MIXTURE-OF-LATENTS FOR FEW-STEP DISCRETE DIFFUSION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete diffusion and flow models can generate high-quality sequences, but they usually require many denoising steps. Distilling them into few-step models is challenging because a single student step predicts many positions independently, which can lose important dependencies between tokens. We show that matching each position separately is not enough to reproduce the teacher's joint transition. Prior work addresses this issue with a shared continuous latent variable, but its training requires an approximation that can fail to make effective use of the latent. We propose **M**ixture-of-**L**atents for **D**iscrete **D**iffusion **D**istillation (MLD3), which instead uses a small finite set of components selected by a learned router. This makes the likelihood of each sampled teacher transition *exactly* computable and allows the student to model correlated updates directly. We also propose a novel sampling approach that exploits the mixture components at a cost close to that of a factorized step. Notably, MLD3 naturally reduces to a standard masked diffusion student with one component and adds negligible parameter overhead as new components are added. Across text, molecules, regulatory DNA, and image generation, MLD3 consistently improves few-step generation by roughly $2\times$ over the baselines, with the largest gains on tasks with stronger dependencies across positions.

1 INTRODUCTION

Discrete diffusion and flow models have recently emerged as a strong alternative to autoregressive models for sequence data (Austin et al., 2021; Lou et al., 2023; Campbell et al., 2024; Gat et al., 2024). Instead of generating one token at a time, they corrupt sequences toward a simple prior and learn to denoise every position in parallel, which has proven effective on text as well as on scientific data such as molecules and DNA (Park et al., 2025). Masked discrete diffusion, in which the prior replaces tokens with an absorbing mask symbol, is the most widely adopted case of this family (Sahoo et al., 2024; Shi et al., 2024; Ou et al., 2025).

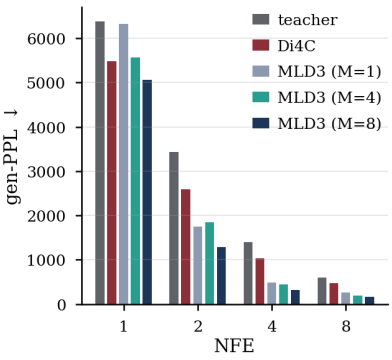

Figure 1: **Few-step generation on WikiText-103**. Generative perplexity, where lower is better, across the number of function evaluations (NFEs). MLD3 achieves lower perplexity than the MDLM teacher and Di4C across the few-step regime.

A significant limitation of discrete state-space models is their slow sampling. High-quality samples typically require a long sequence of denoising steps, and the discrete state space does not directly admit the higher-order solvers used to accelerate continuous diffusion models. A common approach is therefore to distill a pretrained teacher into a student that generates samples in only a few steps, either by compressing the teacher's multi-step trajectory (Deschenaux & Gulcehre, 2025; Hayakawa et al., 2024) or by learning a straighter source-to-target coupling (Yoo et al., 2026; Liu et al., 2022; Park et al., 2025).

However, in all of these students, a single step samples all positions *independently* from a product of per-position categorical distributions. This factorization is exact when a step reveals one position at a time, but a few-step sampler must reveal several positions in each step, and positions revealed together are rarely independent. A factorized student model can match every per-position marginal

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.

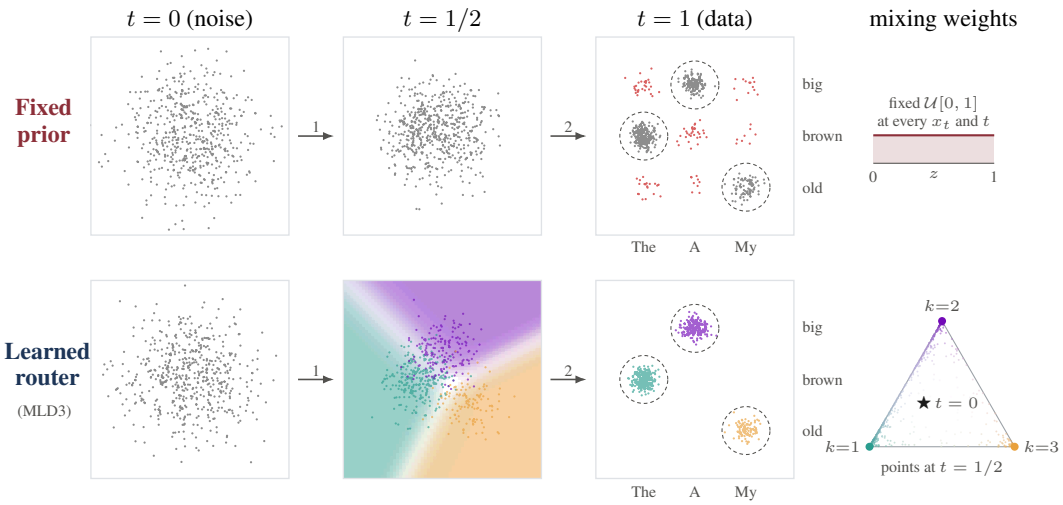

Figure 2: **Toy illustration of correlated few-step generation and the role of the learnable latent**. A single step has to fill 2 positions of the same sentence at once. Each axis represents one position in the sequence, and the 3 circled clusters are the only coherent ways to fill them, namely "The . . . brown" as in *The quick brown fox*, "A . . . big" as in *A very big dog*, and "My . . . old" as in *My dear old friend*. Both students generate the two positions from noise at $t = 0$ to data at $t = 1$, with the positions conditionally independent given a shared latent. **Top**: under the surrogate collapse of Section 3.3, the latent does not distinguish the modes, and hence each position follows its marginal distribution *independently*. This produces probability mass on incoherent cross-mode combinations, such as "A" with "old" (in red). **Bottom**: MLD3 uses a finite latent $k \in \{1, 2, 3\}$ with state-dependent router weights $w_k(x_t, t)$. The colored regions at $t = \frac{1}{2}$ show which component the router favors for each intermediate state, and the simplex on the right shows the corresponding mixture weights. Sharing the selected component across positions couples their predictions jointly in a denoising step that concentrates the samples on the coherent data modes for few-step generation.

of the teacher's transition while still failing to reproduce its joint structure, and its step divergence is at least the transition's total correlation (TC), which no factorized student can remove (Watanabe, 1960; Csiszár, 1975; Huang et al., 2022). Figure 2 illustrates this failure in an example sequence with two positions to be filled, where matching the coordinate-wise marginals places probability on incoherent cross-mode combinations, while conditioning both positions on a shared latent can preserve the joint modes.

A solution is known in principle. Conditioning a step on a latent variable shared across positions preserves factorization within each component while allowing dependence after the latent is marginalized out. The resulting mixture of factorized kernels can represent different coherent continuations, as illustrated in Fig. 2. Di4C (Hayakawa et al., 2024) introduced this construction for discrete diffusion using a continuous latent λ. However, evaluating its marginal step likelihood requires an integral over λ that is generally intractable. A common workaround is to average the component-wise divergences rather than evaluate the divergence to the marginal mixture. Di4C anchors its student to the teacher with a *surrogate*, obtained from the convexity bound of Jensen's inequality,

$$\mathrm{KL}\Big(q \,\Big\|\, \mathbb{E}_\lambda\big[p^\theta(\cdot\,;\lambda)\big]\Big) \;\le\; \mathbb{E}_\lambda\Big[\mathrm{KL}\big(q \,\|\, p^\theta(\cdot\,;\lambda)\big)\Big], \tag{1}$$

where q is the teacher's target for the step and $p^\theta(\cdot\,;\lambda)$ is the student's latent kernel conditioned on λ. Because each $p^\theta(\cdot\,;\lambda)$ is factorized, the surrogate scores each component independently and depends on the target only through its per-position marginals. We prove that, even when the target q carries dependence between positions, the minimizer of this surrogate is a factorized kernel, for any prior over the latent and number of mixture components. We call this failure *surrogate collapse*. Increasing the number of latent samples improves estimation of the surrogate but does not change its optimum, which motivates our central question:

Can we learn correlated discrete token updates with a shared latent whose marginal likelihood can be evaluated exactly?

We propose **M**ixture-of-**L**atents for **D**iscrete **D**iffusion **D**istillation (MLD3), a finite mixture of M factorized kernels with a learned router. The finite latent makes the step likelihood exactly computable, enabling joint training of the router and components without sampling the latent. With one component, MLD3 naturally recovers a factorized masked diffusion student, and additional components capture cross-token dependence with little parameter overhead. Across six datasets spanning text, molecules, DNA, and image tokens, MLD3 improves few-step generation, including generative perplexity on WikiText-103 (Fig. 1).

Our **main contributions** can be summarized as follows:

1. We analyze two limitations of few-step distillation of discrete state-space models: the irreducible dependence error of factorized student transitions, and the collapse of the convexity bound through which latent-conditioned students are anchored to their teacher.
2. We propose MLD3, a finite mixture of factorized transition kernels with a learned router that enables exact marginalization over the shared latent during training.
3. We introduce an efficient spread-consensus sampler that exploits the learned mixture while evaluating only a small number of components per step.
4. Across six benchmarks in text, molecules, DNA, and images, MLD3 improves few-step generation most strongly when the teacher's transitions contain cross-position dependence.

2 Preliminaries and Background

2.1 Distillation of Masked Diffusion Models

Notation. We consider sequences $\boldsymbol{x} = (x^1, \dots, x^L)$ of length L over a vocabulary $\mathcal{S}$ of size V, and corrupted sequences additionally use an absorbing mask symbol $\texttt{[MASK]} \notin \mathcal{S}$, so that a partially masked sequence belongs to $\mathcal{X} = (\mathcal{S} \cup \{\texttt{[MASK]}\})^L$. Superscripts index positions, subscripts index time, $\boldsymbol{X}_s$ and $\boldsymbol{x}_s$ denote the random sequence produced by a step and its realization, and $\mathcal{M}(\boldsymbol{x}_t)$ is the set of positions still masked in $\boldsymbol{x}_t$. App. Table 1 further summarizes all the notation.

Transition kernel. A fully masked sequence at time $t = 0$ is unmasked to clean data at $t = 1$, under a monotone schedule $\kappa_t \in [0, 1]$ that keeps each clean token independently with probability κ_t (Austin et al., 2021; Sahoo et al., 2024). Generation starts from the all-mask sequence and applies one network call per denoising step, and a few-step method reduces the number of these function evaluations (NFEs). Because masking is absorbing, a step from time t to a later time $s > t$ keeps the revealed positions fixed and reveals each masked position with probability $\frac{\kappa_s - \kappa_t}{1 - \kappa_t}$. A masked diffusion student model turns its per-position predictions into such a step with a product over positions, which we call a *factorized transition kernel*,

$$p^{\theta}_{s|t}(\boldsymbol{x}_s \mid \boldsymbol{x}_t) \;=\; \prod_{i=1}^{L} P^{\theta}_i(x^i_s \mid \boldsymbol{x}_t, t, s), \tag{2}$$

where P^{θ}_i places all of its probability on x^i_t, that is, on $\boldsymbol{x}_t$ at an already revealed position i, and is a distribution over $\mathcal{S} \cup \{\texttt{[MASK]}\}$ otherwise. MDLM (Sahoo et al., 2024) and absorbing SEDD (Lou et al., 2023) are the canonical instances. We call the kernels of this form the factorized family $\mathcal{F}_1$, and write $p^{\psi}_{s|t}$ for the teacher's transition over the same interval, which we obtain by running the teacher's sampler for several smaller sub-steps.

The analytic reverse step. A masked diffusion model turns a posterior p_0 over clean tokens into a step in closed form. At time t, each masked position is updated according to

$$q(x^i_s = v) \propto p_0(v)\,(m_t - m_s) \quad (v \neq \texttt{[MASK]}), \qquad q(x^i_s = \texttt{[MASK]}) \propto m_s, \tag{3}$$

where $m_t = (1-\varepsilon)(1-\kappa_t)$ is the masking probability under the log-linear schedule, and the revealed positions are copied unchanged to the next step. This schedule sets the rate at which the positions are revealed, which guarantees that a K-step trajectory removes every masked token within K steps.

Distillation. Given a teacher ψ and a K-step sampling schedule $\{t_j = \frac{j}{K}\}_{j=0}^{K}$, we train a student model by drawing a time pair (t, s) with $t < s$ and a partially masked state $\boldsymbol{x}_t$. We then run the

teacher from t to s to obtain a sample $\boldsymbol{x}_s \sim p^{\psi}_{s|t}(\cdot \mid \boldsymbol{x}_t)$, and train the student model to minimize the negative log-likelihood. In expectation over the teacher's samples, this cross-entropy is

$$\mathbb{E}_{\boldsymbol{x}_s \sim q}\big[-\log p^{\theta}_{s|t}(\boldsymbol{x}_s \mid \boldsymbol{x}_t)\big] \;=\; \mathrm{H}(q) \;+\; \mathrm{KL}\big(q \,\big\|\, p^{\theta}_{s|t}(\cdot \mid \boldsymbol{x}_t)\big), \qquad q = p^{\psi}_{s|t}(\cdot \mid \boldsymbol{x}_t), \tag{4}$$

and since $\mathrm{H}(q)$ does not depend on θ, minimizing the cross-entropy minimizes the Kullback-Leibler (KL) divergence from the teacher's transition to the student's at each step. We call this the *step-KL objective*, and the theoretical results below are related to its expected form at a fixed $\boldsymbol{x}_t$.

2.2 Factorization error

The positions that are correlated in the data but revealed in the same step are sampled *independently* under Eq. 2. They can thus disagree more often than under the teacher, for example, by placing a plural subject beside a singular verb. The standard measure of this error is the total correlation $\mathrm{TC}(q) = \mathrm{KL}(q \,\|\, \prod_i q_i) = \sum_i \mathrm{H}(q_i) - \mathrm{H}(q) \geq 0$ of a distribution q on $\mathcal{X}$ with marginals q_i, which is zero if and only if q factorizes across positions. Writing $\Pi(q) = \prod_{i=1}^{L} q_i$ for the product of the marginals of q, we can state the error as an identity.

