling seeds. **Bold**: best overall per column; underline: best among rows with a reported 131B-token training budget.

Type	Method	Train	Len.	16	32	64	128
mask	SEDD (Lou et al., 2024)	682B	1024	630.8	275.6	158.2	**107.7**
	MDLM (Sahoo et al., 2024)	262B	1024	702.6	287.3	160.4	107.8
	GIDD ($p_{\text{unif}} = 0.0$) (von Rütte et al., 2025)	131B	512	2980.2	1036.9	564.2	421.7
mixture	GIDD ($p_{\text{unif}} = 0.1$) (von Rütte et al., 2025)	131B	512	708.3	399.8	288.1	251.6
	GIDD ($p_{\text{unif}} = 0.2$) (von Rütte et al., 2025)	131B	512	763.0	445.0	346.5	308.1
uniform	Neural CTMC (Li et al., 2026)	131B	512	596.3	265.2	200.4	189.2
	SEDD (Lou et al., 2024)	131B	512	942.3	361.8	246.7	207.3
	GIDD (von Rütte et al., 2025)	131B	512	2194.9	483.5	279.4	232.1
	M2S (Bayes, cosine)	131B	512	**261.0**	**190.6**	**155.7**	146.5

Note: (1) The losses of SEDD (mask) and GIDD ($p_{\text{unif}} = 0.0$) are equivalent to the MDLM loss. (2) Checkpoint and sampler details are given in Appendix C.3.

6 Conclusion

This work identifies Bayes realizability as a structural requirement for discrete score vectors. Positive SEDD scores define valid reverse CTMC rates, but the complete vector need not be induced by any clean-token posterior. We characterize this gap and introduce M2S, which predicts the one-site clean-token posterior and maps it to all candidate scores through the known forward kernel. The construction applies to known coordinate-wise kernels under a mild support condition, restricts uniform-corruption scores to the bridge polytope, and recovers MD4 under absorbing-mask corruption. Empirically, M2S improves FID over SEDD under identical MNIST and CIFAR-10 settings and outperforms the evaluated pure-uniform baselines on OpenWebText at every tested sampling budget. A fixed-checkpoint intervention further shows that projecting SEDD scores onto the bridge polytope removes the observed negative sampler weights and improves generative PPL. These results establish Bayes realizability as a practical design principle for score-based discrete diffusion.

AI use statement

Generative AI tools were used to assist with drafting and language polishing, organizing the related literature, and improving the presentation and interpretation of experimental results. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

References

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*, volume 34, pp. 17981–17993. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.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.

Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *Computer Vision – ECCV 2022*, volume 13683 of *Lecture Notes in Computer Science*, pp. 170–188. Springer, 2022. doi: 10.1007/978-3-031-20050-2_11. URL https://doi.org/10.1007/978-3-031-20050-2_11.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28266–28279. Curran Associates, Inc., 2022. doi: 10.

52202/068431-2049. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b5b528767aa35f5b1a60fe0aaeca0563-Abstract-Conference.html.

Rodrigo Casado Noguerales, Bernhard Schölkopf, Thomas Hofmann, and Aran Raoufi. What does a discrete diffusion model learn? *arXiv preprint arXiv:2607.05381*, 2026. URL https://arxiv.org/abs/2607.05381.

Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. PixelSNAIL: An improved autoregressive generative model. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 864–872. PMLR, 2018. URL https://proceedings.mlr.press/v80/chen18h.html.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. URL https://arxiv.org/abs/1904.10509.

Laurent Condat. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158 (1–2):575–585, 2016. doi: 10.1007/s10107-015-0946-6. URL https://doi.org/10.1007/s10107-015-0946-6.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279, New York, NY, USA, 2008. Association for Computing Machinery. doi: 10.1145/1390156.1390191. URL https://doi.org/10.1145/1390156.1390191.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. https://skylion007.github.io/OpenWebTextCorpus, 2019.

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. *arXiv preprint arXiv:2605.22765*, 2026. URL https://arxiv.org/abs/2605.22765.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. DiffusionBERT: Improving generative masked language models with diffusion models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4521–4534, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.248. URL https://aclanthology.org/2023.acl-long.248/.

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*, volume 34, pp. 12454–12465. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/67d96d458abdef21792e6d8e590244e7-Abstract.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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. *arXiv preprint arXiv:2604.15694*, 2026. URL https://arxiv.org/abs/2604.15694.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

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*, volume 35, pp. 34532–34545. Curran Associates, Inc., 2022. doi: 10.52202/068431-2502. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/df04a35d907e894d59d4eab1f92bc87b-Abstract-Conference.html.

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

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4055–4064. PMLR, 2018. URL https://proceedings.mlr.press/v80/parmar18a.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

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.pdf.

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*, volume 37, pp. 130136–130184. Curran Associates, Inc., 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJrFC6ceg.

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*, volume 37, pp. 103131–103167. Curran Associates, Inc., 2024. doi: 10.52202/079017-3277. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html.

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

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

Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with PixelCNN decoders. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4790–4798. Curran Associates, Inc., 2016a. URL https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1747–1756. PMLR, 2016b. URL https://proceedings.mlr.press/v48/oord16.html.

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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61810–61843. PMLR, 2025. URL https://proceedings.mlr.press/v267/von-rutte25a.html.

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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76716–76753. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25cw.html.

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

Appendix

CONTENTS

A Proofs **12**

B Algorithms **16**

B.1 Training Algorithm 16

B.2 Sampling Algorithms 17

C Experimental Details **18**

C.1 MNIST Experiments 18

C.2 CIFAR-10 Experiments 18

C.3 OpenWebText Experiments 20

D SEDD Bayes Realizability: Projection Repair, Violation Rates, and Sample Quality **22**

D.1 Projection Repair 22

D.2 Bayes-Realizability Violation Rates 23

D.3 Sample Quality after Projection Repair 24

D.4 Projected SEDD Sampler 24

E Sample Generation **25**

A PROOFS

Assumption A.1 (Support condition)**.** Scores and posteriors are evaluated only at states x with $p_t(x) > 0$. For every site i, candidate y, and clean token z in the site-i data support,

$$P_t^{(i)}(x^i \mid z) = 0 \implies P_t^{(i)}(y \mid z) = 0.$$

Theorem 4.1. *Under Assumption A.1, for any site i, candidate token y, and noisy state x with $p_t(x) > 0$, let $\pi^\star(z) = \Pr(X_0^i = z \mid X_t = x)$. For every $z \in \mathrm{supp}(\pi^\star)$, we have $P_t^{(i)}(x^i \mid z) > 0$. Then*

$$s_i^\star(x, t; y) = \mathbb{E}\left[\frac{P_t^{(i)}(y \mid X_0^i)}{P_t^{(i)}(x^i \mid X_0^i)} \,\middle|\, X_t = x\right] = \sum_{z \in \mathrm{supp}(\pi^\star)} \pi^\star(z)\, \frac{P_t^{(i)}(y \mid z)}{P_t^{(i)}(x^i \mid z)}.$$

Proof of Theorem 4.1. For each clean token z, define

$$A_z := \sum_{x_0 :\, x_0^i = z} p_{\mathrm{data}}(x_0) \prod_{j \neq i} P_t^{(j)}(x^j \mid x_0^j), \qquad P_z := P_t^{(i)}(x^i \mid z).$$

These quantities collect all contributions outside site i. They give

$$p_t(x) = \sum_z A_z P_z, \qquad p_t(x^{i \to y}) = \sum_z A_z P_t^{(i)}(y \mid z).$$

The one-site posterior satisfies

$$\Pr(X_0^i = z \mid X_t = x) = \frac{A_z P_z}{p_t(x)}.$$

Let $\mathcal{Z}_\star = \mathrm{supp}(\pi^\star) = \{z : A_z P_z > 0\}$. For every $z \in \mathcal{Z}_\star$, we have $P_z > 0$, so define

$$g(z) := \frac{P_t^{(i)}(y \mid z)}{P_t^{(i)}(x^i \mid z)}$$

on $\mathcal{Z}_\star$. Expanding the conditional expectation over its support gives

$$\begin{aligned}\mathbb{E}[g(X_0^i) \mid X_t = x] &= \sum_{z \in \mathcal{Z}_\star} \pi^\star(z) g(z) \\ &= \sum_{z \in \mathcal{Z}_\star} \frac{A_z P_z}{p_t(x)} \frac{P_t^{(i)}(y \mid z)}{P_z} \\ &= \frac{1}{p_t(x)} \sum_{z \in \mathcal{Z}_\star} A_z P_t^{(i)}(y \mid z).\end{aligned}$$

If $z \notin \mathcal{Z}_\star$, then either $A_z = 0$ or $P_z = 0$. In the latter case, whenever $A_z > 0$, the token z occurs in the clean support at site i, so Assumption A.1 implies $P_t^{(i)}(y \mid z) = 0$. Thus $A_z P_t^{(i)}(y \mid z) = 0$ outside $\mathcal{Z}_\star$, and the restricted sum can be extended to all z. Hence

$$\mathbb{E}[g(X_0^i) \mid X_t = x] = \frac{\sum_z A_z P_t^{(i)}(y \mid z)}{p_t(x)} = \frac{p_t(x^{i \to y})}{p_t(x)}.$$

The first equality above is the posterior expansion in the theorem, and the last equality is the concrete score $s_i^\star(x, t; y)$. □

Theorem 4.2. *For the uniform CTMC with $\alpha_t = e^{-\sigma(t)}$ and $\dot\sigma(t) \geq 0$, let $x_0 \sim p_{\mathrm{data}}$, $t \sim \mathrm{Unif}[0,1]$, and $x_t \sim q_t(\cdot \mid x_0)$. For $y \neq x_t^i$, define $r_i(x_0, x_t, t; y) = P_t^{(i)}(y \mid x_0^i)/P_t^{(i)}(x_t^i \mid x_0^i)$. For $s, r > 0$, define $h(s,r) = s - r\log s + r\log r - r = r\big(s/r - 1 - \log(s/r)\big) \geq 0$, with equality if and only if $s = r$, and extend this definition continuously to $r = 0$ by $h(s,0) = s$. Define the M2S objective*

$$\mathcal{L}_{\mathrm{M2S}}(\theta) = \mathbb{E}_{t, x_0, x_t}\left[\sum_{i=1}^{L} \sum_{y \neq x_t^i} w_{t,i}(x_t^i, y)\, h\big((B\mu_\theta^i)_y, r_i(x_0, x_t, t; y)\big)\right],$$

$$w_{t,i}(x_t^i, y) = Q_t^{(i)}(y, x_t^i) = \frac{\dot\sigma(t)}{K}\ (y \neq x_t^i).$$

Let $p_{\theta,0}$ be the time-zero marginal obtained by initializing the reverse process from p_{ref} and replacing $s_i^\star$ with $s_{\theta,i}$ in its rates. Then

$$\mathbb{E}_{x_0 \sim p_{\mathrm{data}}}[-\log p_{\theta,0}(x_0)] \leq \mathcal{L}_{\mathrm{M2S}}(\theta) + \mathbb{E}_{x_0 \sim p_{\mathrm{data}}} D_{\mathrm{KL}}(q_T(\cdot \mid x_0) \,\|\, p_{\mathrm{ref}})\,.$$

Proof of Theorem 4.2. For $u = s/r > 0$, the inequality $u - 1 - \log u \geq 0$ gives $h(s,r) = r(u - 1 - \log u) \geq 0$, with equality only at $u = 1$.

Theorem 3.6 of Lou et al. (2024), based on the continuous-time ELBO of Campbell et al. (2022), includes the target-only normalization $r(\log r - 1) = r\log r - r$. Its integrand is therefore exactly $h(s,r)$. Since $T = 1$, the theorem gives, for every x_0,

$$\begin{aligned}-\log p_{\theta,0}(x_0) \leq \int_0^1 \mathbb{E}_{x_t \sim q_t(\cdot \mid x_0)}\left[\sum_{\widetilde{x} \neq x_t} Q_t(\widetilde{x}, x_t)\, h\left(s_\theta(x_t, t; \widetilde{x}), \frac{q_t(\widetilde{x} \mid x_0)}{q_t(x_t \mid x_0)}\right)\right] \mathrm{d}t \\ + D_{\mathrm{KL}}(q_T(\cdot \mid x_0) \,\|\, p_{\mathrm{ref}})\,,\end{aligned}$$

written here in our row-generator convention.

The coordinate-wise generator has a nonzero off-diagonal entry only when $\widetilde{x} = x_t^{i \to y}$ for some i and $y \neq x_t^i$. For this pair,

$$Q_t(\widetilde{x}, x_t) = Q_t^{(i)}(y, x_t^i), \qquad \frac{q_t(\widetilde{x} \mid x_0)}{q_t(x_t \mid x_0)} = \frac{P_t^{(i)}(y \mid x_0^i)}{P_t^{(i)}(x_t^i \mid x_0^i)} = r_i(x_0, x_t, t; y).$$

For the uniform process, $Q_t^{(i)}(y, x_t^i) = \dot{\sigma}(t)/K$ for $y \neq x_t^i$. Hence the integral, after averaging over $x_0 \sim p_{\text{data}}$, is exactly $\mathcal{L}_{\text{M2S}}(\theta)$ because $t \sim \text{Unif}[0, 1]$. Averaging the pointwise bound proves Eq. (9). □

Proposition 4.3. *For fixed t, x, i, y with $s_i^\star(x, t; y) > 0$, the conditional risk $R_x(s) := \mathbb{E}[h(s, r_i) \mid X_t = x]$ satisfies $R_x(s) - R_x(s_i^\star) = h(s, s_i^\star) \geq 0$, with equality if and only if $s = s_i^\star$.*

Proof of Proposition 4.3. Condition on $X_t = x$ and write $s^\star = s_i^\star(x, t; y)$. Since $\mathbb{E}[r_i \mid X_t = x] = s^\star$ by Theorem 4.1,

$$\begin{aligned} R_x(s) &:= \mathbb{E}[h(s, r_i) \mid X_t = x] \\ &= s - s^\star \log s + C_x, \end{aligned}$$

where C_x is independent of s. Moreover,

$$R_x'(s) = 1 - \frac{s^\star}{s}, \qquad R_x''(s) = \frac{s^\star}{s^2} > 0.$$

Thus R_x is strictly convex and uniquely minimized at $s^\star$, with

$$R_x(s) - R_x(s^\star) = s - s^\star - s^\star \log \frac{s}{s^\star} = h(s, s^\star).$$

□

Theorem 4.4. *For fixed t, x, i, let $\pi^\star$ be the clean posterior from Theorem 4.1, let $\mathcal{Y}_+ = \{y \neq x^i : w_{t,i}(x^i, y) > 0\}$, and let B_+ contain the rows of B indexed by $\mathcal{Y}_+$. Assume $s_i^\star(x, t; y) > 0$ for every $y \in \mathcal{Y}_+$. Then μ minimizes the conditional M2S risk if and only if $B_+\mu = B_+\pi^\star$. Moreover, B_+ is injective on Δ^{K-1} if and only if*

$$\ker(B_+) \cap \{v \in \mathbb{R}^K : \mathbf{1}^\top v = 0\} = \{0\}, \qquad \textit{equivalently} \qquad \operatorname{rank}\begin{bmatrix} B_+ \\ \mathbf{1}^\top \end{bmatrix} = K.$$

When this rank condition holds, the unique minimizer is $\mu = \pi^\star$. For the uniform kernel, $\dot{\sigma}(t) > 0$ makes every candidate positively weighted, and $\alpha_t \in (0, 1)$ makes the augmented matrix in Eq. (10) full rank, and hence $\mu^\star = \pi^\star = \mathbb{E}[e_{X_0^i} \mid X_t = x]$.

Proof of Theorem 4.4. By Theorem 4.1,

$$B\pi^\star = s_i^\star(x, t; \cdot).$$

Write $w_y = w_{t,i}(x^i, y)$ and let $\mathcal{R}_x(\mu)$ denote the conditional M2S risk summed over $y \in \mathcal{Y}_+$. Applying Proposition 4.3 to each candidate gives, for every μ with finite conditional risk,

$$\begin{aligned} \mathcal{R}_x(\mu) - \mathcal{R}_x(\pi^\star) &= \sum_{y \in \mathcal{Y}_+} w_y \, h\big((B\mu)_y, (B\pi^\star)_y\big) \\ &\geq 0. \end{aligned}$$

Here every w_y and $(B\pi^\star)_y = s_i^\star(x, t; y)$ is positive. Since $h(s, r) = 0$ if and only if $s = r$, equality holds exactly when

$$B_+\mu = B_+\pi^\star.$$

This proves the first claim.