Theorem 1 (Factorization error)**.** *Let q be any distribution on $\mathcal{X}$ with marginals q_i. For every factorized kernel $p = \prod_i p_i \in \mathcal{F}_1$, we have the decomposition*

$$\mathrm{KL}(q \,\|\, p) \;=\; \mathrm{TC}(q) \;+\; \sum_{i=1}^{L} \mathrm{KL}(q_i \,\|\, p_i), \qquad \textit{hence} \qquad \min_{p \in \mathcal{F}_1} \mathrm{KL}(q \,\|\, p) \;=\; \mathrm{TC}(q), \tag{5}$$

with the minimum attained if and only if $p = \Pi(q)$. For a fixed target transition q, this minimum depends only on the teacher and on the step size.

The proof is provided in App. B.2. Eq. 5 is the classical Pythagorean identity for the information projection onto the set of product distributions (Csiszár, 1975; Cover et al., 1991), where TC is the total correlation of Watanabe (1960). Here, q is the teacher's transition from t to s, and $\mathrm{TC}(q)$ is determined by the step size, which is fixed once the number of steps is chosen. It typically grows with the step size (Figure 10(c)), which is fixed once the number of steps is chosen. The teacher pays this cost only over its much smaller sub-steps, because it builds up the dependence over many small steps that a few-step student model must reproduce within one step.

Even when each of its sub-steps is factorized, the teacher produces a correlated transition over the interval of one step of the student model. Writing $\boldsymbol{x}_{u_0} = \boldsymbol{x}_t, \boldsymbol{x}_{u_1}, \ldots, \boldsymbol{x}_{u_T} = \boldsymbol{x}_s$ for the states that the teacher passes through, we see that the coarse transition $q(\boldsymbol{x}_s \mid \boldsymbol{x}_t)$ marginalizes over the intermediate ones. Conditional on the trajectory, the positions of $\boldsymbol{x}_s$ are independent, because each position either keeps the value assigned by an earlier sub-step or is sampled from the last factorized kernel in the chain, and summing over trajectories introduces dependence between the positions. The coarse transition of the teacher is thus already a mixture of product distributions, and the trajectory is the latent variable that indexes its components.

3 Mixture-of-Latents for Discrete Diffusion Distillation

We propose **M**ixture-of-**L**atents for **D**iscrete **D**iffusion **D**istillation (MLD3), which reformulates the single product model of Eq. 2 by introducing a finite mixture of M products,

$$p^{\theta}_{s|t}(\boldsymbol{x}_s \mid \boldsymbol{x}_t) \;=\; \sum_{k=1}^{M} w^{\theta}_k(\boldsymbol{x}_t, t) \prod_{i=1}^{L} P^{\theta}_{i,k}\big(x^i_s \mid \boldsymbol{x}_t, t\big), \tag{6}$$

where the router, parameterized by a network head, assigns the weight w^{θ}_k to component k at the state $\boldsymbol{x}_t$, and $P^{\theta}_{i,k}$ is the distribution of component k at position i, obtained by applying the analytic reverse step of Eq. 3 to a clean-token posterior. Each component is thus a factorized kernel, and we call $p^{\theta}_{s|t}$ a mixture kernel. The network takes $(\boldsymbol{x}_t, t, k)$ as input, which allows one trained model to run at any number of steps without retraining. We write $\mathcal{F}_M$ for the resulting family of kernels. To sample one step, we draw a single index $k \sim w^{\theta}(\boldsymbol{x}_t, t)$ shared by all positions, and then sample

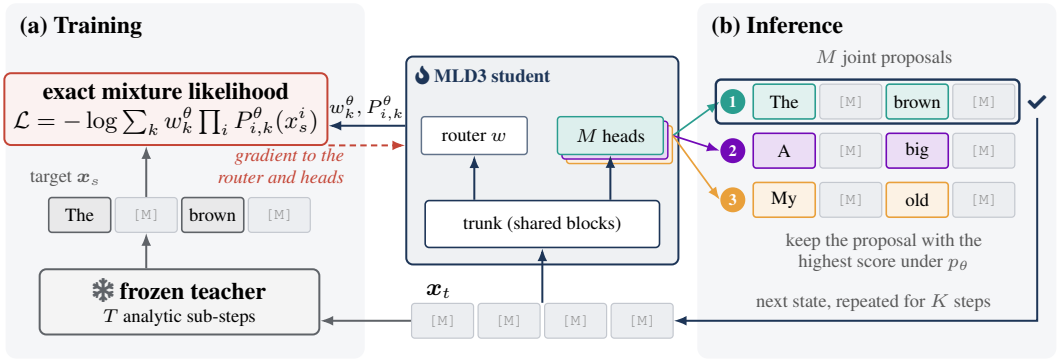

Figure 3: **Training and inference in MLD3.** (a) From a masked state $\boldsymbol{x}_t$, the frozen teacher takes T analytic sub-steps to produce the target $\boldsymbol{x}_s$, which the student model learns to reach in a single forward pass. The trunk feeds a router, which gives the weights w_k^θ, and M component heads, which give the factorized distributions $P_{i,k}^\theta$. (b) At each of the K sampling steps, each evaluated component gives one joint proposal. We score the proposals under the mixture with Eq. 9 and advance the denoising trajectory with the highest-scoring component via our consensus sampler.

every masked position independently from component k. This single sequence-level draw couples all positions revealed in the step.

Because the latent index is finite, the step likelihood is a closed-form sum of M terms. Let $\ell_k(\boldsymbol{x}_s) = \sum_{i \in \mathcal{M}(\boldsymbol{x}_t)} \log P_{i,k}^\theta(x_s^i \mid \boldsymbol{x}_t, t)$ be the log-probability that component k assigns to the newly revealed tokens. Then, we have

$$\log p_{s|t}^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t) \;=\; \log \sum_{k \in [M]} w_k^\theta(\boldsymbol{x}_t, t)\, e^{\ell_k(\boldsymbol{x}_s)}. \tag{7}$$

Here, the sum over positions in ℓ_k is *inside* the summation over components. Eq. 7 thus scores the probability that a single shared component explains all revealed tokens jointly. The expression is differentiable and requires no sampling of k during training. Moreover, the family $\mathcal{F}_M$ grows with M and contains the factorized family, which makes a larger mixture at least as expressive as a smaller one, and a single-component student model naturally reduces to an MDLM student model. A finite index also makes the mixture measurable, because the joint distribution of the index and the step is explicit, so the information that the index carries about a step is available in closed form and is at most $\log M$ (Proposition 6). App. B.6 bounds the correlation captured by a mixture of M components and how many components a transition requires.

3.1 Distillation Objective

We train the student model against a frozen teacher ψ without access to the data likelihood. The objective combines a transition loss, which provides the signal about dependence between positions, with an auxiliary loss and a router regularizer, both defined in App. C.4. All losses are taken in expectation over a distribution ν on the pair of times spanning a step. Given $\boldsymbol{x}_t$, we advance the teacher from t to s in T analytic sub-steps of Eq. 3, and keep the final sequence of this rollout as a hard target. Algorithm 1 states the full procedure, and Figure 3(a) illustrates a training step in MLD3.

The student model scores the target with Eq. 7, and the transition loss is the negative log-likelihood,

$$\mathcal{L}_{\text{trans}}(\theta) \;=\; -\mathbb{E}_{(t,s)\sim\nu}\, \mathbb{E}_{\boldsymbol{x}_t}\, \mathbb{E}_{\boldsymbol{x}_s \sim p_{s|t}^\psi(\cdot|\boldsymbol{x}_t)} \Big[\log \sum_k w_k^\theta(\boldsymbol{x}_t, t)\, e^{\ell_k(\boldsymbol{x}_s)} \Big]. \tag{8}$$

By Eq. 4, this loss equals the step-KL objective up to a constant in θ, and its minimizer is the KL projection of the teacher's transition onto $\mathcal{F}_M$. Intuitively, the finite index allows us to sum over k in closed form. The mixing weights then appear inside the logarithm and receive a gradient, and a component that explains the teacher's samples better than the others increases its own weight. Reducing the objective at every step also controls the quality of the final samples, because the per-step divergences add up along the trajectory and bound the divergence of the final sequence (Proposition 1, with the caveat of Remark 1).

Algorithm 1 MLD3 training.

Require: teacher ψ; student θ with M components; time distribution ν; sub-steps T; weights λ_0, λ

repeat
  $\boldsymbol{x}_1 \sim$ data; $(t, s) \sim \nu$ with $t < s$
  $\boldsymbol{x}_t \leftarrow$ mask each x_1^i with probability $1 - \kappa_t$
  $\mathcal{M} \leftarrow \{i : x_t^i = \texttt{[MASK]}\}$
  $\boldsymbol{x}_s \leftarrow$ teacher rollout from t to s in T sub-steps
  $\big(w, \{p_{0,k}\}_{k=1}^{M}\big) \leftarrow \theta(\boldsymbol{x}_t, t)$ ▷ $C(M)$ passes
  for $k = 1, \dots, M$ **do**
    $P_{i,k} \leftarrow$ reverse-step distribution from $p_{0,k}$ at position i, for $i \in \mathcal{M}$ ▷ Eq. 3
    $\ell_k \leftarrow \sum_{i \in \mathcal{M}} \log P_{i,k}(x_s^i)$
  end for
  $\mathcal{L}_{\text{trans}} \leftarrow -\operatorname{logsumexp}_{k=1}^{M}\big(\log w_k + \ell_k\big)$
  for $i \in \mathcal{M}$ **do**
    $\bar{P}_i \leftarrow \sum_{k=1}^{M} w_k P_{i,k}$
  end for
  $\theta \leftarrow \text{OptimizerStep}(\theta,\ \mathcal{L}_{\text{trans}})$ ▷ Eq. 45
until convergence

Algorithm 2 MLD3 inference.

Require: student θ with M components; length L; steps K; evaluated components m

$\boldsymbol{x} \leftarrow [\texttt{[MASK]}]^L$; $r \sim \mathcal{U}\{0, \dots, L-1\}$
$\pi \leftarrow \text{rotate}\big(\text{bitrev}(L),\ r\big)$ ▷ spread order
for $i = 1, \dots, K$ **do**
  $t \leftarrow i/K$; $\mathcal{M} \leftarrow \{\, \ell : x^\ell = \texttt{[MASK]} \,\}$
  $\big(w, \{p_{0,k}\}\big) \leftarrow \theta(\boldsymbol{x}, t)$ on the shared blocks
  $\mathcal{E} \leftarrow$ top-m components by w
  $\tilde{w} \leftarrow w_{\mathcal{E}} / \sum_{k \in \mathcal{E}} w_k$
  $n \leftarrow$ expected reveals at this step
  $\mathcal{R} \leftarrow$ first n positions of $\mathcal{M}$ in the order π
  for $j \in \mathcal{E}$ **do**
    $\boldsymbol{x}^{(j)} \leftarrow$ fill $\mathcal{R}$ from $p_{0,j}$ via Eq. 3
    $S_j \leftarrow \sum_{\ell \in \mathcal{M}} \log \sum_{k \in \mathcal{E}} \tilde{w}_k P_{i,k}^{\theta}(x^{(j),\ell})$ ▷ Eq. 9
  end for
  $j^\star \leftarrow \arg\max_{j \in \mathcal{E}} S_j$
  $x^\ell \leftarrow x^{(j^\star),\ell}$ for all $\ell \in \mathcal{R}$
end for
return $\boldsymbol{x}$

3.2 Inference

We sample from MLD3 with the analytic reverse step of Eq. 3, starting from the all-mask sequence. At step i of a K-step trajectory, we run the trunk at $t = i/K$ and, for each evaluated component, drop the mask entry from its softmax and obtain a clean-token posterior p_0. Each step then makes three choices, as summarized in Algorithm 2. We call the resulting sampler *spread consensus*.

Spread order. The analytic step reveals each masked position independently. Hence, the number revealed per step is random, and its variance grows with the number of masked positions. We instead reveal exactly the number expected under the schedule, taken in a fixed *spread order*. Before sampling, we order the positions by the bit-reversal permutation of $\{0, \dots, L-1\}$ and rotate this order by a random offset. Each step reveals the next masked positions in this order, and positions revealed together are therefore far apart in the sequence.

Component selection. A step evaluates only the m components with the largest router weights, which we denote $\mathcal{E}$, and renormalizes the mixture over them. Components share the first stack of blocks and differ only in their component-specific blocks, so a step costs $D - L_{\text{lat}} + mL_{\text{lat}}$ block passes, and $m = M$ evaluates the full mixture. A single component may remove the correlation property learned by the mixture, while a wider route raises the computational cost of a step.

Consensus scoring. We keep a single chain and search over $\mathcal{E}$. One forward pass produces the router weights and the posteriors of all components in $\mathcal{E}$, and each component j produces a joint proposal $\boldsymbol{x}^{(j)}$ via Eq. 3. We score each proposal by the log-probability that the mixture over $\mathcal{E}$ assigns to it at every currently masked position, including those that stay masked after the step:

$$S\big(\boldsymbol{x}^{(j)}\big) \;=\; \sum_{\ell \in \mathcal{M}(\boldsymbol{x}_t)} \log \sum_{k \in \mathcal{E}} \tilde{w}_k(\boldsymbol{x}_t, t)\, P_{i,k}^{\theta}\big(x^{(j),\ell} \,\big|\, \boldsymbol{x}_t, t\big), \tag{9}$$

where $\tilde{w}_k$ are the router weights renormalized over $\mathcal{E}$. The chain advances with the highest-scoring proposal: each component proposes a continuation, and the mixture chooses among them, possibly picking a different component at each step (Figure 3(b)). Because the components differ by design, a single forward pass generates systematically different proposals to compare at every step.