We next prove the rank characterization. If $\mu, \nu \in \Delta^{K-1}$ satisfy $B_+\mu = B_+\nu$, then $v = \mu - \nu$ satisfies

$$B_+ v = 0, \qquad \mathbf{1}^\top v = \mathbf{1}^\top \mu - \mathbf{1}^\top \nu = 0.$$

Thus the null-space condition in Eq. (10) implies $v = 0$ and hence $\mu = \nu$, so B_+ is injective on Δ^{K-1}.

For the converse, suppose there exists a nonzero $v \in \ker(B_+)$ with $\mathbf{1}^\top v = 0$. Let

$$\bar{\mu} = \frac{1}{K}\mathbf{1}, \qquad 0 < \varepsilon < \frac{1}{K\|v\|_\infty}, \qquad \mu^\pm = \bar{\mu} \pm \varepsilon v.$$

The choice of ε ensures that every coordinate of $\mu^\pm$ is positive, while $\mathbf{1}^\top v = 0$ gives $\mathbf{1}^\top \mu^\pm = 1$. Hence $\mu^+, \mu^- \in \Delta^{K-1}$ and $\mu^+ \neq \mu^-$. However,

$$B_+\mu^+ - B_+\mu^- = 2\varepsilon B_+ v = 0,$$

so B_+ is not injective on Δ^{K-1}. This proves

$$B_+ \text{ is injective on } \Delta^{K-1} \quad \Longleftrightarrow \quad \ker(B_+) \cap \ker(\mathbf{1}^\top) = \{0\}.$$

Finally,

$$\ker \begin{bmatrix} B_+ \\ \mathbf{1}^\top \end{bmatrix} = \ker(B_+) \cap \ker(\mathbf{1}^\top).$$

Because the augmented matrix has K columns, its kernel is trivial if and only if its column rank is K. This establishes both equivalences in Eq. (10). When they hold, $B_+\mu = B_+\pi^\star$ forces $\mu = \pi^\star$.

For the uniform kernel, $\dot{\sigma}(t) > 0$ makes every candidate $y \neq x^i$ positively weighted, so $B_+ = B$. Moreover, $\alpha_t \in (0,1)$ implies $\rho_t \in (0,1)$. Let $k = x^i$ and suppose $\mu, \nu \in \Delta^{K-1}$ satisfy $B\mu = B\nu$. For any $\eta \in \Delta^{K-1}$, summing Eq. (7) over $y \neq k$ gives

$$\begin{aligned} F(\eta) &:= \sum_{y \neq k} \big((B\eta)_y - 1\big) \\ &= (K-1)(\rho_t - 1)\eta_k + (\rho_t^{-1} - 1)(1 - \eta_k). \end{aligned}$$

The coefficient of η_k is

$$(K-1)(\rho_t - 1) - (\rho_t^{-1} - 1) = (\rho_t - 1)\big((K-1) + \rho_t^{-1}\big),$$

which is nonzero for $\rho_t \in (0,1)$. Since $B\mu = B\nu$ implies $F(\mu) = F(\nu)$, it follows that $\mu_k = \nu_k$. For each $y \neq k$, Eq. (7) then gives

$$\begin{aligned} 0 &= (B\mu)_y - (B\nu)_y \\ &= (\rho_t - 1)(\mu_k - \nu_k) + (\rho_t^{-1} - 1)(\mu_y - \nu_y) \\ &= (\rho_t^{-1} - 1)(\mu_y - \nu_y). \end{aligned}$$

Because $\rho_t^{-1} - 1 \neq 0$, we obtain $\mu_y = \nu_y$ for every $y \neq k$. Thus $\mu = \nu$, proving that the uniform bridge is injective on Δ^{K-1}; by the equivalence above, its augmented matrix has rank K. The unique minimizer is therefore $\mu = \pi^\star = \mathbb{E}[e_{X_0^i} \mid X_t = x]$. □

Proposition 4.5. *Let $\pi^\star = \Pr(X_0^i = \cdot \mid X_t = x)$ and let $s^\star = (s_i^\star(x,t;y))_{y \neq k} = B_{t,k}\pi^\star$. For the uniform kernel with $\alpha_t \in (0,1)$, the projection in Eq.* (11) *uniquely recovers the clean posterior:*

$$\mu^{\mathrm{proj}}(s^\star) = \pi^\star.$$

Proof of Proposition 4.5. Because $\pi^\star \in \Delta^{K-1}$ and $s^\star = B_{t,k}\pi^\star$, choosing $\mu = \pi^\star$ in Eq. (11) gives objective value zero. Hence every minimizer $\widehat{\mu}$ satisfies $B_{t,k}\widehat{\mu} = s^\star = B_{t,k}\pi^\star$. The uniform bridge is injective on Δ^{K-1} when $\alpha_t \in (0,1)$ by Theorem 4.4, so $\widehat{\mu} = \pi^\star$. Thus the minimizer is unique and equals the clean posterior. □

Theorem 4.6. *Let m be an absorbing mask. For every clean token $z \neq m$, let $P_t(z \mid z) = \alpha_t$ and $P_t(m \mid z) = 1 - \alpha_t$, and let $P_t(m \mid m) = 1$, with no transitions between distinct clean tokens. Suppose that each μ_θ^i is a distribution over the clean vocabulary, so $[\mu_\theta^i]_m = 0$. Then*

$$\mathcal{L}_{\mathrm{M2S}}^{\mathrm{abs}}(\theta) = \mathcal{L}_{\mathrm{MD4}}(\theta) := \mathbb{E}_{t,x_0,x_t}\left[\sum_{i:\, x_t^i = m} \frac{-\dot{\alpha}_t}{1-\alpha_t}\left(-\log[\mu_\theta^i(x_t,t)]_{x_0^i}\right)\right].$$

Proof of Theorem 4.6. **Masked sites.** Fix a site with $x_t^i = m$ and let $c_t = \alpha_t/(1-\alpha_t)$. The bridge and conditional target reduce to

$$s_{\theta,i}(x_t,t;y) = c_t[\mu_\theta^i(x_t,t)]_y, \qquad r_i(x_0,x_t,t;y) = c_t \mathbf{1}\{y = x_0^i\}.$$

Using $h(s,0) = s$, substitution into the inner score-entropy sum in Eq. (8) gives

$$\sum_{y \neq m}\Big[c_t[\mu_\theta^i]_y - c_t\mathbf{1}\{y = x_0^i\}\log(c_t[\mu_\theta^i]_y) + c_t\mathbf{1}\{y = x_0^i\}\log c_t - c_t\mathbf{1}\{y = x_0^i\}\Big].$$

Because μ_θ^i is supported on the clean vocabulary, $\sum_{y\neq m}[\mu_\theta^i]_y = 1$. The first and last terms therefore cancel after summation, and the two occurrences of $\log c_t$ also cancel. The remaining loss is

$$-c_t \log[\mu_\theta^i]_{x_0^i}.$$

The Kolmogorov equation gives the clean-to-mask rate $Q_t^{(i)}(y,m) = -\dot{\alpha}_t/\alpha_t$, independent of the clean candidate y. Therefore,

$$Q_t^{(i)}(y,m)\, c_t = \frac{-\dot{\alpha}_t}{\alpha_t}\frac{\alpha_t}{1-\alpha_t} = \frac{-\dot{\alpha}_t}{1-\alpha_t}.$$

Thus, summing over masked sites and taking the expectation gives exactly Eq. (15).

Unmasked sites. Now let $x_t^i = x_0^i = k \neq m$. For every candidate $y \neq k$, $Q_t^{(i)}(y,k) = 0$: clean-to-clean rates vanish, and the absorbing state has no outgoing rate. Thus every candidate weight is zero, so unmasked sites make no contribution. This completes the reduction. □

B Algorithms

B.1 Training Algorithm

During training, a clean token z is corrupted according to the uniform kernel

$$q_t(y \mid z) = \alpha_t \mathbf{1}\{y = z\} + \frac{1-\alpha_t}{K}, \qquad \alpha_t = 1 - t, \tag{16}$$

where y is the corrupted token, K is the vocabulary size, and $t \in [0,1]$.

Algorithm 1 summarizes the training procedure. Given a clean sequence x_0 and a diffusion time t, we sample x_t from the forward kernel in Eq. (16). The model predicts the clean-token posterior, converts it to concrete scores through the M2S bridge, and is optimized with the loss in Eq. (8).

Algorithm 1 M2S Training

Require: model f_θ, data distribution p_{data}, vocabulary size K, learning rate η

while not converged **do**
  Sample $x_0 \sim p_{\mathrm{data}}$ and $t \sim \mathrm{Unif}[0,1]$
  Corrupt x_0 with Eq. (16) to obtain x_t
  Predict $\mu_\theta^i = \mathrm{softmax}(f_\theta^i(x_t,t))$ for all sites i
  Compute $s_{\theta,i}(x_t,t;y)$ for all i and $y \neq x_t^i$ using Eq. (6)
  Compute r_i and $w_{t,i}$ from the known forward process as in Eq. (8)
  $\theta \leftarrow \theta - \eta\nabla_\theta \mathcal{L}_{\mathrm{M2S}}(\theta)$
end while
return θ

B.2 Sampling Algorithms

M2S pairs Euler and Bayes updates with either a linear or cosine time grid. This gives four samplers: Euler-linear, Euler-cosine, Bayes-linear, and Bayes-cosine.

For M sampling steps, let $t_0 > t_1 > \cdots > t_M$ denote the reverse-time grid:

$$t_m^{\text{lin}} = 1 - \frac{m}{M}, \qquad t_m^{\cos} = \cos\Big(\frac{\pi m}{2M}\Big), \qquad m = 0, \dots, M, \tag{17}$$

Let $\Delta_m = t_m - t_{m+1}$. Both grids start at $t_0 = 1$ and end at $t_M = 0$. All samplers initialize X_{t_0} from the uniform distribution and update all sites in parallel. At step m, the model predicts the clean-token posterior $\mu_{\theta,i}$ from the current sequence $x = X_{t_m}$ and maps it to scores using Eq. (7). Euler sampling uses the reverse rates

$$R_{m,i}(y) = Q^{(i)}_{t_m}(y, x^i)\, s_{\theta,i}(x, t_m; y), \qquad y \neq x^i. \tag{18}$$

Algorithm 2 M2S Euler Sampling (Linear or Cosine Grid)

Require: model f_θ, steps M, grid type $g \in \{\text{lin}, \cos\}$, vocabulary size K, sequence length L
Construct $\{t_m\}_{m=0}^M$ from Eq. (17)
Sample $X^i_{t_0} \sim \text{Unif}(\{1, \dots, K\})$ for all sites i
for $m = 0, \dots, M-1$ **do**
  $x \leftarrow X_{t_m}$ and $\Delta_m \leftarrow t_m - t_{m+1}$
  Predict $\mu_{\theta,i}$ and compute scores using Eq. (7)
  Compute $R_{m,i}$ using Eq. (18)
  for all sites i **in parallel do**
    $p_i(y) \leftarrow \Delta_m R_{m,i}(y)$ for $y \neq x^i$
    $p_i(x^i) \leftarrow 1 - \sum_{y \neq x^i} p_i(y)$
    Sample $X^i_{t_{m+1}} \sim \text{Categorical}(p_i)$
  end for
end for
return X_{t_M}

Bayes sampling draws directly from the finite-time posterior instead of discretizing the reverse rates. For the current token $k = X^i_{t_m}$, marginalizing over $\mu_{\theta,i}$ gives

$$\widetilde{p}_{m,i}(y) = \sum_{z=1}^{K} [\mu_{\theta,i}]_z \, \frac{q_{t_{m+1}}(y \mid z)\, q_{t_m|t_{m+1}}(k \mid y)}{q_{t_m}(k \mid z)}. \tag{19}$$

Algorithm 3 M2S Bayes Sampling (Linear or Cosine Grid)

Require: model f_θ, steps M, grid type $g \in \{\text{lin}, \cos\}$, vocabulary size K, sequence length L
Construct $\{t_m\}_{m=0}^M$ from Eq. (17)
Sample $X^i_{t_0} \sim \text{Unif}(\{1, \dots, K\})$ for all sites i
for $m = 0, \dots, M-1$ **do**
  $x \leftarrow X_{t_m}$
  Predict $\mu_{\theta,i}$ from (x, t_m)
  for all sites i **in parallel do**
    Compute $\widetilde{p}_{m,i}(y)$ for all y using Eq. (19)
    Normalize $p_{m,i} \leftarrow \widetilde{p}_{m,i} / \sum_y \widetilde{p}_{m,i}(y)$
    Sample $X^i_{t_{m+1}} \sim \text{Categorical}(p_{m,i})$
  end for
end for
return X_0

C Experimental Details

C.1 MNIST Experiments

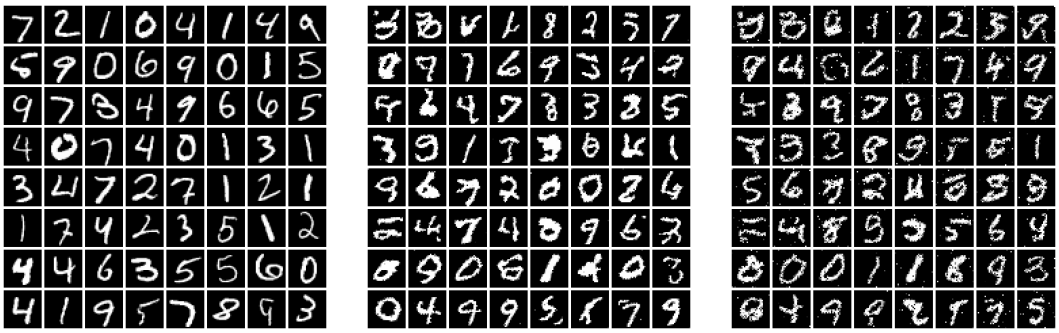
We compare M2S and SEDD on 28×28 MNIST images quantized to $K = 256$ gray levels. For a fair comparison, both methods use the same width-64 U-Net architecture and parameter count, optimization hyperparameters (AdamW with a learning rate of 2×10^{-4}), uniform transition kernel and forward process $\alpha_t = e^{-t}$, 50-epoch training budget and generate samples with the same sampler, while differing only in the training parameterization: M2S predicts the posterior mean, whereas SEDD predicts the log-score directly. We generate 5000 samples for each model and compute FID against the first 5000 MNIST test images.

Table 3 compares the two methods with 50, 100, and 200 sampling steps while holding all other settings fixed. M2S lowers FID by more than 52 points at every budget.

Table 3: MNIST FID (↓) across sampling budgets.

Objective	50 steps	100 steps	200 steps
M2S (ours)	**72.80**	**72.87**	**73.44**
SEDD	127.18	126.81	126.19

Figure 4 shows the first 64 uncurated samples from the 200-step runs. M2S produces clearer strokes and fewer isolated bright pixels than SEDD, consistent with the FID comparison.



(a) MNIST test data (b) M2S, FID 73.4 (c) SEDD, FID 126.2

Figure 4: Qualitative comparison on MNIST.

C.2 CIFAR-10 Experiments

Training setup. We follow the CIFAR-10 (Krizhevsky, 2009) setup of MD4 (Shi et al., 2024). Each 32×32 RGB image is represented as a length-3072 sequence with 256 states per color channel. We use the same approximately 28M-parameter U-Net with self-attention, AdamW with learning rate 4×10^{-4} and weight decay 0.01. The warmup covers 25,600 image presentations, equal to 100 MD4 batch-256 updates and 6.25 of our large-batch updates, and is followed by cosine learning-rate decay. The model is trained without data augmentation. M2S retains the clean-token posterior head but uses pure-uniform corruption and maps its output to concrete scores as described in Section 4.

MD4 trains for 2M updates with batch size 256. We use a global batch size of 4096 for 125,000 updates, matching its total budget of 512M image presentations (10,240 CIFAR-10-equivalent epochs). The resulting 28,427,520-parameter checkpoint achieves a CIFAR-10 test BPD upper bound of 3.129. Re-evaluating SEDD under exactly the same test protocol gives 3.173, so M2S lowers test BPD by approximately 0.045.