3.3 Surrogate collapse

A continuous latent makes the step likelihood an integral over λ, hence the divergence to the teacher cannot be evaluated directly. Di4C therefore trains the term that anchors its student to the teacher through the convexity of the KL divergence, which moves the average over the latent outside the

divergence as in Eq. 1, and then bounds the result by a sum over positions (Hayakawa et al., 2024, App. A.1). Written with our finite index in place of λ and with the target of Eq. 8, the bound scores each component separately,

$$\mathcal{L}_{\mathrm{surr}}(\theta) \;=\; -\,\mathbb{E}_{(t,s)\sim\nu}\,\mathbb{E}_{\boldsymbol{x}_t}\,\mathbb{E}_{\boldsymbol{x}_s\sim p^{\psi}_{s|t}(\cdot|\boldsymbol{x}_t)}\,\mathbb{E}_{\hat{k}\sim w^{\theta}(\boldsymbol{x}_t,t)}\big[\ell_{\hat{k}}(\boldsymbol{x}_s)\big]. \tag{10}$$

The two objectives thus differ only in whether the average over components is taken inside or outside the logarithm. At a fixed $\boldsymbol{x}_t$, we write $\mathcal{L}_{\mathrm{exact}}(p)$ and $\mathcal{L}_{\mathrm{surr}}(p)$ for the inner expectations of Eqs. 8 and 10 under the kernel $p = p^{\theta}_{s|t}(\cdot \mid \boldsymbol{x}_t)$, so that the two losses are their averages over (t, s) and $\boldsymbol{x}_t$. At every $\boldsymbol{x}_s$ with $p^{\theta}_{s|t}(\boldsymbol{x}_s \mid \boldsymbol{x}_t) > 0$, the two are related by the identity

$$\mathbb{E}_{\hat{k}\sim w}\big[-\ell_{\hat{k}}(\boldsymbol{x}_s)\big] = -\log p^{\theta}_{s|t}(\boldsymbol{x}_s \mid \boldsymbol{x}_t) + \mathrm{KL}\big(w \big\| p^{\theta}(\cdot \mid \boldsymbol{x}_s)\big), \quad p^{\theta}(k \mid \boldsymbol{x}_s) = \frac{w_k\, e^{\ell_k(\boldsymbol{x}_s)}}{p^{\theta}_{s|t}(\boldsymbol{x}_s \mid \boldsymbol{x}_t)} \tag{11}$$

where $p^{\theta}(k \mid \boldsymbol{x}_s)$ is the posterior over the component. Eq. 11 is the evidence lower bound with the variational posterior frozen at the prior w, so minimizing the bound minimizes the exact objective plus the penalty $\mathrm{KL}(w \,\|\, p^{\theta}(\cdot \mid \boldsymbol{x}_s))$. This penalty vanishes if and only if every component with positive weight assigns the same probability to $\boldsymbol{x}_s$, that is, when the step carries no evidence about the component.

Theorem 2 (Surrogate collapse)**.** *Fix $\boldsymbol{x}_t$, let q be the teacher's transition over the masked positions, and let $M \geq 1$. For $p \in \mathcal{F}_M$ with weights w and components $P^1, \ldots, P^M$, define the bound of Eq. 10 and the exact objective of Eq. 8 at $\boldsymbol{x}_t$,*

$$\mathcal{L}_{\mathrm{surr}}(p) = \mathbb{E}_{\boldsymbol{x}_s\sim q}\,\mathbb{E}_{\hat{k}\sim w}\big[-\ell_{\hat{k}}(\boldsymbol{x}_s)\big], \qquad \mathcal{L}_{\mathrm{exact}}(p) = \mathbb{E}_{\boldsymbol{x}_s\sim q}\big[-\log p(\boldsymbol{x}_s)\big].$$

(i) The bound depends on the target only through its marginals. *For every $p \in \mathcal{F}_M$,*

$$\begin{aligned}\mathcal{L}_{\mathrm{surr}}(p) \;&=\; \mathrm{H}(q) + \sum_{k=1}^{M} w_k\,\mathrm{KL}\big(q \,\big\|\, P^k\big) \;=\; \sum_i \mathrm{H}(q_i) + \sum_{k=1}^{M} w_k \sum_i \mathrm{KL}\big(q_i \,\big\|\, P^k_i\big) \\ &\geq\; \mathrm{H}(q) + \mathrm{TC}(q),\end{aligned} \tag{12}$$

with equality if and only if $P^k = \Pi(q)$ for every k with $w_k > 0$.

(ii) Every minimizer of the bound is factorized. *Whether w is held fixed or optimized, every minimizer p of $\mathcal{L}_{\mathrm{surr}}$ over $\mathcal{F}_M$ satisfies $p = \Pi(q)$, $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) = 0$, and $\mathrm{KL}(q \,\|\, p) = \mathrm{TC}(q)$.*

(iii) The exact objective uses the latent. $\min_{p\in\mathcal{F}_M} \mathcal{L}_{\mathrm{exact}}(p) = \mathrm{H}(q) + \mathcal{D}_M(q)$, *where* $\mathcal{D}_M(q) = \min_{p\in\mathcal{F}_M} \mathrm{KL}(q \,\|\, p)$. *If $M \geq 2$ and $\mathrm{TC}(q) > 0$, then $\mathcal{D}_M(q) < \mathrm{TC}(q)$, so $\min_{p\in\mathcal{F}_M} \mathcal{L}_{\mathrm{exact}}(p) < \min_{p\in\mathcal{F}_M} \mathcal{L}_{\mathrm{surr}}(p)$ and every minimizer $p^\star$ of $\mathcal{L}_{\mathrm{exact}}$ has $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) > 0$.*

The proof is in App. B.4. Even with unlimited data and a perfect optimizer, the bound thus returns the factorized student of Theorem 1 for any M and any weights, because it scores a mixture by the average score of its components, whereas the exact objective rewards a mixture as soon as the teacher's step contains any dependence.

Relation to Di4C. Di4C and MLD3 use the same kind of student, a mixture of factorized kernels, but with different formulations. A student learns dependence between positions only if a loss term presents a correlated target and averages over the latent in a way that lets the student use it. In Di4C, the terms tied to the teacher compare the student with the teacher's one-step prediction, which is factorized, so they do not teach sequential dependence (Proposition 5). The terms with correlated targets use either the student's own prediction after one teacher sub-step or the data, and they estimate the average over the latent from R samples, which penalizes an informative latent for every finite R (Proposition 3). In MLD3, the target is the teacher's correlated transition over the whole step, and the finite latent makes the average exact. Theorem 2 shows why both are needed since the convexity bound would discard the dependence of this target.

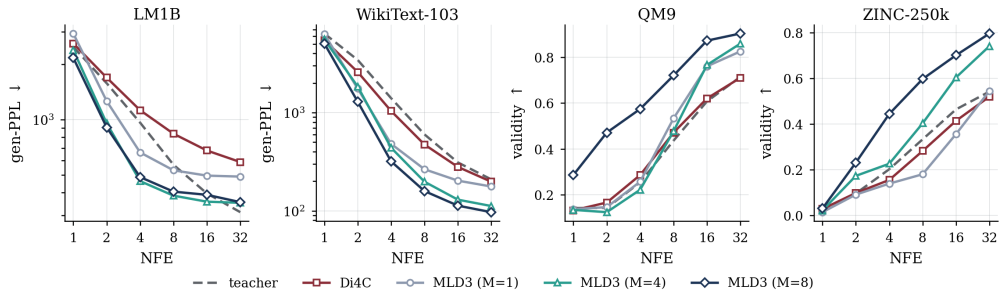

Figure 4: Few-step generation on the text (↓ gen-PPL) and molecule (↑ validity) datasets for the MDLM teacher, Di4C, and MLD3 students with $M \in \{1, 4, 8\}$.

4 Experiments

4.1 Setup

Datasets. We evaluated MLD3 for unconditional generation on 6 datasets across 4 domains, namely LM1B (Chelba et al., 2013) and WikiText-103 (Merity et al., 2016) for text, QM9 (Ramakrishnan et al., 2014) and ZINC-250k (Irwin et al., 2012) for molecules, the Drosophila STARR-seq enhancers (DeepSTARR) of De Almeida et al. (2022) for regulatory DNA, and ImageNet (Chang et al., 2022) for images. App. C and D further present dataset statistics and the experimental details.

Model training and baselines. We trained an MDLM teacher (Sahoo et al., 2024) on every dataset except ImageNet, for which we used the pretrained MaskGIT model as the teacher. The text teachers were trained for 1M steps on WikiText-103 and 200k steps on LM1B. We trained MLD3 students with $M \in \{1, 4, 8\}$ components and $K = 4$ steps. The student model with $M = 1$ serves as the *control*, since it shares the objective, the teacher, and the number of training steps with the mixtures. Our primary baseline is Di4C (Hayakawa et al., 2024), which we distilled from the same teacher as our students on these datasets. We retrained Di4C from random initialization to match the training setup of our students, whereas Di4C warm-starts its student from the teacher. We sampled from MLD3 students with the spread consensus, from the factorized models with the analytic reverse step of Eq. 3, and from Di4C with its ancestral sampler.

Metrics. For text, we report the generative perplexity (gen-PPL) under GPT-2 Large (Radford et al., 2019), the unigram entropy H_1 of the samples, and the transition information $\widehat{I}$, which estimates the mutual information between the component index and the step and which a finite mixture provides in closed form. For molecules, we report the fraction of valid molecules, known as validity (Park et al., 2025). App. C defines these metrics and their estimators.

4.2 Language modeling

As shown in Figure 4, the gain grows with M on both text datasets. On WikiText-103, the $M = 8$ student model improves on the $M = 1$ control and the $M = 4$ student model at every NFE. The $M = 8$ student model reduces gen-PPL from 265 to 159 at 8 NFEs, a 40% reduction against the control, and it improves on the teacher at all NFEs. Similarly, the $M = 8$ student model has 0.74 nats of transition information, while the factorized control has none. It also outperforms Di4C at all NFEs, by a factor of 3 at 8 NFEs and 2.0 at 32. On LM1B, the $M = 8$ student model reduces gen-PPL by 23% from 529 to 405 at 8 NFEs, and improves on the teacher at every NFE up to 16.

4.3 Molecule generation

Molecular strings couple their positions through hard chemical constraints. For example, a ring that opens but does not close invalidates the whole string, which makes validity a direct measure of whether the dependence within a step is respected. As shown in Figure 4, the gains of the mixtures are largest in the regime of very few steps and persist at all step counts. On QM9, our $M = 8$ student model reaches 0.471 validity at 2 NFEs against 0.166 for Di4C and 0.146 for the control, factors of 2.8 and 3.2, and it stays ahead of both at every NFE, at 0.903 against 0.710 and 0.825 at 32. On ZINC-250k, the same student model reaches 0.231 at 2 NFEs against 0.098 for Di4C, and 0.796 against 0.520 at 32. The $M = 4$ student model trails the control on QM9 up to 8 NFEs, and it leads the control at all NFEs on ZINC-250k.

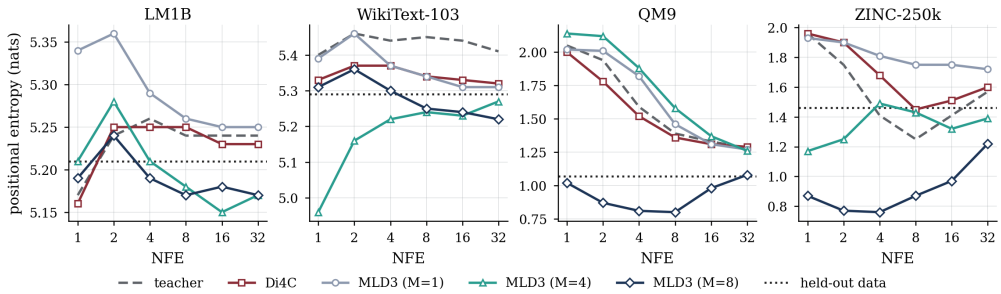

Figure 5: Entropy of the per-position marginal over a batch of samples across NFEs on the LM1B, WikiText-103, QM9, and ZINC-250k datasets. The dotted line marks held-out data for reference.

4.4 Additional results

Sample diversity. A student model can improve a sample-quality metric by concentrating its output distribution instead of modeling the data better. We therefore measure the entropy of the per-position marginal over a batch of samples and compare it with held-out data, as shown in Figure 5. On WikiText-103 and QM9, the factorized models have a higher entropy than the data at every NFE. Averaged over 4 to 8 NFEs, our $M = 8$ student model differs from the data by 0.03 nats on WikiText-103 and 0.26 nats on QM9, against 0.06 to 0.55 nats for the teacher and the control.

Regulatory DNA and image generation. We also evaluated MLD3 on the regulatory DNA (DeepSTARR) and ImageNet-256 benchmarks (details in App. D). On DeepSTARR, the $M = 4$ mixture improves the Fréchet biological distance (FBD) over its control by up to 21%, a smaller gain than on text or molecules, because the teacher's distilled steps contain almost no dependence between positions (App. E.3). The Di4C student model on this dataset also ignores its latent entirely, which we verify directly by sweeping the latent of the trained model (App. D.1). On ImageNet-256, where the students distill a MaskGIT teacher, our $M = 8$ student model improves the Fréchet Inception distance (FID) over Di4C at every NFE (App. D.2).

The overall gains of MLD3 are large on text and molecules and small on regulatory DNA and image tokens, and we quantify this difference via the teacher's predictions. In App. E.3, we measure the fraction of the entropy of a teacher step that is shared among the positions it reveals together. With 4 steps, this fraction is 10 to 41% on text and molecules and 1 to 4% on DNA and images, and the gain of the mixture over its control follows the same ranking (Figure 10). This measurement serves as a quick metric to run on the teacher before training an MLD3 student. Hence, a dataset whose teacher steps contain only a few percent of dependence is already well served by a factorized denoiser.

Ablations. App. E reports the ablations that isolate the contributions of individual design choices in MLD3. The sweep over the number of student model steps K shows that the number of steps a student is trained for fixes its quality at the smallest NFEs and matters little once the sampler is given more steps. Moreover, we decoded the MLD3 students with our spread consensus at $m = 2$, which costs 12 transformer-block passes per step at $M = 1$ and 16 at $M > 1$, where one block pass is a single evaluation of one transformer block.

Analysis of the learned mixture. We trained the same MLD3 trunk with the Di4C objective in place of ours, with the same teacher, $M = 8$, and number of training steps (App. D.4). Under our objective, the components of the QM9 student are 0.53 nats apart in pairwise KL divergence, and the router moves its weight between them along the trajectory (Figure 7). Under the Di4C objective, the components coincide to within 0.002 nats on all 3 datasets, and the router stays uniform, since it's not trainable under its objective. A denoising step therefore carries at most 0.004 nats of dependence between positions, compared to 1.2 to 11.2 nats under our approach.

5 Conclusion

We presented MLD3, a novel distillation approach for discrete state space models in which each denoising step of the student model is a finite mixture of factorized kernels. We first identified a failure mode in the convexity bound that prior work trains on, and then proposed a finite latent in MLD3. Our approach makes the step likelihood a closed-form solution with learnable mixing weights as routers. The extensive experiments on unconditional text and molecule generation show large few-step gains.

Reproducibility Statement

We will make the code publicly available on GitHub and provide installation scripts with documentation to address any dependency issues. Furthermore, all the datasets used in this paper are publicly available, with the experimental details provided in the Appendix. We hope that this will support and accelerate future research and development.

Ethics Statement

This work adheres to the ICLR Code of Ethics. Our research focuses on computational methods for language and biological sequence design. We acknowledge that biological sequence design technologies could have dual-use implications. We have designed our experiments to focus on clearly beneficial applications and have not explored potentially harmful uses. All datasets used in this work are publicly available and have been previously published with appropriate ethical considerations. No human subjects were involved in this research. The computational methods developed here are general frameworks that require domain expertise and appropriate oversight for real-world applications. We encourage responsible use of these methods and recommend that practitioners consider ethical implications and potential risks when applying sequence design technologies in practice.

AI use statement

Generative AI tools were used for manuscript language editing and for assistance with code drafting and debugging. All reported numerical results were produced by the authors' computational pipelines and checked against saved outputs. The authors reviewed and revised all AI-assisted text and code and take responsibility for the final manuscript, analyses, and artifacts.

References

Mansoor Ahmed, Yue-Tsz Fan, Hemanth Venkateswara, and Murray Patterson. Latent-kernel discrete flow maps for few-step generation. *arXiv preprint arXiv:2607.27529*, 2026.

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pp. 50726–50753, 2025.

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

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models. 2022.

Nicholas Boffi, Michael Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *Advances in Neural Information Processing Systems*, 38: 33346–33382, 2026.

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv preprint arXiv:2406.07507*, 2024.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pp. 10–21, 2016.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. 2015.

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

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

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

Chen-Hao Lance Chao, Wei-Fang Sun, Hanwen Liang, Chun-Yi Lee, and Rahul Krishnan. Beyond masked and unmasked: Discrete diffusion models via partial masking. *Advances in Neural Information Processing Systems*, 38:97458–97500, 2026.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

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

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Bernardo P De Almeida, Franziska Reiter, Michaela Pagani, and Alexander Stark. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature genetics*, 54(5):613–624, 2022.

Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast LLMs via self-distillation through time. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.21035.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Daniil Dmitriev, Zhihan Huang, and Yuting Wei. Efficient sampling with discrete diffusion models: Sharp and adaptive guarantees. *arXiv preprint arXiv:2602.15008*, 2026.

Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. *arXiv preprint arXiv:2502.02150*, 2025.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. Distillation of discrete diffusion through dimensional correlations. *arXiv preprint arXiv:2410.08709*, 2024.

Peter Holderrieth, Douglas Chen, Luca Eyring, Ishin Shah, Giri Anantharaman, Yutong He, Zeynep Akata, Tommi Jaakkola, Nicholas Matthew Boffi, and Max Simchowitz. Diamond maps: Efficient reward alignment via stochastic flow maps. *arXiv preprint arXiv:2602.05993*, 2026.

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

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52 (7):1757–1768, 2012.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *International Conference on Learning Representations*, volume 2024, pp. 44493–44525, 2024.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

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. *International Conference on Learning Representations*, 2025.

Mingue Park, Jisung Hwang, Seungwoo Yoo, Kyeongmin Yeo, and Minhyuk Sung. Pairflow: Closed-form source-target coupling for few-step generation in discrete flow models. *arXiv preprint arXiv:2512.20063*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023. arXiv:2212.09748.

Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S Albergo. Discrete flow maps. *arXiv preprint arXiv:2604.09784*, 2026.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):140022, 2014.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, 2021.

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

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *arXiv preprint arXiv:2506.10892*, 2025.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*, 2024.

Zhengyan Wan, Yidong Ouyang, Liyan Xie, Fang Fang, Hongyuan Zha, and Guang Cheng. Discrete guidance matching: Exact guidance for discrete flow matching. *arXiv preprint arXiv:2509.21912*, 2025.

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

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. Energy-based diffusion language models for text generation. In *International Conference on Learning Representations*, volume 2025, pp. 33769–33789, 2025.

Jaehoon Yoo, Wonjung Kim, and Seunghoon Hong. Redi: Rectified discrete flow. *Advances in Neural Information Processing Systems*, 38:81651–81683, 2026.

Yunxiao Zhao and Changxiao Cai. Adaptation to intrinsic dependence in diffusion language models. *arXiv preprint arXiv:2602.20126*, 2026.

APPENDIX

A Related work **15**

B Proofs **16**

B.1 Notation, conventions, and auxiliary lemmas 16

B.2 Factorization error 18

B.3 The per-step objective and the final sample 19

B.4 Surrogate collapse 20

B.5 Information identity 27

B.6 Capacity of a mixture and the number of components 28

C Experimental details **29**

C.1 Text 29

C.2 Molecules 29

C.3 Metrics and estimators 29

C.4 Implementation details 30

C.5 Detailed molecule results 32

D Additional experiments **32**

D.1 Regulatory DNA 32

D.2 Image generation 33

D.3 Sample diversity 34

D.4 Quantitative analysis of the learned mixture components 34

E Ablations **37**

E.1 Number of student steps 37

E.2 Alternative sampling approaches 38

E.3 When to distill discrete diffusion models with MLD3. 39

F Example generations **41**

F.1 Text 41

F.2 Molecules 45

F.3 Regulatory DNA 45

F.4 Images 47

A Related work

Few-step discrete generation. Discrete diffusion and flow models corrupt a sequence toward a simple source and learn to reverse the corruption in small steps (Austin et al., 2021; Campbell et al., 2022; Lou et al., 2023; Campbell et al., 2024; Gat et al., 2024), with masked diffusion with a simplified objective as the standard instance (Sahoo et al., 2024; Shi et al., 2024; Ou et al., 2025). Later work extends the probability paths and the guidance of these models and analyzes the error of coarse sampling (Sahoo et al., 2025; Stark et al., 2024; Wan et al., 2025; Feng et al., 2025; Chao et al., 2026; Dmitriev et al., 2026; Zhao & Cai, 2026). Few-step generation is commonly achieved in three ways. *Distillation* compresses the sampling chain of a teacher into a student, either through the teacher's own multi-step targets (Deschenaux & Gulcehre, 2025) or through a consistency objective that regresses the student onto itself at a nearby time (Sahoo et al., 2025; Hayakawa et al., 2024). *Rectification* keeps the student factorized and changes the coupling between source and target so that fewer steps suffice, iteratively (Yoo et al., 2026) or in closed form (Park et al., 2025), following rectified flow in continuous state (Liu et al., 2022). A third line leaves the discrete state space and learns a *flow map or consistency model on a continuous relaxation* of the tokens (Potaptchik et al., 2026; Lee et al., 2026), transferring the objectives of continuous-state flow maps (Song & Dhariwal, 2024; Kim et al., 2024; Boffi et al., 2024; 2026). For instance, stochastic flow maps sample the clean state instead of predicting its mean (Holderrieth et al., 2026). Potaptchik et al. (2026) and Lee et al. (2026) are also motivated by the limitation that we study in this paper, namely that a distilled discrete step updates each token independently. These approaches do not use the discrete transition kernel, while we keep the kernel and change its form. This makes the student compatible with pretrained masked diffusion models and the step likelihood exact.

Dependence within a step. Three main approaches aim to capture the dependence between the positions of one step. Di4C (Hayakawa et al., 2024) conditions a masked diffusion student on a continuous latent shared across positions, so that the step is a mixture of product kernels. It shows that such mixtures represent any distribution, once a sequence of positive probability receives its own component. Its objective anchors the student to the teacher through the convexity bound of Eq. 1 and a per-position marginal loss. The intended dependence enters through a consistency term, whose average over the latent is estimated from samples with control variates. Discrete copula diffusion (Liu et al., 2025) and energy-based diffusion language models (Xu et al., 2025) instead correct the independent sampler at inference with an auxiliary autoregressive or energy model, and block diffusion (Arriola et al., 2025) generates blocks of tokens autoregressively. MLD3 shares the mixture-of-products kernel idea with Di4C and differs in the latent, target, and objective. The latent is reformulated as a finite mixture with a learned state-dependent router. The target is the teacher's transition over the whole step instead of the student's own posterior at an adjacent time, while the average over the latent is a closed-form sum rather than a bound or an estimate. A single network produces the mixture components, where sampling does not need an auxiliary model. In non-autoregressive translation, the multimodality problem (Gu et al., 2017) was traced to the conditional total correlation that is dropped in a conditionally independent model (Huang et al., 2022). Moreover, the router in MLD3 is analogous to the gate in a mixture-of-experts layer (Shazeer et al., 2017; Riquelme et al., 2021; Jiang et al., 2024). It differs in that it selects one component for the whole sequence to couple the positions, and its weights enter the objective in closed form rather than through sparse routing.

Objectives. The factorization error has several formal treatments. Theorem 1 states the error of a single step as the information projection onto product distributions (Csiszár, 1975; Watanabe, 1960). Hayakawa et al. (2024) prove that product models need a number of steps inversely proportional to the target error. Yoo et al. (2026) show that one rectification step reduces the conditional total correlation. On the objective side, the convexity bound of Eq. 1 is the evidence lower bound of a latent-variable model with its variational posterior frozen at the prior. Its minimizers therefore ignore the latent, similar to posterior collapse in variational autoencoders (Bowman et al., 2016). Burda et al. (2015) showed that a single-sample bound under-uses the latent and tightened it with more samples. We show in MLD3 that a finite latent prevents the estimate altogether, similar to how Bao et al. (2022) replaced an approximate reverse variance by its closed form. Consistency objectives (Song & Dhariwal, 2024; Kim et al., 2024) regress a student onto itself at another time and need a boundary term at the clean end to anchor the induction, and Di4C follows this design in discrete state. However, MLD3 regresses onto the multiple evaluations of a frozen teacher over an interval and does not need an anchor.

B PROOFS

B.1 NOTATION, CONVENTIONS, AND AUXILIARY LEMMAS

Throughout this appendix, the vocabulary $\mathcal{S}$ is finite, sequences have L positions, and the state space is $\mathcal{X} = (\mathcal{S} \cup \{\texttt{[MASK]}\})^L$. We report all entropies, divergences, and mutual information in nats. Table 1 summarizes the notation used in the paper.