Optimization trace. The two parameterizations are trained with the same complete score-entropy objective: M2S maps posterior logits to a log-score through the uniform bridge, whereas SEDD

predicts the log-score directly, after which both use the same time distribution, corruption kernel, token sum, and batch mean. We log the all-reduced raw objective over all 16 ranks at every optimizer step and convert step u to $4096u/50{,}000$ CIFAR-10-equivalent epochs. M2S starts higher and optimizes more slowly early in training, but its trailing 625-step mean makes its final crossover below SEDD at equivalent epoch 3963.6 and remains lower thereafter. Over the final 5,000 updates, the mean raw loss is 6770.1 for M2S and 6901.9 for SEDD. This curve is an optimization diagnostic, distinct from the held-out BPD estimates above.

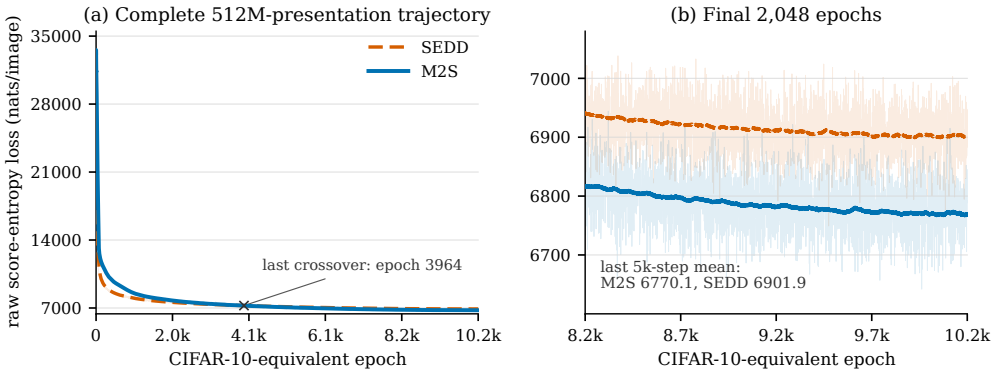


Figure 5: **M2S converges to a lower CIFAR-10 training objective.** Panel (a) includes every optimizer step in the 512M-image-presentation run; panel (b) enlarges the final 2,048 equivalent epochs. Faint curves show raw per-step losses (subsampled only for rendering), and bold curves show trailing 625-step means, a 51.2-equivalent-epoch window. Both axes use equivalent epoch rather than wall-clock time or dataloader passes.

Generation protocol. We compare the EMA weights at update 125,000 for M2S and SEDD. Both checkpoints have 28,427,520 parameters and were trained without augmentation using the same optimizer, schedule, global batch size, and 512M image-presentation budget. For each method, we generate 50,000 images with the same 256-step categorical Euler discretization, cosine time grid, eight-GPU rank assignment, per-GPU batch size 256, and seed 20260723. Every rank-local batch is independently seeded from its first global sample index, so corresponding M2S and SEDD images use the same initial noise and categorical random-number stream even after a restart. We do not apply a final clean-token projection, rejection, or sample filtering. The generation paths differ only in the learned model output and the corresponding reverse-rate construction: M2S maps a posterior through the bridge, whereas SEDD exponentiates a directly predicted log-score.

Table 4: **M2S improves both likelihood and sample quality in the controlled CIFAR-10 comparison.** BPD is the four-pass variational upper bound on all 10,000 test images. FID-50k uses `pytorch-fid` v0.3.0 with 2,048-dimensional Inception pool-3 features against all 50,000 CIFAR-10 training images. Lower is better.

Method	Admissible score set	Test BPD (↓)	FID-50k (↓)
M2S (ours)	$\mathcal{P}_{t,k} = B_{t,k}(\Delta^{K-1})$	$\leq \mathbf{3.129}$	**28.09**
SEDD	$\mathbb{R}^{K-1}_{>0}$	≤ 3.173	42.83



(a) M2S, FID-50k 28.09 (b) SEDD, FID-50k 42.83

Figure 6: **M2S suppresses the isolated chromatic artifacts visible in SEDD under a fully paired sampler.** Each panel contains generated indices 0–63 from its corresponding 50,000-image FID run, in index order, without ranking, filtering, or manual selection. Corresponding positions use the same seeded random-number stream, 256-step Euler discretization, and cosine grid.

Connection to Bayes realizability. Theorem 4.1 shows that the true score vector is induced by one clean-token posterior. Because M2S predicts that posterior in the simplex, its entire off-diagonal score vector lies in the bridge polytope $\mathcal{P}_{t,k}$ from Eq. (12) at every network iterate. SEDD enforces positivity coordinate by coordinate but does not enforce this joint constraint. Under finite data, capacity, and optimization, the M2S parameterization therefore removes non-Bayesian score degrees of freedom and promotes a coherent reverse-rate field. In the controlled experiment, this structural restriction coincides with a 14.74-point FID reduction (34.4%), the lower test BPD in Table 4, and fewer isolated high-saturation pixels and local texture breaks in Figure 6. Both parameterizations contain the population-optimal score, so the theorem alone does not imply a universal FID ordering; the paired result is empirical evidence consistent with the proposed finite-training mechanism.

C.3 OpenWebText Experiments

M2S training. We train a 169.7M-parameter diffusion Transformer on OpenWebText with GPT-2 (Radford et al., 2019) tokenization and a maximum sequence length of 512. We use the uniform forward process and the M2S loss in Eq. (8), setting the time-sampling cutoff to 10^{-3}. The global batch size is 1584 across 24 H100 GPUs. We use AdamW with a peak learning rate of 5×10^{-4}, 3200 linear warmup steps, cosine decay, gradient clipping at 1.0, and bf16 precision. The model is trained on 131B tokens.

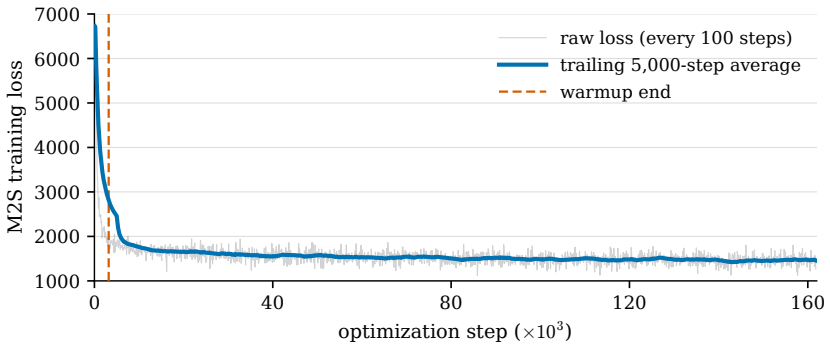


Figure 7: **M2S training loss over 161.8k optimization steps.** The gray curve shows raw loss recorded every 100 steps, the blue curve shows its trailing 5,000-step average, and the dashed line marks the end of warmup.

GenPPL evaluation. For each method, sampling-step budget, and seed, we draw 1024 unconditional samples and score 512 tokens per sample with the same Gemma 2 9B evaluator. For checkpoints with `max_len`=1024, we evaluate only the first 512 generated tokens; all other checkpoints generate 512-token sequences. We report GenPPL as the exponentiated average next-token NLL assigned by the evaluator. To evaluate the baselines with their intended sampling procedures, MDLM, GIDD, SEDD, and Neural CTMC each use the sampler specified in the corresponding paper (Sahoo et al., 2024; von Rütte et al., 2025; Lou et al., 2024; Li et al., 2026). The sampling algorithms used for M2S are described in Appendix B. Generated samples are shown in Appendix E.

Extended evaluator results. Tables 5 and 6 report the arithmetic mean and sample standard deviation over five sampling seeds at each budget. The two evaluator tables use the same generated samples.

Table 5: Gemma 2 9B generative PPL on OpenWebText (lower is better). Standard deviations are shown in parentheses. Bold marks the best result among checkpoints with maximum length 512; † marks the overall best. M2S uses Bayes sampling on the cosine grid; the pure-uniform SEDD row uses Euler sampling.