Table 1: Notation used in the paper.

symbol	meaning
$\mathcal{S}, V$	token vocabulary and its size, excluding [MASK]
[MASK]	absorbing mask symbol, index V
L	sequence length
$\mathcal{X}$	state space $(\mathcal{S} \cup \{\texttt{[MASK]}\})^L$
$\boldsymbol{x} = (x^1, \dots, x^L)$	a sequence, with x^i its i-th token
$t < s$	start and end times of a step, with $t = 0$ fully masked and $t = 1$ clean
κ_t	probability that a position is unmasked at time t
K	number of student sampling steps. The step boundaries are $t_j = \frac{j}{K}$ for $j = 0, \dots, K$, and a training example on the sampling schedule uses one interval $[t_j, t_{j+1}]$
T	number of analytic teacher sub-steps used to construct one distillation target from t to s. The sweep over K keeps the product KT fixed at 16
NFE	number of function evaluations, i.e., network calls, used by a sampler
$p^{\psi}_{s\|t}(\cdot \mid \boldsymbol{x}_t)$	teacher transition from t to s, written q for short
$p^{\theta}_{s\|t}(\cdot \mid \boldsymbol{x}_t)$	student transition from t to s
q_i	i-th position marginal of q
$\Pi(q)$	product of the marginals of q, $\prod_i q_i$
M	number of mixture components
$k \in [M]$	component index
$w^{\theta}_k(\boldsymbol{x}_t, t)$	router weight of component k
$\mathcal{E}$, $\tilde{w}_k$	components a step evaluates and the router weights renormalized over them
$P^{\theta}_{i,k}(\cdot \mid \boldsymbol{x}_t, t)$	position-i factor of component k
$\boldsymbol{X}_s, \boldsymbol{x}_s$	random sequence at the end of a step and its realization
$\bar{p}_i$, $\bar{P}_i$	position-i marginal and denoiser averaged over components with the router weights
Z_i, $\tilde{P}_i$	probability of a clean token, $1 - \bar{P}_i(\texttt{[MASK]})$, and the denoiser renormalized over the clean tokens, $\tilde{P}_i = \frac{\bar{P}_i}{Z_i}$
$\mathcal{M}(\boldsymbol{x}_t)$	the positions still masked at $\boldsymbol{x}_t$
$\ell_k(\boldsymbol{x}_s)$	component log likelihood $\sum_{i \in \mathcal{M}(\boldsymbol{x}_t)} \log P^{\theta}_{i,k}(x^i_s \mid \boldsymbol{x}_t, t)$
$\tilde{\ell}_k(\boldsymbol{x}_s)$	log-likelihood of the tokens revealed by a step under component k, renormalized over the clean tokens
$\mathrm{sg}[\cdot]$	stop-gradient operator, which blocks the gradient from flowing through its argument
$\mathcal{F}_M$	family of mixture kernels with at most M components
$\mathcal{D}_M(q)$	smallest divergence from q to $\mathcal{F}_M$, $\inf_{p \in \mathcal{F}_M} \mathrm{KL}(q \,\|\, p)$
$\delta_{\boldsymbol{x}}$	point mass on the sequence $\boldsymbol{x}$
$\mathrm{TC}(q)$	total correlation of a transition
$\mathrm{TC}(\boldsymbol{X} \mid Z)$	conditional total correlation given a latent Z
$\mathcal{L}_{\mathrm{exact}}(p)$, $\mathcal{L}_{\mathrm{surr}}(p)$	exact objective and latent-separable bound at a fixed state
$p(k \mid \boldsymbol{x}_s)$	posterior over the component given the step
μ_j, ν_j	distributions of the state at time t_j under the teacher and under the student
$\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t)$, $\widehat{I}$	transition information and its estimate
ρ_K, κ_K	relative dependence and relative reverse gap of the K-step targets of a teacher
$\mathcal{B}_{\tilde{P}}$, $\omega(t)$	the masked-diffusion evidence lower bound under $\tilde{P}$ and its time weight
D, L_{lat}	trunk depth and number of blocks evaluated once per component
$C(M)$	number of transformer-block passes per forward pass over each component, $C(M) = (D - L_{\mathrm{lat}}) + ML_{\mathrm{lat}}$
H, KL, I	entropy, Kullback-Leibler divergence, mutual information

We fix the corrupted state $\boldsymbol{x}_t$, and each statement below holds pointwise in $\boldsymbol{x}_t$ unless written as an outer expectation explicitly. We write $q = p^{\psi}_{s|t}(\cdot \mid \boldsymbol{x}_t)$ for the teacher's transition from $\boldsymbol{x}_t$. The mixture family with M components is the set of kernels

$$\mathcal{F}_M \;=\; \Big\{ p \,:\, p(\boldsymbol{x}_s) = \sum_{k=1}^{M} w_k \prod_{i=1}^{L} P_i^k(x_s^i), \quad w \in \Delta^{M-1}, \quad P_i^k \in \Delta(\mathcal{S} \cup \{\texttt{[MASK]}\}) \Big\}, \tag{13}$$

where Δ^{M-1} is the probability simplex over M indices and $\Delta(\mathcal{S} \cup \{\texttt{[MASK]}\})$ is the set of categorical distributions over the extended vocabulary. We assume one property of the kernels and fix two conventions. First, the *carry-over* constraint of Eq. 2 holds for the teacher and for every component. At a position i that $\boldsymbol{x}_t$ already reveals, each P_i^k and each q_i is therefore the point mass $\delta_{x_t^i}$, because masking is absorbing. Second, we use the conventions $0 \log 0 = 0$ and $\mathrm{KL}(\mu \,\|\, \nu) = +\infty$ whenever $\nu(x) = 0$ for some x with $\mu(x) > 0$, so that the identities below remain valid when a component assigns zero mass to a token the teacher can produce, and every identity between divergences is understood in $[0, +\infty]$. We write $\mathrm{supp}(\mu) = \{x : \mu(x) > 0\}$ for the support of a distribution μ.

Definition 1 (Total correlation of a transition)**.** *For a distribution q on $\mathcal{X}$ with marginals q_i,*

$$\mathrm{TC}(q) \;=\; \mathrm{KL}\Big(q \,\Big\|\, \textstyle\prod_{i=1}^{L} q_i\Big) \;=\; \textstyle\sum_{i=1}^{L} \mathrm{H}(q_i) \;-\; \mathrm{H}(q) \;\geq\; 0, \tag{14}$$

and it vanishes if and only if q factorizes across positions.

The second equality in Eq. 14 follows by expanding the divergence and marginalizing each summand,

$$\mathrm{KL}\Big(q \,\Big\|\, \prod_{i} q_i\Big) \;=\; \mathbb{E}_q\big[\log q(\boldsymbol{X})\big] - \sum_{i=1}^{L} \mathbb{E}_q\big[\log q_i(X^i)\big] \;=\; -\mathrm{H}(q) + \sum_{i=1}^{L} \mathrm{H}(q_i), \tag{15}$$

where the last step uses $\mathbb{E}_q[\log q_i(X^i)] = \mathbb{E}_{q_i}[\log q_i] = -\mathrm{H}(q_i)$, since the i-th summand depends on the sequence only through position i. The non-negativity and the equality condition follow from applying Gibbs' inequality, which we state as Lemma 1 below, to the pair $(q, \Pi(q))$.

Lemma 1 (Gibbs' inequality)**.** *For distributions μ, ν on a finite set, $\mathrm{KL}(\mu \,\|\, \nu) \geq 0$, with equality if and only if $\mu = \nu$.*

Proof. If $\nu(x) = 0$ for some x with $\mu(x) > 0$, the divergence is $+\infty$ by convention, and the claim holds trivially. We may therefore assume that $\nu > 0$ on $\mathrm{supp}(\mu)$. Since $\log$ is strictly concave, Jensen's inequality yields $-\mathrm{KL}(\mu \,\|\, \nu) = \mathbb{E}_\mu\big[\log \frac{\nu}{\mu}\big] \leq \log \mathbb{E}_\mu\big[\frac{\nu}{\mu}\big] = \log \sum_{x \in \mathrm{supp}(\mu)} \nu(x) \leq \log 1 = 0$. Equality in the Jensen step requires $\frac{\nu}{\mu}$ to be constant on $\mathrm{supp}(\mu)$, and equality in the last step requires ν to place all its probability on $\mathrm{supp}(\mu)$. Together, the two conditions imply $\mu = \nu$. □

Lemma 2 (Conditional additivity of entropy)**.** *Let $(k, \boldsymbol{X})$ be a pair of random variables such that the coordinates $X^1, \ldots, X^L$ of $\boldsymbol{X}$ are conditionally independent given k. Then*

$$\mathrm{H}(\boldsymbol{X} \mid k) = \sum_{i=1}^{L} \mathrm{H}(X^i \mid k).$$

Proof. By the chain rule of entropy, $\mathrm{H}(\boldsymbol{X} \mid k) = \sum_{i=1}^{L} \mathrm{H}(X^i \mid X^1, \ldots, X^{i-1}, k)$. Given k, the coordinates are independent. Hence, we may drop the conditioning on the earlier coordinates, and each summand becomes $\mathrm{H}(X^i \mid k)$. □

Lemma 3 (Data processing)**.** *Let k and $\boldsymbol{X}$ be jointly distributed random variables, and let f be a deterministic function of $\boldsymbol{X}$. Then*

$$\mathrm{I}(k; f(\boldsymbol{X})) \leq \mathrm{I}(k; \boldsymbol{X}).$$

Proof. We expand the mutual information between k and the pair $(\boldsymbol{X}, f(\boldsymbol{X}))$ with the chain rule in two orders. On the one hand, $\mathrm{I}(k; \boldsymbol{X}, f(\boldsymbol{X})) = \mathrm{I}(k; \boldsymbol{X})$, because $f(\boldsymbol{X})$ is determined by

$\boldsymbol{X}$. On the other hand, $\mathrm{I}(k; \boldsymbol{X}, f(\boldsymbol{X})) = \mathrm{I}(k; f(\boldsymbol{X})) + \mathrm{I}(k; \boldsymbol{X} \mid f(\boldsymbol{X})) \geq \mathrm{I}(k; f(\boldsymbol{X}))$, because conditional mutual information is non-negative. Comparing the two expansions establishes $\mathrm{I}(k; \boldsymbol{X}) \geq \mathrm{I}(k; f(\boldsymbol{X}))$. □

Lemma 4 (Chain rule of the divergence). *Let μ and ν be distributions of a pair (Y_1, Y_2) on a finite set, with marginals μ_1 and ν_1 of Y_1 and conditionals $\mu_{2|1}(\cdot \mid y_1)$ and $\nu_{2|1}(\cdot \mid y_1)$ of Y_2 given $Y_1 = y_1$. Then*

$$\mathrm{KL}(\mu \,\|\, \nu) = \mathrm{KL}(\mu_1 \,\|\, \nu_1) + \mathbb{E}_{y_1 \sim \mu_1}\Big[\mathrm{KL}\big(\mu_{2|1}(\cdot \mid y_1) \,\big\|\, \nu_{2|1}(\cdot \mid y_1)\big)\Big],$$

and in particular $\mathrm{KL}(\mu_1 \,\|\, \nu_1) \leq \mathrm{KL}(\mu \,\|\, \nu)$.

Proof. If $\nu(y_1, y_2) = 0 < \mu(y_1, y_2)$ for some pair, then either $\nu_1(y_1) = 0 < \mu_1(y_1)$, or $\mu_1(y_1) > 0$ and $\nu_{2|1}(y_2 \mid y_1) = 0 < \mu_{2|1}(y_2 \mid y_1)$, so both sides equal $+\infty$. Otherwise, $\mu(y_1, y_2) = \mu_1(y_1)\,\mu_{2|1}(y_2 \mid y_1)$ and $\nu(y_1, y_2) = \nu_1(y_1)\,\nu_{2|1}(y_2 \mid y_1)$ are positive on the support of μ, so

$$\mathrm{KL}(\mu \,\|\, \nu) = \sum_{y_1, y_2} \mu(y_1, y_2) \log \frac{\mu_1(y_1)}{\nu_1(y_1)} + \sum_{y_1, y_2} \mu(y_1, y_2) \log \frac{\mu_{2|1}(y_2 \mid y_1)}{\nu_{2|1}(y_2 \mid y_1)}.$$

Summing the first term over y_2 gives $\mathrm{KL}(\mu_1 \,\|\, \nu_1)$, and writing $\mu(y_1, y_2) = \mu_1(y_1)\,\mu_{2|1}(y_2 \mid y_1)$ in the second term and summing over y_2 first gives the expected conditional divergence. The inequality follows because the second term is non-negative by Lemma 1. □

Lemma 5 (Joint convexity of the divergence). *Let π be a distribution on a finite index set $\mathcal{Z}$, and for every $z \in \mathcal{Z}$ let μ_z and ν_z be distributions on the same finite set. Then*

$$\mathrm{KL}\Big(\sum_z \pi_z\, \mu_z \,\Big\|\, \sum_z \pi_z\, \nu_z\Big) \leq \sum_z \pi_z\, \mathrm{KL}(\mu_z \,\|\, \nu_z).$$

Proof. For non-negative numbers a_z and b_z, the log-sum inequality (Cover et al., 1991, Thm. 2.7.1) states that $\sum_z a_z \log \frac{a_z}{b_z} \geq \big(\sum_z a_z\big) \log \frac{\sum_z a_z}{\sum_z b_z}$, under the conventions of App. B.1. Applying it at every point x with $a_z = \pi_z\, \mu_z(x)$ and $b_z = \pi_z\, \nu_z(x)$, and summing over x, provides the claim. □

B.2 Factorization error

Theorem 1 (factorization error). *Let q be any distribution on $\mathcal{X}$ with marginals q_i. For every factorized kernel $p = \prod_i p_i \in \mathcal{F}_1$, we have the decomposition*

$$\mathrm{KL}(q \,\|\, p) = \mathrm{TC}(q) + \sum_{i=1}^{L} \mathrm{KL}(q_i \,\|\, p_i), \qquad \textit{hence} \qquad \min_{p \in \mathcal{F}_1} \mathrm{KL}(q \,\|\, p) = \mathrm{TC}(q),$$

with the minimum attained if and only if $p = \Pi(q)$. For a fixed target transition q, this minimum depends only on the teacher and on the step size.

Proof. We first treat the case in which both sides of the decomposition are infinite, then establish the decomposition by inserting the product of the marginals of q inside the logarithm, and finally derive the minimizer from that decomposition.

Step 1 (the infinite case). Suppose that $p_i(v) = 0$ for some position i and some token v with $q_i(v) > 0$. The right-hand side is then $+\infty$, because $\mathrm{KL}(q_i \,\|\, p_i) = +\infty$ by the convention of App. B.1, and the remaining summands are non-negative. The left-hand side is also $+\infty$, because $q_i(v) > 0$ implies that some $\boldsymbol{x}$ with $q(\boldsymbol{x}) > 0$ has $x^i = v$, and at this $\boldsymbol{x}$, the product $p(\boldsymbol{x}) = \prod_j p_j(x^j)$ contains the factor $p_i(v) = 0$. Hence, $p(\boldsymbol{x}) = 0$ while $q(\boldsymbol{x}) > 0$, and $\mathrm{KL}(q \,\|\, p) = +\infty$ by the same convention. Both sides thus agree. We may therefore assume from now on that each p_i is strictly positive on $\mathrm{supp}(q_i)$. Then p is strictly positive on $\mathrm{supp}(q)$, and each sum below is finite.

Step 2 (the decomposition). Since $q(\boldsymbol{x}) > 0$ implies $q_i(x^i) > 0$ for all i, we may multiply and divide by $\prod_i q_i(x^i)$ inside the logarithm,

$$\begin{aligned}
\mathrm{KL}(q \,\|\, p) &= \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{\prod_i p_i(x^i)} && \text{(definition of the divergence)} \\
&= \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log\Big(\frac{q(\boldsymbol{x})}{\prod_i q_i(x^i)} \cdot \frac{\prod_i q_i(x^i)}{\prod_i p_i(x^i)} \Big) && \text{(insert } \Pi(q)\text{, which is positive on } \mathrm{supp}(q)) \\
&= \mathrm{TC}(q) \;+\; \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \sum_{i=1}^{L} \log \frac{q_i(x^i)}{p_i(x^i)}. && \text{(Definition 1; logarithm of a product)}
\end{aligned} \tag{16}$$

We may split the sum, because Step 1 makes both resulting sums finite. The i-th inner summand depends on $\boldsymbol{x}$ only through position i. Hence, we group the outer sum by the value v at that position and obtain

$$\sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q_i(x^i)}{p_i(x^i)} \;=\; \sum_{v} \Big(\sum_{\boldsymbol{x}\,:\,x^i = v} q(\boldsymbol{x}) \Big) \log \frac{q_i(v)}{p_i(v)} \;=\; \sum_{v} q_i(v) \log \frac{q_i(v)}{p_i(v)} \;=\; \mathrm{KL}(q_i \,\|\, p_i), \tag{17}$$

since the inner sum over all $\boldsymbol{x}$ with $x^i = v$ equals the marginal probability $q_i(v)$. Substituting Eq. 17 into Eq. 16 for all i produces the stated identity.

Step 3 (the minimization). The first term of the identity does not depend on p. By Lemma 1, each summand $\mathrm{KL}(q_i \,\|\, p_i)$ is non-negative and vanishes if and only if $p_i = q_i$. Within $\mathcal{F}_1$, we may choose the factor at each position independently of the others. The sum thus attains its minimum value of zero if and only if $p_i = q_i$ for all i. The minimizer $\Pi(q)$ is itself a member of $\mathcal{F}_1$. Hence, the infimum is attained and equals $\mathrm{TC}(q)$. Finally, $\mathrm{TC}(q)$ is a function of q alone, and q is determined by the teacher and the pair of times (t, s). Hence, the minimum does not depend on any property of the student. □

B.3 The per-step objective and the final sample

The step-KL objective of Eq. 4 is defined one step at a time. The following proposition shows that the sum of the per-step divergences along the sampling schedule bounds the divergence between the teacher's and the student's distributions of the final sequence.

Proposition 1 (Per-step divergences bound the final divergence)**.** *Let $0 = t_0 < t_1 < \cdots < t_K = 1$ be the boundaries of the sampling schedule, and for $j = 1, \ldots, K$ let $q_j(\cdot \mid \boldsymbol{x}) = p^{\psi}_{t_j|t_{j-1}}(\cdot \mid \boldsymbol{x})$ and $p_j(\cdot \mid \boldsymbol{x}) = p^{\theta}_{t_j|t_{j-1}}(\cdot \mid \boldsymbol{x})$ be the teacher's and the student's transitions over the j-th interval. Let both chains start from the same distribution μ_0 of the initial state, and let μ_j and ν_j be the distributions of the state at time t_j under the teacher and under the student. Then*

$$\mathrm{KL}(\mu_K \,\|\, \nu_K) \;\le\; \sum_{j=1}^{K} \mathbb{E}_{\boldsymbol{x}\sim\mu_{j-1}} \Big[\mathrm{KL}\big(q_j(\cdot \mid \boldsymbol{x}) \,\big\|\, p_j(\cdot \mid \boldsymbol{x}) \big) \Big]. \tag{18}$$

Proof. Let $\mu^{(j)}$ and $\nu^{(j)}$ be the joint distributions of $(\boldsymbol{X}_{t_0}, \ldots, \boldsymbol{X}_{t_j})$ under the teacher and under the student. Both chains are Markov, so the conditional distribution of $\boldsymbol{X}_{t_j}$ given $(\boldsymbol{X}_{t_0}, \ldots, \boldsymbol{X}_{t_{j-1}})$ is $q_j(\cdot \mid \boldsymbol{x}_{t_{j-1}})$ under the teacher and $p_j(\cdot \mid \boldsymbol{x}_{t_{j-1}})$ under the student. Lemma 4, applied with $Y_1 = (\boldsymbol{X}_{t_0}, \ldots, \boldsymbol{X}_{t_{j-1}})$ and $Y_2 = \boldsymbol{X}_{t_j}$, gives

$$\mathrm{KL}\big(\mu^{(j)} \,\big\|\, \nu^{(j)}\big) \;=\; \mathrm{KL}\big(\mu^{(j-1)} \,\big\|\, \nu^{(j-1)}\big) \;+\; \mathbb{E}_{\boldsymbol{x}\sim\mu_{j-1}} \Big[\mathrm{KL}\big(q_j(\cdot \mid \boldsymbol{x}) \,\big\|\, p_j(\cdot \mid \boldsymbol{x}) \big) \Big],$$

where the expectation over the past reduces to an expectation over $\boldsymbol{x}_{t_{j-1}} \sim \mu_{j-1}$, because both conditional distributions depend on the past only through $\boldsymbol{x}_{t_{j-1}}$. Since $\mu^{(0)} = \nu^{(0)} = \mu_0$, induction over j gives $\mathrm{KL}(\mu^{(K)} \,\|\, \nu^{(K)}) = \sum_{j=1}^{K} \mathbb{E}_{\boldsymbol{x}\sim\mu_{j-1}} [\mathrm{KL}(q_j(\cdot \mid \boldsymbol{x}) \,\|\, p_j(\cdot \mid \boldsymbol{x}))]$. Finally, μ_K and ν_K are the marginals of $\mu^{(K)}$ and $\nu^{(K)}$ at time t_K, so $\mathrm{KL}(\mu_K \,\|\, \nu_K) \le \mathrm{KL}(\mu^{(K)} \,\|\, \nu^{(K)})$ by the inequality in Lemma 4. □

Remark 1 (Relation to the training objective). *The j-th term on the right of Eq. 18 is the step-KL objective of Eq. 4 for the j-th interval of the sampling schedule, evaluated at states drawn from the teacher's marginal μ_{j-1}, whereas the transition loss of Eq. 8 draws the states by corrupting data. The two coincide when the teacher's reverse process reproduces the forward marginals. The proposition holds for every student kernel, factorized or not, and Pinsker's inequality turns Eq. 18 into a bound on the total variation distance between the two distributions of the final sequence, which is the form of the guarantee that Hayakawa et al. (2024, Thm. 2) give for consistency-type losses.*

B.4 Surrogate collapse

Setting. Throughout this section, we fix the state $\boldsymbol{x}_t$ and the pair of times (t, s), and every distribution is a distribution over the tokens at the masked positions $\mathcal{M}(\boldsymbol{x}_t)$. Because masking is absorbing, a position that $\boldsymbol{x}_t$ already reveals keeps its token under the teacher and under every component. Such a position contributes a factor of 1 to every likelihood below and can be omitted. Hence, all summations over positions are over $\mathcal{M}(\boldsymbol{x}_t)$, and Definition 1 and Theorem 1 apply verbatim to distributions over $\mathcal{M}(\boldsymbol{x}_t)$. We write q for the teacher's transition over the masked positions and q_i for its marginal at position i. A kernel $p \in \mathcal{F}_M$ has weights $w \in \Delta^{M-1}$ and components $P^k = \prod_i P_i^k$, so that $p = \sum_k w_k P^k$ and $\ell_k(\boldsymbol{x}_s) = \log P^k(\boldsymbol{x}_s)$. At the fixed state, the two objectives compared in Section 3.3 are

$$\mathcal{L}_{\mathrm{exact}}(p) \;=\; \mathbb{E}_{\boldsymbol{x}_s \sim q}\big[-\log p(\boldsymbol{x}_s)\big] \;=\; \mathrm{H}(q) + \mathrm{KL}(q \,\|\, p), \tag{19}$$

$$\mathcal{L}_{\mathrm{surr}}(p) \;=\; \mathbb{E}_{\boldsymbol{x}_s \sim q}\, \mathbb{E}_{\hat{k} \sim w}\big[-\ell_{\hat{k}}(\boldsymbol{x}_s)\big] \;=\; \sum_{k=1}^{M} w_k\, \mathbb{E}_{\boldsymbol{x}_s \sim q}\big[-\log P^k(\boldsymbol{x}_s)\big], \tag{20}$$

where the second equality in Eq. 19 is Eq. 4, and the second equality in Eq. 20 holds because $\hat{k}$ is drawn independently of $\boldsymbol{x}_s$. Eqs. 8 and 10 are the expectations of these two quantities over the time pair and the state. All identities below hold in $[0, +\infty]$ under the conventions of App. B.1. In particular, both objectives equal $+\infty$ at a kernel that assigns zero probability to some $\boldsymbol{x}_s \in \mathrm{supp}(q)$, and such a kernel cannot be a minimizer. For every $\boldsymbol{x}_s$ with $p(\boldsymbol{x}_s) > 0$, the posterior over the component given the step is

$$p(k \mid \boldsymbol{x}_s) \;=\; \frac{w_k\, P^k(\boldsymbol{x}_s)}{p(\boldsymbol{x}_s)}, \qquad k \in [M]. \tag{21}$$

Lemma 6 (Variational identity). *Let $p \in \mathcal{F}_M$ have weights w and components $P^1, \dots, P^M$, and let $\boldsymbol{x}_s$ satisfy $p(\boldsymbol{x}_s) > 0$. Then*

$$\mathbb{E}_{\hat{k} \sim w}\big[-\ell_{\hat{k}}(\boldsymbol{x}_s)\big] \;=\; -\log p(\boldsymbol{x}_s) \;+\; \mathrm{KL}\big(w \,\big\|\, p(\cdot \mid \boldsymbol{x}_s)\big). \tag{22}$$

The divergence on the right is non-negative, and it vanishes if and only if $P^k(\boldsymbol{x}_s) = p(\boldsymbol{x}_s)$ for every k with $w_k > 0$.

Proof. If $P^k(\boldsymbol{x}_s) = 0$ for some k with $w_k > 0$, then $p(k \mid \boldsymbol{x}_s) = 0$ by Eq. 21, and both sides of Eq. 22 equal $+\infty$. Otherwise, Eq. 21 gives $\log p(k \mid \boldsymbol{x}_s) = \log w_k + \ell_k(\boldsymbol{x}_s) - \log p(\boldsymbol{x}_s)$ for every k with $w_k > 0$. Substituting this into the definition of the divergence yields

$$\begin{aligned}\mathrm{KL}\big(w \,\big\|\, p(\cdot \mid \boldsymbol{x}_s)\big) \;&=\; \sum_{k:\, w_k > 0} w_k\big[\log w_k - \log p(k \mid \boldsymbol{x}_s)\big] \;=\; \sum_{k:\, w_k > 0} w_k\big[\log p(\boldsymbol{x}_s) - \ell_k(\boldsymbol{x}_s)\big] \\ &=\; \log p(\boldsymbol{x}_s) + \mathbb{E}_{\hat{k} \sim w}\big[-\ell_{\hat{k}}(\boldsymbol{x}_s)\big],\end{aligned}$$

which is Eq. 22. By Lemma 1, the divergence is non-negative and vanishes if and only if $p(\cdot \mid \boldsymbol{x}_s) = w$, and by Eq. 21 this holds if and only if $P^k(\boldsymbol{x}_s) = p(\boldsymbol{x}_s)$ for every k with $w_k > 0$. □

Theorem 2 (Surrogate collapse). *Fix $\boldsymbol{x}_t$, let q be the teacher's transition over the masked positions, and let $M \geq 1$. For $p \in \mathcal{F}_M$ with weights w and components $P^1, \dots, P^M$, define the bound of Eq. 10 and the exact objective of Eq. 8 at $\boldsymbol{x}_t$,*

$$\mathcal{L}_{\mathrm{surr}}(p) = \mathbb{E}_{\boldsymbol{x}_s \sim q}\, \mathbb{E}_{\hat{k} \sim w}\big[-\ell_{\hat{k}}(\boldsymbol{x}_s)\big], \qquad \mathcal{L}_{\mathrm{exact}}(p) = \mathbb{E}_{\boldsymbol{x}_s \sim q}\big[-\log p(\boldsymbol{x}_s)\big].$$

(i) The bound depends on the target only through its marginals. *For every* $p \in \mathcal{F}_M$,

$$\mathcal{L}_{\mathrm{surr}}(p) \;=\; \mathrm{H}(q) + \sum_{k=1}^{M} w_k \,\mathrm{KL}\big(q \,\big\|\, P^k\big) \;=\; \sum_i \mathrm{H}(q_i) + \sum_{k=1}^{M} w_k \sum_i \mathrm{KL}\big(q_i \,\big\|\, P_i^k\big) \;\geq\; \mathrm{H}(q) + \mathrm{TC}(q), \tag{23}$$

with equality if and only if $P^k = \Pi(q)$ *for every* k *with* $w_k > 0$.

(ii) Every minimizer of the bound is factorized. *Whether* w *is held fixed or optimized, every minimizer* p *of* $\mathcal{L}_{\mathrm{surr}}$ *over* $\mathcal{F}_M$ *satisfies* $p = \Pi(q)$, $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) = 0$, *and* $\mathrm{KL}(q \,\|\, p) = \mathrm{TC}(q)$.

(iii) The exact objective uses the latent. $\min_{p\in\mathcal{F}_M} \mathcal{L}_{\mathrm{exact}}(p) = \mathrm{H}(q) + \mathcal{D}_M(q)$, *where* $\mathcal{D}_M(q) = \min_{p\in\mathcal{F}_M} \mathrm{KL}(q \,\|\, p)$. *If* $M \geq 2$ *and* $\mathrm{TC}(q) > 0$, *then* $\mathcal{D}_M(q) < \mathrm{TC}(q)$, *so* $\min_{p\in\mathcal{F}_M} \mathcal{L}_{\mathrm{exact}}(p) < \min_{p\in\mathcal{F}_M} \mathcal{L}_{\mathrm{surr}}(p)$ *and every minimizer* $p^\star$ *of* $\mathcal{L}_{\mathrm{exact}}$ *has* $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) > 0$.