Type	Method	Train	Len.	Sampling steps				
				16	32	64	128	256
mask	SEDD	682B	1024	630.793 (7.926)	275.590 (4.928)	158.236 (3.333)	107.719† (1.001)	79.246 (0.966)
	MDLM	262B	1024	702.586 (12.028)	287.290 (3.186)	160.357 (2.475)	107.756 (1.721)	77.767† (1.526)
mixture	GIDD ($p_{\text{unif}} = 0.0$)	131B	512	2980.247 (98.176)	1036.871 (33.130)	564.206 (23.818)	421.737 (13.676)	368.391 (11.134)
	GIDD ($p_{\text{unif}} = 0.1$)	131B	512	708.316 (14.833)	399.776 (5.931)	288.078 (7.464)	251.572 (7.899)	230.243 (2.287)
	GIDD ($p_{\text{unif}} = 0.2$)	131B	512	763.013 (21.389)	445.039 (5.149)	346.500 (8.506)	308.073 (5.285)	287.921 (6.380)
continuous	DUO	524B	1024	272.043 (5.837)	197.862 (1.812)	171.522 (2.292)	156.682 (3.371)	150.533 (3.075)
uniform	SEDD	131B	512	942.300 (43.081)	361.819 (8.608)	246.701 (6.430)	207.270 (0.598)	197.105 (2.578)
	GIDD	131B	512	2194.880 (123.065)	483.540 (12.116)	279.374 (8.078)	232.096 (6.802)	209.727 (2.702)
	Neural CTMC	131B	512	596.250 (18.507)	265.181 (6.327)	200.394 (3.306)	189.205 (3.541)	194.956 (4.384)
	M2S (ours)	**131B**	**512**	**260.997**† (3.437)	**190.589**† (1.168)	**155.743**† (1.508)	**146.490** (3.407)	**141.740** (1.195)

Table 6: GPT-2 Large generative PPL on the same OpenWebText samples (lower is better). Standard deviations are shown in parentheses. Bold marks the best result among checkpoints with maximum length 512; † marks the overall best. M2S uses Bayes sampling on the cosine grid; the pure-uniform SEDD row uses Euler sampling.

Type	Method	Train	Len.	Sampling steps				
				16	32	64	128	256
mask	SEDD	682B	1024	230.571 (1.500)	136.543 (1.691)	95.625 (1.052)	74.196 (0.624)	60.045 (0.400)
	MDLM	262B	1024	243.047 (2.291)	137.317 (0.633)	93.965† (0.829)	72.243† (1.035)	57.952† (0.987)
mixture	GIDD ($p_{\text{unif}} = 0.0$)	131B	512	566.066 (10.590)	297.354 (4.299)	205.251 (5.606)	169.339 (3.386)	155.875 (3.181)
	GIDD ($p_{\text{unif}} = 0.1$)	131B	512	239.518 (2.855)	166.977 (1.344)	136.454 (2.193)	124.020 (2.357)	118.165 (1.459)
	GIDD ($p_{\text{unif}} = 0.2$)	131B	512	245.696 (4.107)	176.264 (1.199)	150.196 (2.064)	140.027 (2.121)	134.277 (1.256)
continuous	DUO	524B	1024	132.835† (1.205)	108.972 (0.844)	100.086 (1.525)	95.529 (1.502)	93.433 (1.005)
uniform	SEDD	131B	512	334.007 (10.592)	171.825 (2.967)	130.487 (2.213)	115.904 (0.616)	112.059 (0.664)
	GIDD	131B	512	574.170 (15.752)	202.289 (4.054)	139.239 (1.750)	123.097 (1.535)	115.513 (0.909)
	Neural CTMC	131B	512	227.430 (3.214)	135.640 (2.144)	113.019 (0.772)	109.279 (0.727)	111.516 (1.070)
	M2S (ours)	**131B**	**512**	**134.526** (**1.115**)	**108.765**† (**0.904**)	**95.384** (**0.326**)	**91.197** (**1.308**)	**89.204** (**0.499**)

D SEDD Bayes Realizability: Projection Repair, Violation Rates, and Sample Quality

This appendix develops and evaluates projection repair for SEDD scores. We first distinguish affine reconstruction, which reveals whether a score has a valid clean-token posterior, from simplex-constrained projection, which maps an non-realizable score to the nearest Bayes-realizable one. We then measure Bayes-realizability violations during SEDD sampling. Finally, we compare sample quality before and after projection repair at a fixed checkpoint and describe the projected SEDD sampler.

D.1 Projection Repair

For current token k, time t, and an SEDD candidate-score vector s, we seek the valid clean-token posterior whose induced score is closest to s:

$$\mu^{\text{proj}}(s) = \arg\min_{\mu \in \Delta^{K-1}} \|B_{t,k}\mu - s\|_2^2, \qquad s^{\text{proj}}(s) = B_{t,k}\mu^{\text{proj}}(s).$$

Thus $s^{\text{proj}}(s)$ is the Euclidean projection of s onto the set of scores produced by valid posteriors, $\mathcal{P}_{t,k} = B_{t,k}(\Delta^{K-1})$. Euclidean projection onto the probability simplex is a standard optimization primitive (Duchi et al., 2008; Condat, 2016). Our objective measures distance after the linear score map $B_{t,k}$, so Section D.4 derives a specialized solver for the uniform bridge.

The simplex constraint is what makes this projection nontrivial. The matrix $B_{t,k} \in \mathbb{R}^{(K-1)\times K}$ has full row rank, so if μ were allowed to be an arbitrary real vector, every $s \in \mathbb{R}^{K-1}$ could be written exactly as $B_{t,k}\mu$ and the projection would leave s unchanged. Even after imposing $\mathbf{1}^\top \widetilde{\mu} = 1$, every s has a unique affine inverse, but this inverse may contain negative entries. Requiring $\widetilde{\mu} \in \Delta^{K-1}$, including nonnegativity, is what restricts the score to $\mathcal{P}_{t,k}$.

For the uniform bridge, this affine inverse has a closed form. Let $n = K - 1$, $d_t = 1 - \rho_t$, and $b_t = d_t/\rho_t$. Then

$$m(s) = \frac{\sum_{y \neq k}(s_y - \rho_t)}{d_t(n + \rho_t^{-1})}, \qquad \widetilde{\mu}_k = 1 - m(s), \qquad \widetilde{\mu}_y = \frac{s_y - \rho_t - d_t m(s)}{b_t}.$$

It is important to distinguish reconstruction from projection. The SEDD head produces

$$s_y^{\mathrm{SEDD}} = \exp(a_{\theta,y}), \qquad y \neq k.$$

Applying the affine inverse above gives a unique signed vector $\widetilde{\mu}(s^{\mathrm{SEDD}})$ with $\mathbf{1}^\top \widetilde{\mu} = 1$. Reapplying the bridge defines the reconstruction

$$s^{\mathrm{rec}} := B_{t,k}\widetilde{\mu}(s^{\mathrm{SEDD}}) = s^{\mathrm{SEDD}}.$$

The last equality is an algebraic identity on the full affine hyperplane, so s^{rec} is a numerical consistency check, not a repair. In particular, a negative entry of $\widetilde{\mu}$ can coexist with $s^{\mathrm{rec}} = s^{\mathrm{SEDD}}$.

Bayes realizability requires $\widetilde{\mu} \in \Delta^{K-1}$. Consequently, only the simplex-constrained projection changes a score outside the bridge polytope:

$$\mu^{\mathrm{proj}} := \mu^{\mathrm{proj}}(s^{\mathrm{SEDD}}), \qquad s^{\mathrm{proj}} := B_{t,k}\mu^{\mathrm{proj}}.$$

Thus $s^{\mathrm{proj}} = s^{\mathrm{SEDD}}$ if and only if $\widetilde{\mu} \in \Delta^{K-1}$.

Section D.4 derives the scalar-threshold solver used by the projected SEDD sampler.

D.2 Bayes-Realizability Violation Rates

We use three sets of samples for different purposes: four trajectories for the diagnostic in Figure 2(b), 128 trajectories to measure Bayes-realizability violations, and 1,024 paired sequences to evaluate projection repair. Because the three analyses use different units, their reported percentages should be interpreted separately.

Table 7: Data used in the three SEDD score analyses. The first two rows inspect SEDD scores while leaving the generated trajectories unchanged. The third applies projection repair during sampling.