Proof. Steps 1 and 2 prove part (i), by expressing the bound as the exact objective plus a divergence and then proving Eq. 23 and its equality case. Step 3 proves part (ii), and Steps 4 to 6 prove part (iii).

Step 1 (the bound equals the exact objective plus a divergence). If $p(\boldsymbol{x}_s) = 0$ for some $\boldsymbol{x}_s \in \mathrm{supp}(q)$, both objectives equal $+\infty$. Otherwise, Lemma 6 applies at every $\boldsymbol{x}_s \in \mathrm{supp}(q)$, and taking the expectation over $\boldsymbol{x}_s \sim q$ gives

$$\mathcal{L}_{\mathrm{surr}}(p) \;=\; \mathcal{L}_{\mathrm{exact}}(p) \;+\; \mathbb{E}_{\boldsymbol{x}_s\sim q}\Big[\mathrm{KL}\big(w \,\big\|\, p(\cdot \mid \boldsymbol{x}_s)\big)\Big]. \tag{24}$$

The second term is non-negative by Lemma 1, so the bound is an upper bound on the exact objective, which is the inequality of Eq. 1 written with a finite index, and by Lemma 6 the bound is tight if and only if all components with positive weight assign the same probability to $\boldsymbol{x}_s$ for q-almost every $\boldsymbol{x}_s$.

Step 2 (Eq. 23 and its equality case). For each k, the cross-entropy decomposition and Theorem 1, applied to the product distribution P^k, give

$$\mathbb{E}_{\boldsymbol{x}_s\sim q}\big[-\log P^k(\boldsymbol{x}_s)\big] \;=\; \mathrm{H}(q)+\mathrm{KL}(q \,\|\, P^k), \qquad \mathrm{KL}(q \,\|\, P^k) \;=\; \mathrm{TC}(q)+\sum_i \mathrm{KL}(q_i \,\|\, P_i^k) \;\geq\; \mathrm{TC}(q), \tag{25}$$

with equality if and only if $P_i^k = q_i$ for every i, that is, if and only if $P^k = \Pi(q)$. Substituting into Eq. 20 and using $\sum_k w_k = 1$ gives $\mathcal{L}_{\mathrm{surr}}(p) = \mathrm{H}(q) + \sum_k w_k \,\mathrm{KL}(q \,\|\, P^k) \geq \mathrm{H}(q) + \mathrm{TC}(q)$, which is Eq. 23. Equality holds if and only if Eq. 25 is an equality for every k with $w_k > 0$, that is, if and only if $P^k = \Pi(q)$ for every such k.

Step 3 (the minimizers of the bound). The kernel with $P^k = \Pi(q)$ for every k belongs to $\mathcal{F}_M$ and attains the value $\mathrm{H}(q) + \mathrm{TC}(q)$, which is therefore the minimum of $\mathcal{L}_{\mathrm{surr}}$ over $\mathcal{F}_M$. By Step 2, a kernel p is a minimizer if and only if $P^k = \Pi(q)$ for every k with $w_k > 0$, and then $p = \sum_k w_k P^k = \Pi(q)$. Under such a p, the conditional distribution of $\boldsymbol{X}_s$ given k equals $\Pi(q)$ for every k with $w_k > 0$, so $\boldsymbol{X}_s$ is independent of k and $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) = 0$. Finally, $\mathrm{KL}(q \,\|\, p) = \mathrm{KL}(q \,\|\, \Pi(q)) = \mathrm{TC}(q)$ by Definition 1.

Step 4 (the exact objective). By Eq. 19, minimizing $\mathcal{L}_{\mathrm{exact}}$ over $\mathcal{F}_M$ amounts to minimizing $\mathrm{KL}(q \,\|\, p)$ over $\mathcal{F}_M$. The minimum is attained, because $\mathcal{F}_M$ is a compact subset of a finite-dimensional space and $p \mapsto \mathrm{KL}(q \,\|\, p)$ is lower semicontinuous, so $\min_{p\in\mathcal{F}_M} \mathcal{L}_{\mathrm{exact}}(p) = \mathrm{H}(q) + \mathcal{D}_M(q)$. A factorized kernel belongs to $\mathcal{F}_M$ for every M, since we may take all components equal to it, and hence $\mathcal{D}_M(q) \leq \mathcal{D}_1(q) = \mathrm{TC}(q)$ by Theorem 1.

Step 5 (the strict gap). Assume $M \geq 2$ and $\mathrm{TC}(q) > 0$. Then $q \neq \Pi(q)$ by Definition 1, and since both are probability distributions on the same finite set, there is a sequence $\boldsymbol{x}^\star$ with $q(\boldsymbol{x}^\star) > \Pi(q)(\boldsymbol{x}^\star)$. In particular, $q(\boldsymbol{x}^\star) > 0$. Hence, $q_i(x^{\star i}) > 0$ for every i and $\Pi(q)(\boldsymbol{x}^\star) > 0$. For $\eta \in [0, 1)$, define the product distribution $h_\eta = \prod_i \big[(1-\eta)\, q_i + \eta\, \delta_{x^{\star i}}\big]$, and for $\varepsilon \in [0, 1)$, define $p_\varepsilon = (1-\varepsilon)\,\Pi(q) + \varepsilon\, h_\eta$, which belongs to $\mathcal{F}_2 \subseteq \mathcal{F}_M$. On $\mathrm{supp}(q)$, we have $p_\varepsilon \geq (1-\varepsilon)\,\Pi(q) > 0$, so $\varepsilon \mapsto \mathrm{KL}(q \,\|\, p_\varepsilon)$ is differentiable on $[0, 1)$, with

$$\frac{d}{d\varepsilon}\mathrm{KL}(q \,\|\, p_\varepsilon)\Big|_{\varepsilon=0} \;=\; -\sum_{\boldsymbol{x}} q(\boldsymbol{x})\, \frac{h_\eta(\boldsymbol{x}) - \Pi(q)(\boldsymbol{x})}{\Pi(q)(\boldsymbol{x})} \;=\; 1 - \sum_{\boldsymbol{x}} q(\boldsymbol{x})\, \frac{h_\eta(\boldsymbol{x})}{\Pi(q)(\boldsymbol{x})}. \tag{26}$$

As $\eta \to 1$, $h_\eta(\boldsymbol{x}) \to \mathbf{1}\{\boldsymbol{x} = \boldsymbol{x}^\star\}$ for every $\boldsymbol{x}$, so the sum in Eq. 26 converges to $q(\boldsymbol{x}^\star)/\Pi(q)(\boldsymbol{x}^\star) > 1$. Fix η close enough to 1 that the derivative in Eq. 26 is negative. Then $\mathrm{KL}(q \,\|\, p_\varepsilon) < \mathrm{KL}(q \,\|\, p_0) = \mathrm{TC}(q)$ for all sufficiently small $\varepsilon > 0$, and therefore $\mathcal{D}_M(q) \le \mathcal{D}_2(q) < \mathrm{TC}(q)$.

Step 6 (the exact minimizers use the latent). Let $M \ge 2$ and $\mathrm{TC}(q) > 0$, and let $p^\star$ minimize $\mathcal{L}_{\mathrm{exact}}$ over $\mathcal{F}_M$. Then $\mathrm{KL}(q \,\|\, p^\star) = \mathcal{D}_M(q) < \mathrm{TC}(q) = \min_{p \in \mathcal{F}_1} \mathrm{KL}(q \,\|\, p)$ by Step 5 and Theorem 1, so $p^\star \notin \mathcal{F}_1$. By Proposition 6, whose proof does not depend on this theorem, $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) = 0$ would place $p^\star$ in $\mathcal{F}_1$. Hence, $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) > 0$ at $p^\star$. □

Corollary 1 (The bound cannot improve on a factorized student)**.** *Fix $\boldsymbol{x}_t$ and q as in Theorem 2. For every $M \ge 1$,*

$$\min_{p \in \mathcal{F}_M} \mathcal{L}_{\mathrm{surr}}(p) \;=\; \min_{p \in \mathcal{F}_1} \mathcal{L}_{\mathrm{exact}}(p) \;=\; \mathrm{H}(q) + \mathrm{TC}(q),$$

and every minimizer of $\mathcal{L}_{\mathrm{surr}}$ over $\mathcal{F}_M$ lies in $\mathcal{F}_1$ and has $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) = 0$. If $M \ge 2$ and $\mathrm{TC}(q) > 0$, then $\min_{p \in \mathcal{F}_M} \mathcal{L}_{\mathrm{exact}}(p) < \min_{p \in \mathcal{F}_M} \mathcal{L}_{\mathrm{surr}}(p)$, and every minimizer of $\mathcal{L}_{\mathrm{exact}}$ over $\mathcal{F}_M$ lies outside $\mathcal{F}_1$ and has $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) > 0$.

Proof. The first equality combines Step 3 of the proof of Theorem 2 with Theorem 1, by which $\min_{p \in \mathcal{F}_1} \mathrm{KL}(q \,\|\, p) = \mathrm{TC}(q)$. By Step 3, every minimizer of the bound equals $\Pi(q)$, which lies in $\mathcal{F}_1$ and has $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) = 0$. The strict inequality follows from Steps 4 and 5, and the last claim is Step 6. □

Remark 2 (Number of latent draws)**.** *Theorem 2 holds for every weight vector w, in particular for the uniform distribution over any finite set of draws from a prior. The collapse therefore follows from the position of the average in Eq. 1, outside the logarithm, and not from the number of draws used to evaluate it. Proposition 3 below treats an average inside the logarithm that is estimated from finitely many draws.*

Theorem 2 describes the global minimizers of the two objectives. The next proposition describes their behavior near a collapsed kernel, which is where a student starts when its components are initialized identically.

Proposition 2 (Second-order expansion at a collapsed kernel)**.** *Fix $\boldsymbol{x}_t$ and q as in Theorem 2, let $M \ge 1$, and let $w \in \Delta^{M-1}$ satisfy $w_k > 0$ for every k. Let $p_0 \in \mathcal{F}_M$ be the collapsed kernel with weights w and components $P^k = \Pi(q)$ for every k. Call a family $a = (a_i^k)_{k \in [M],\, i \in \mathcal{M}(\boldsymbol{x}_t)}$ of functions $a_i^k : \mathcal{S} \cup \{\texttt{[MASK]}\} \to \mathbb{R}$ an* admissible direction *if, for all k, i, and v,*

$$\sum_{v'} q_i(v')\, a_i^k(v') \;=\; 0, \qquad \sum_{k'} w_{k'}\, a_i^{k'}(v) \;=\; 0, \qquad a_i^k(v) \;=\; 0 \;\text{ whenever } q_i(v) = 0. \tag{27}$$

For an admissible direction a and $|\varepsilon| \max_{k,i,v} |a_i^k(v)| < 1$, let $p_\varepsilon \in \mathcal{F}_M$ have weights w and components $P_i^{k,\varepsilon}(v) = q_i(v)\big(1 + \varepsilon\, a_i^k(v)\big)$, and write $\mathcal{B}(\varepsilon) = \mathcal{L}_{\mathrm{surr}}(p_\varepsilon)$ and $\mathcal{E}(\varepsilon) = \mathcal{L}_{\mathrm{exact}}(p_\varepsilon)$. For positions $i \ne j$, let q_{ij} be the marginal of q on (i, j) and $C_{ij}(v, v') = q_{ij}(v, v') - q_i(v)\, q_j(v')$. Then the following three statements hold as $\varepsilon \to 0$.

1. $\mathcal{B}(\varepsilon) \;=\; \mathcal{B}(0) + \dfrac{\varepsilon^2}{2} \displaystyle\sum_k w_k \sum_i \sum_v q_i(v)\, a_i^k(v)^2 + O(\varepsilon^3).$

2. $\mathcal{E}(\varepsilon) \;=\; \mathcal{E}(0) - \varepsilon^2 \displaystyle\sum_k w_k \sum_{i<j} \sum_{v,v'} C_{ij}(v, v')\, a_i^k(v)\, a_j^k(v') + O(\varepsilon^3)$, *and* $\mathcal{E}(0) = \mathcal{B}(0)$.

3. *If $a \ne 0$, then $\mathcal{B}(\varepsilon) > \mathcal{B}(0)$ for all sufficiently small $\varepsilon \ne 0$. If $M \ge 2$ and $C_{ij} \ne 0$ for some pair $i \ne j$, then there is an admissible direction a with $\mathcal{E}(\varepsilon) < \mathcal{E}(0)$ for all sufficiently small $\varepsilon \ne 0$.*

Proof. By Eq. 27, each $P_i^{k,\varepsilon}$ is a probability distribution with the same support as q_i, so $p_\varepsilon(\boldsymbol{x}_s) > 0$ for every $\boldsymbol{x}_s \in \mathrm{supp}(q)$, and every quantity below is finite. Throughout, $O(\varepsilon^3)$ denotes a remainder bounded by $c\,|\varepsilon|^3$ for a constant c that depends on q, w, and a but not on $\boldsymbol{x}_s$. Such remainders arise

from the expansion $\log(1+x) = x - x^2/2 + O(x^3)$, which holds uniformly for $|x| \le 1/2$, and from products of finitely many bounded terms.