Data	Purpose	Reported unit
4 SEDD trajectories	Negative-weight diagnostic in Figure 2(b)	Step-averaged candidate-score and position rates
128 SEDD trajectories	Bayes-realizability violation rates	8,388,608 position-step score vectors
1,024 paired sequences	Sample quality before and after projection repair	67,633,152 position updates and sequence quality

For the first two analyses, we used the original SEDD scores throughout sampling. At each state and position, we recorded the complete score vector and checked whether every coordinate lay in the required interval and whether the complete vector was induced by a valid clean-token posterior. These checks did not alter the generated trajectories. Only the projection experiment in Section D.3 replaced s^{SEDD} with s^{proj} during sampling.

Sampler-weight violations. We ran four length-512 sequences with the 128-step SEDD sampler. At each step we evaluated the SEDD score s^{SEDD} and its projection on the same current states. Averaged over steps, 9.781% of off-diagonal candidate scores violated the coordinate box, and 37.632% of positions had at least one negative pre-normalization weight. Projection removed all observed negative weights, giving the diagnostic in Figure 2(b).

For the 128-trajectory analysis, we used the partition in Eq. (13). To distinguish material violations from numerical effects near the boundary, we split $\mathcal{C}_{t,k} \setminus \mathcal{P}_{t,k}$ at $\varepsilon_\mu = 10^{-6}$. Table 8 reports the resulting four classes over 512 positions and 128 sampler steps per trajectory.

Table 8: SEDD score vectors over 128 trajectories, using $\mathcal{P}_{t,k} = B_{t,k}(\Delta^{K-1}) \subsetneq \mathcal{C}_{t,k} \subsetneq \mathbb{R}^{K-1}_{>0}$. The boundary row separates numerical from material violations.

Position-step class	Criterion	Fraction	Count
Outside $\mathcal{C}_{t,k}$	$s^{\mathrm{SEDD}} \notin \mathcal{C}_{t,k}$	25.0177%	2,098,635
In $\mathcal{C}_{t,k} \setminus \mathcal{P}_{t,k}$, material	$s^{\mathrm{SEDD}} \in \mathcal{C}_{t,k} \setminus \mathcal{P}_{t,k}$ and $\min \widetilde{\mu} < -10^{-6}$	56.2310%	4,716,994
In $\mathcal{C}_{t,k} \setminus \mathcal{P}_{t,k}$, boundary	$s^{\mathrm{SEDD}} \in \mathcal{C}_{t,k} \setminus \mathcal{P}_{t,k}$ and $-10^{-6} \le \min \widetilde{\mu} < 0$	18.7419%	1,572,188
In $\mathcal{P}_{t,k}$	$s^{\mathrm{SEDD}} \in \mathcal{P}_{t,k}$	0.0094%	791

Across position-steps, 25.0177% lay outside $\mathcal{C}_{t,k}$, 56.2310% were material violations in $\mathcal{C}_{t,k} \setminus \mathcal{P}_{t,k}$, 18.7419% formed its numerical boundary band, and only 0.0094% lay in $\mathcal{P}_{t,k}$. The 56.2310% rate was counted directly because the marginal coordinate and posterior-sign violations overlap (1.2584%).

D.3 Sample Quality after Projection Repair

To isolate the effect of projection repair, we kept the pure-uniform SEDD checkpoint and sampler fixed and changed only the score passed to the sampler. It received either s^{SEDD} or $s^{\mathrm{proj}} = B_{t,k}\mu^{\mathrm{proj}}$. For each setting, we generated 1,024 sequences of length 512 with 128 sampling steps. Each pair started from the same initial noise and used the same random numbers and final denoising update.

With projected scores, none of the observed pre-normalization sampler weights was negative. We decoded the samples with the GPT-2 tokenizer and measured external GenPPL using the same Gemma 2 9B evaluator as in the main language experiment. Table 9 summarizes the comparison; confidence intervals are estimated by sequence bootstrap, with samples paired across the two settings.

Table 9: Effect of projection repair at a fixed SEDD checkpoint. The checkpoint and sampler are identical in both rows; only the score input changes. Intervals are 95% sequence-bootstrap confidence intervals.

Score input	Average NLL	External GenPPL	95% GenPPL interval
s^{SEDD}	5.31615	203.60	[196.64, 210.45]
$s^{\mathrm{proj}} = B_{t,k}\mu^{\mathrm{proj}}$	5.16518	175.07	[169.10, 181.19]

Using projected scores lowered external GenPPL from 203.60 to 175.07, a 14.0% reduction. The paired bootstrap consistently favored projection, and all generated sequences were distinct in both settings, so the gain was not explained by sequence collapse.

Because the checkpoint and sampler were unchanged, this comparison isolates the effect of projection repair at inference time for this checkpoint. It does not imply that post-hoc projection is equivalent to training M2S; external GenPPL is also an external measure of sample quality, not the model's likelihood.

D.4 Projected SEDD Sampler

The uniform bridge permits an exact reduction of the simplex-constrained projection to a scalar threshold. Let $u = \mu_{-k}$, $z = s^{\mathrm{SEDD}}_{-k} - \rho_t \mathbf{1}$, $n = K - 1$, $d = 1 - \rho_t$, and $b = d/\rho_t$. Since $\mu_k = 1 - \mathbf{1}^\top u$, Eq. (7) gives

$$\min_{u \ge 0,\ \mathbf{1}^\top u \le 1} \left\| (bI + d\mathbf{1}\mathbf{1}^\top)u - z \right\|_2^2 . \tag{20}$$

For the objective in Eq. (20), define

$$c = b^2, \qquad h = d^2\left(n + \frac{2}{\rho_t}\right), \qquad g_y = bz_y + d\sum_{v \ne k} z_v .$$

The KKT conditions give

$$u_y = \frac{[g_y - \theta]_+}{c}.$$

If the mass constraint is inactive, the scalar threshold satisfies $\sum_{y \neq k} u_y = \theta/h$ with $\theta \in [0, h]$. If it is active, it satisfies $\sum_{y \neq k} u_y = 1$ with $\theta \in [h, \max_{y \neq k} g_y]$. At the branch point $\theta = h$, let $m_h = c^{-1} \sum_{y \neq k} [g_y - h]_+$. The mass constraint is active exactly when $m_h > 1$. In either branch, the mass on the left decreases monotonically while the target on the right is nondecreasing, so batched bisection solves all positions without sorting the vocabulary. Here $[a]_+ = \max\{a, 0\}$. The resulting projection is applied at every position before the standard SEDD update, as summarized in Algorithm 4.

Algorithm 4 Projected SEDD Sampler

Require: trained SEDD model, reverse-time grid $t_0 > \cdots > t_M$, terminal distribution

Sample X_{t_0} from the terminal distribution
for $m = 0, \ldots, M-1$ **do**
  $x \leftarrow X_{t_m}$
  Evaluate the SEDD score s_i^{SEDD} for every position i
  for all positions i **in parallel do**
    $k \leftarrow x^i$ and $\mu_i^{\mathrm{proj}} \leftarrow \mu^{\mathrm{proj}}(s_i^{\mathrm{SEDD}})$ using Eq. (11)
    $s_i^{\mathrm{proj}} \leftarrow B_{t_m,k} \mu_i^{\mathrm{proj}}$
  end for
  Apply the standard SEDD update with s^{proj} to sample $X_{t_{m+1}}$
end for
Apply the standard final denoising update
return X_0

E SAMPLE GENERATION

The following passages are randomly selected excerpts from samples generated by the M2S model using 128 sampling steps.

Sample 1

The European Union (EU) said the European Union would "fight" Russia's aggression against ISIS with "love for neighbour" and achieve a "defaulted" nuclear agreement and prevent the launch of a nuclear weapon.

Among the top pro-Moscow-Kremlin claims about the Minsk peace agreement were that in 2014, however the war a far east pro-Russian separatists had stopped their long and violent clashes with Russia.

That was still false now, the EU leaders said in a joint statement delivered by German Chancellor Angela Merkel late on Monday. They claimed that ceasefire negotiations would continue after 2014, but added that their "government has to compromise," before rushing to Minsk.

This came during a brief discussion with President Merkel at a joint news conference of the Joint Forces Committee of the Russian Aerospace Forces.

Sample 2

A Republican proposal to change Senate rules would force the Senate budget in line for 10 percent, but under current law the regs spending is $77 million.

The budget proposal submitted on Tuesday by Sen. Marco Rubio, R-Fla., the Senate's third stiffer member and didn't include the first shortfall, a Republican said, one that he now faces would cost a Republican incumbent $3,000 in the U.S. Senate, the unnamed Republican said.

While Senate Democrats were unavailable for comment about the Rubio proposal, the Senate's budget from 2009 to fiscal 2013 was $120 million, with Senate Majority Leader Harry Reid, D-Nev., calling for savings to narrow the gap.

Whitehouse communications director Matt Lunt responded, saying that while any new bill would be pork, full funding would have been required, if the usual funds made financial sense.

Leavint Reid, D-Nev., had not responded to questions about the proposal.

According to the Senate, Rubio would have requested $14 million—a sum that became somewhat problematic in 2011. In the final 15 days that he had given into campaign finance, he closed a public college and policy research board, which had no students. The board of education was appointed by Obama.

Sample 3

Apple has revealed a new 5K tablet with a redesigned, bezel-less Touch Cover design. It's a bit cumbersome and a little more of a departure than the first redesign will be expected. It looks harmless, especially considering launching the invitation ring on the new 5800 later this year, with another refreshed version soon.

What we should see in the huge next iPhone 5S is a tablet stand, costing $420,000–$850,000, along with a full top shelf smartphone. Such mini-boxes will serve them the power down and thus the memories, and possibly drive ports are all possible.

You'd most likely wish they'd wrap this thing up, rather than with their own logo but at least that's what we can expect. Apple will also offer the keyboard that pairs with it like you're forced to the top of its sides. Obviously we find that as a nice gesture and you'd think that wouldn't have to be available during launch too, but if you take the $300–$400, the setup will fit very nicely.

Apple has also prepared a new iPad, just so you can have it with it too, but this idea is a myth. Not bad, since the iPhone offers a 4K resolution to 4k video and up to 12GB should be very respectable, but its not why the phone comes with Android 4.1/2.0 OS. Stay tuned for more hands-on details about the latest, and we'll report back to iPad TV Line in the future for more info.

Sample 4

Junior Robinson is a genius. He's always been blessed with natural size and speed. Lemon speed is about any player. But the speed he cultivated in the offseason, and he has averaged just six minutes in practice ice in a season starting the past three is at all impressive.

And it's not something you can't. People can carry a a couple of minutes in five preseason games, able to attend a freeze from assembling the NHL roster. They are versatile, guys.

The 24-year-old former 18th overall selection chose Detroit as the first-round pick of the 2012 NHL Pending Draft, two years after averaging 25 points in 24 games in various stints at 82 points, and starting since this past October, for Detroit, in the same spot some of his most promising stints in short-re 30 games, and 24 points in 27, for a rookie season, he performed very well.

He returned this offseason, maximizing his own potential by one season at the NHL—subject to arbitration until 2012 at his due date. In order for the Red Wings to make use of roster room, they need just a couple of 25-man reinforcements in their rotation, without shipping John O'Neil or others for a roster spot.

Sample 5

The year is once again in 2015, as the number of layoffs has been a record year high. Consequently, the banking giant Chase & Chase Bank recently reported that the sector had lost more than 19,000 jobs over the past three years.

Read more

Markers, who are criticised for feeling a stagnation, usually cite strong employment growth trends to speak to its current trends. "Today, the unemployment rate is down by 330 per cent since 2010," said in its annual report. "And that depends on the shifts in opportunities in business."

However, in 2014–2015, one Fargo Ontario office employing 17,000 from was downsized—by 5,200—it backtracked this. It eliminated more than two-thirds of production, filling out the need among its banking sector employees.

Earlier, the Bank had hinted that it may tighten its finances more. It released evidence that a low in wage workers caused the unemployment cut by almost $14,000 to workers before a tax of 0.25 percent. This would push the GDP worse in 2015. The recovery could not have been struck without these defects.

As Fargo is expected to come out soon about the disinitation of its economy, the report, however, states the need to change infrastructure in a more sustainable way.

The sale of Fargo operating Business Hummer division was taken in as the expense by example cited above.

Sample 6

The main web apps on the desktop are 8GB allowed KB. They have a 40% download speed, and get a chunk of that from the Google Play Store, you can bet that Google also intends to bring the fun back into the store by having the same much functionality as in Firefox browsers like Chrome and Safari.

On the mobile side, we have Hoothing, Magazines, Cycle Enhancement Browser, and Goggles apps, though this technically means you can upgrade later this year if you tried and missed before. Full apps can go on now too.

Apps are also available on self-link for third-party developers to help with the Web apps as well as the Mobile settings also. However, such details were not available at this stage of the above launch.

Chrome users, however, need to enable options even for Android. That's been going on for many years now, with the Google Chrome desktop live app (which has been live for nearly a year now) and the mobile browsers on the U.S. and Chinese Web versions. Still enough stuff, you know?

Apps first launched in early September will now be supported including on Google's website, therefore making the Google Chrome app easier to sign up. There is still an open beta. All things are ready for everyone, always and soon.

Google is testing out the redboards in China and other countries and is testing the locations available on the site.

Sample 7

When you sign up for an AWS Elastic Load Network (ELT), get connected with your email or even the online language in your applications, it gets you go immediately. This may have only gotten so rare in the past, but the full power of computing is up for grabs in cloud computing and with its acquisition by PxCom, Inc. its cloud-delivered services has always been unusually designed with one important one: customer delivery in mind.

"As the dominant broadband-only provider in the world, Seattle-based AWS has particular plans to modernize its computing infrastructure and give it real-time on-demand service and production-ready broadband customer delivery, with direct access to its delivery systems," its agency says in a release.

It also understands that it can be a layer used by its other large providers and coordinate their services to their customers's data centers and connect them in their case at home, an indication that not all of the customer experiences in the cloud are necessarily different from each of the largest ISPs' portfolio companies.

P&Com, founded in partnership with Yormia and Escape Networks in January of 2014, is a provider of enterprise cloud service for "cloud" services AWS, Microsoft Azure, Amazon Dropbox, and Google Docs.

Sample 8

Microsoft has announced it will announce and released of AI assistant in around the holidays.

Last Thursday, Microsoft revealed that its AI assistant device, which was shown as a prototype over at CES, "is the successor to Google tablet- AI operating system with a text and call interface." It will likely be paired with the upcoming Microsoft-powered Xbox One.

Now there is another hugely interesting and interesting rumor floating around, and it got execs caught at CES anyway. An official Windows blog post is claiming that the new software will feature in users' mobile systems, and it will provide a "personal voice assistant for device hubs with limited sharing space," Microsoft said. Unlike Siri, it works with older storage devices, but there stored multiple SD cards or at least a volume on them. That's obviously a waste of time, as your device can be gone out-of-state after being hooked up to the SD card.

The AI device is all tied to Microsoft's tech future, the AI+ operating system. Siri that will be capable of communicating with Internet-connected PCs, and the goal will be to work with iOS, Mac, and other groups of devices in common, require AI+ using operating system.

We said, but wait. The full video is available here. μ

Sample 9

With a new law pending before Congress, California faces an uphill labor battle, chief executive officer Xavier Becerra said recently.

There is an uproar over California's pay rate for the corporate behemoth that sells health care, which makes home appliances and other products.

Western leaders on Capitol Hill still struggling with lower wages in countries that lean to union workers and rely on pliers' compensation. But here were signs of a change of heart that amounts to a higher rate for California, partly because taxpayers in the state have invested heavily in company labor.

The federal labor department said rates in its factories in Tennessee were also getting higher. "As there has been such a measure of growth in the states, labor costs have risen," Mr. Cosino said in ODFO's fact sheet March 20.

Sample 10

The International Maoist-China forum called in return for "a world leader in international stability and international goods and services".

"China is proposing that its vision of state-denialing is a new machine, based on international cooperation, and its own market, rather than a country who would use its culture and opportunity to lead," Hu Bingkao told participants, in an announcement of the convention on Saturday.

Saturday, as the party celebrated the first anniversary of the start of the Maoist Commission, which was held in October when it warned of economic and social unrest in the country.

The government, at the time felt there was no warning about the new trend, which included the rise of crack economy, ultra-high unemployment rates, levels of corruption, and the growth of militant groups.

The convention took place just north of the Helfen, where also collected in some over 60 addresses take place annually.

"Our society will be part of the discussion, but our ideological reason is simply the objective," the Bingkao said.

Meanwhile, Hu said People needed to grow the environment, dealing with drug dealing, and said that the methods to boost social stability was needed to respond to problems for the population.