Step 1 (statement 1). By Step 2 of the proof of Theorem 2, $\mathcal{B}(\varepsilon) = \mathrm{H}(q) + \mathrm{TC}(q) + \sum_k w_k \sum_i \mathrm{KL}\big(q_i \,\|\, P_i^{k,\varepsilon}\big)$. For fixed k and i, write $a = a_i^k$. Then

$$\mathrm{KL}\big(q_i \,\|\, P_i^{k,\varepsilon}\big) \;=\; -\sum_v q_i(v)\,\log\big(1+\varepsilon\, a(v)\big) \;=\; -\varepsilon \sum_v q_i(v)\, a(v) + \frac{\varepsilon^2}{2}\sum_v q_i(v)\, a(v)^2 + O(\varepsilon^3),$$

and the linear term vanishes by the first condition in Eq. 27. Summing over i and averaging over k gives statement 1, since $\mathcal{B}(0) = \mathrm{H}(q) + \mathrm{TC}(q)$.

Step 2 (statement 2). For $\boldsymbol{x}_s \in \mathrm{supp}(q)$,

$$p_\varepsilon(\boldsymbol{x}_s) \;=\; \sum_k w_k \prod_i q_i(x_s^i)\big(1+\varepsilon\, a_i^k(x_s^i)\big) \;=\; \Pi(q)(\boldsymbol{x}_s) \sum_k w_k \prod_i \big(1+\varepsilon\, a_i^k(x_s^i)\big).$$

Expanding the product, $\prod_i(1+\varepsilon a_i) = 1+\varepsilon\sum_i a_i + \varepsilon^2 \sum_{i<j} a_i a_j + O(\varepsilon^3)$, and the second condition in Eq. 27 removes the linear term after averaging over k. Hence, $p_\varepsilon(\boldsymbol{x}_s) = \Pi(q)(\boldsymbol{x}_s)\big[1+\varepsilon^2 c(\boldsymbol{x}_s) + O(\varepsilon^3)\big]$ with $c(\boldsymbol{x}_s) = \sum_k w_k \sum_{i<j} a_i^k(x_s^i)\, a_j^k(x_s^j)$, and therefore $-\log p_\varepsilon(\boldsymbol{x}_s) = -\log \Pi(q)(\boldsymbol{x}_s) - \varepsilon^2 c(\boldsymbol{x}_s) + O(\varepsilon^3)$. Taking the expectation over $\boldsymbol{x}_s \sim q$ gives $\mathcal{E}(\varepsilon) = \mathcal{E}(0) - \varepsilon^2\, \mathbb{E}_q[c(\boldsymbol{X}_s)] + O(\varepsilon^3)$, where $\mathcal{E}(0) = \mathbb{E}_q[-\log \Pi(q)(\boldsymbol{X}_s)] = \mathrm{H}(q) + \mathrm{TC}(q) = \mathcal{B}(0)$ by Definition 1, and

$$\mathbb{E}_q[c(\boldsymbol{X}_s)] \;=\; \sum_k w_k \sum_{i<j} \sum_{v,v'} q_{ij}(v,v')\, a_i^k(v)\, a_j^k(v').$$

By the first condition in Eq. 27, $\sum_{v,v'} q_i(v)\, q_j(v')\, a_i^k(v)\, a_j^k(v') = 0$ for every k and every pair $i<j$, so $q_{ij}(v,v')$ may be replaced by $C_{ij}(v,v')$ in the last display, which gives statement 2.

Step 3 (statement 3). If $a \neq 0$, the third condition in Eq. 27 gives $a_i^k(v) \neq 0$ for some (k,i,v) with $q_i(v) > 0$, so the coefficient of ε^2 in statement 1 is positive, and $\mathcal{B}(\varepsilon) > \mathcal{B}(0)$ for all sufficiently small $\varepsilon \neq 0$. Now let $M \ge 2$ and $C_{ij} \neq 0$ for some $i \neq j$. Since $\sum_v C_{ij}(v,v') = 0$ for every v' and $\sum_{v'} C_{ij}(v,v') = 0$ for every v, the matrix C_{ij} has a positive entry $C_{ij}(v_0, v_0') > 0$, and then $q_{ij}(v_0, v_0') > 0$, so $q_i(v_0) > 0$ and $q_j(v_0') > 0$. Define $u_i(v) = \big(\mathbf{1}\{v = v_0\} - q_i(v_0)\big)\,\mathbf{1}\{q_i(v) > 0\}$, define u_j in the same way from v_0', and let $u_l = 0$ at every other position l. Set $a^1 = u$, $a^2 = -(w_1/w_2)\, u$, and $a^k = 0$ for $k \ge 3$. The three conditions in Eq. 27 hold, because $\sum_v q_i(v)\, u_i(v) = q_i(v_0) - q_i(v_0) = 0$ and likewise at j, $w_1 a^1 + w_2 a^2 = 0$, and u vanishes outside the supports. Along this direction, $\sum_k w_k\, a_i^k(v)\, a_j^k(v') = w_1(1+w_1/w_2)\, u_i(v)\, u_j(v')$, and every pair of positions other than (i,j) contributes nothing to statement 2. Moreover, $C_{ij}(v,v') = 0$ whenever $q_i(v) = 0$ or $q_j(v') = 0$, and the rows and columns of C_{ij} sum to zero, so

$$\sum_{v,v'} C_{ij}(v,v')\, u_i(v)\, u_j(v') \;=\; C_{ij}(v_0, v_0') \;>\; 0.$$

Hence, the coefficient of ε^2 in statement 2 equals $-w_1(1+w_1/w_2)\, C_{ij}(v_0, v_0') < 0$, and $\mathcal{E}(\varepsilon) < \mathcal{E}(0)$ for all sufficiently small $\varepsilon \neq 0$. □

Remark 3 (Interpretation). *Both objectives are stationary at a collapsed kernel, since the first-order terms in statements 1 and 2 vanish. Along every admissible direction, the bound has a strict local minimum there, whereas the exact objective decreases along a direction aligned with any pairwise dependence of the target. Under the bound, a kernel that has begun to use its components is pulled back to the collapsed one, and under the exact objective it is pushed away. Dependence of order higher than two enters the expansion of $\mathcal{E}$ at the corresponding higher order in ε.*

Proposition 3 (Marginalization estimated from latent samples). *Fix $\boldsymbol{x}_t$, let q be a distribution on $\mathcal{X}$, for instance the teacher's transition from $\boldsymbol{x}_t$, let λ be a latent variable with an arbitrary prior, and consider a student of the form*

$$p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t) \;=\; \mathbb{E}_\lambda\big[p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t; \lambda)\big], \qquad p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t; \lambda) \;=\; \prod_{i=1}^{L} P_i^\theta\big(x_s^i \mid \boldsymbol{x}_t, \lambda\big),$$

where each factor is positive on the support of q. Write $\mathcal{L}_\infty = -\mathbb{E}_{\boldsymbol{x}_s\sim q}[\log p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t)]$ for the exact objective, and write

$$\widehat{\mathcal{L}}_R \;=\; -\,\mathbb{E}_{\boldsymbol{x}_s\sim q}\,\mathbb{E}_{\lambda_1,\dots,\lambda_R}\Big[\log \tfrac{1}{R}\textstyle\sum_{j=1}^{R} p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t;\lambda_j)\Big]$$

for the objective obtained by estimating the average with R independent draws from the prior. Then the following three statements hold.

1. *$\widehat{\mathcal{L}}_R \geq \mathcal{L}_\infty$ for all $R \geq 1$.*
2. *$\widehat{\mathcal{L}}_{R+1} \leq \widehat{\mathcal{L}}_R$ for all $R \geq 1$.*
3. *Equality $\widehat{\mathcal{L}}_R = \mathcal{L}_\infty$ holds if and only if $p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t;\lambda)$ does not depend on λ, for q-almost every $\boldsymbol{x}_s$.*

Proof. Fix an $\boldsymbol{x}_s$ in the support of q and write $Z_R = \frac{1}{R}\sum_{j=1}^{R} p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t;\lambda_j)$ for the estimate at that $\boldsymbol{x}_s$. The draws are independent and identically distributed, so $\mathbb{E}[Z_R] = p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t)$ for all R, and the positivity assumption keeps $\log Z_R$ finite.

Part 1. The logarithm is concave. Hence, Jensen's inequality implies $\mathbb{E}[\log Z_R] \leq \log \mathbb{E}[Z_R] = \log p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t)$. Negating this inequality and taking the expectation over $\boldsymbol{x}_s \sim q$ produces $\widehat{\mathcal{L}}_R \geq \mathcal{L}_\infty$.

Part 2. The set $\{1,\dots,R+1\}$ has exactly $R+1$ subsets of size R, and each index belongs to exactly R of them. Averaging the R-sample estimates over these subsets therefore returns the $(R+1)$-sample estimate,

$$\mathbb{E}_S\Big[\tfrac{1}{R}\textstyle\sum_{j\in S} p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t;\lambda_j)\Big] \;=\; \tfrac{1}{R+1}\textstyle\sum_S \tfrac{1}{R}\sum_{j\in S} p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t;\lambda_j) \;=\; \tfrac{1}{R+1}\textstyle\sum_{j=1}^{R+1} p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t;\lambda_j) \;=\; Z_{R+1},$$

where S is uniform on those subsets, and the second equality counts each index R times. We apply Jensen's inequality to this inner average, which yields $\log Z_{R+1} \geq \mathbb{E}_S\big[\log \frac{1}{R}\sum_{j\in S} p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t;\lambda_j)\big]$. Taking the expectation over the draws and over $\boldsymbol{x}_s$ then produces $\widehat{\mathcal{L}}_{R+1} \leq \widehat{\mathcal{L}}_R$, which is the bound of Burda et al. (2015) with the prior in the role of the proposal distribution.

Part 3. The logarithm is strictly concave, so equality holds in Part 1 at a given $\boldsymbol{x}_s$ if and only if Z_R is constant almost surely. At $R = 1$, this says that $p^\theta(\boldsymbol{x}_s \mid \boldsymbol{x}_t;\lambda)$ is constant in λ almost surely. At $R \geq 2$, the same conclusion follows by holding $\lambda_2,\dots,\lambda_R$ fixed and varying λ_1 alone. Finally, the equality must hold under the expectation over $\boldsymbol{x}_s$, which establishes the condition at q-almost every $\boldsymbol{x}_s$. □

Remark 4 (Estimated marginalization)**.** *The excess $\widehat{\mathcal{L}}_R - \mathcal{L}_\infty$ is non-negative, non-increasing in R, and zero only when the latent carries no information about the step. An objective that estimates the average over the latent from finitely many draws therefore penalizes the information that the latent carries, and the penalty vanishes only in the limit $R \to \infty$. For a finite index, the sum over the M components in Eq. 7 evaluates $\mathcal{L}_\infty$ itself, so no such penalty arises.*

Remark 5 (The mixing weights)**.** *Hold the components fixed and regard both objectives as functions of the weights $w \in \Delta^{M-1}$. The exact objective is convex in w, because $p(\boldsymbol{x}_s) = \sum_k w_k P^k(\boldsymbol{x}_s)$ is linear in w, the negative logarithm of a positive linear function is convex, and an expectation of convex functions is convex. Its partial derivatives are*

$$\frac{\partial}{\partial w_k}\,\mathcal{L}_{\text{exact}}(p) \;=\; -\,\mathbb{E}_{\boldsymbol{x}_s\sim q}\Big[\frac{P^k(\boldsymbol{x}_s)}{p(\boldsymbol{x}_s)}\Big] \;=\; -\frac{1}{w_k}\,\mathbb{E}_{\boldsymbol{x}_s\sim q}\big[p(k \mid \boldsymbol{X}_s)\big],$$

where the second equality is Eq. 21, so each weight is driven by the posterior probability of its own component. The bound is linear in w by Eq. 20, with coefficients $\mathbb{E}_{\boldsymbol{x}_s\sim q}[-\log P^k(\boldsymbol{x}_s)]$ that depend on q only through its marginals by Eq. 25. A linear function on the simplex attains its minimum at a vertex, so for fixed components the bound is minimized by placing all weight on one component, and Lemma 6 shows that the slack of the bound vanishes at such a weighting as well. The bound thus becomes tight in two ways, through identical components and through a degenerate weighting,

and both remove the mixture. When the latent is drawn from a fixed prior, as in Di4C, no term of the objective depends on the weights, and Theorem 2 applies with w equal to that prior. A fixed prior also restricts the family, since the kernels of $\mathcal{F}_M$ with prescribed weights form a proper subset of $\mathcal{F}_M$ for $M \geq 2$, so the divergence attainable with fixed weights is at least $\mathcal{D}_M(q)$.

Proposition 4 (Terms that admit a factorized optimum)**.** *Fix $\boldsymbol{x}_t$ and let J assign a real number to each kernel $p \in \mathcal{F}_M$, as one term of a training objective does. We say that J* admits a factorized optimum *if*

$$\inf_{p \in \mathcal{F}_M} J(p) \;=\; \inf_{p \in \mathcal{F}_1} J(p),$$

so that restricting the student to a single factorized kernel costs nothing on that term. Let r be a distribution on $\mathcal{X}$ that does not depend on θ, with position marginals $r_1, \ldots, r_L$, write $\bar{P}_i = \sum_k w_k P_i^k$ for the i-th position marginal of p, and write $\mathcal{D}_M(r) = \inf_{p \in \mathcal{F}_M} \mathrm{KL}(r \,\|\, p)$ as in Theorem 2. Then the following three statements hold.

1. *If $J(p) = \sum_{i=1}^{L} \mathrm{KL}\big(r_i \,\big\|\, \bar{P}_i\big)$, then $J(p) = J(\Pi(p))$ for all $p \in \mathcal{F}_M$, and J admits a factorized optimum with infimum 0.*
2. *The bound of Eq. 10 admits a factorized optimum, and the kernel it induces at its minimum lies in $\mathcal{F}_1$.*
3. *If $J(p) = \mathrm{KL}(r \,\|\, p)$, then $\inf_{p \in \mathcal{F}_1} J(p) = \mathrm{TC}(r)$ and $\inf_{p \in \mathcal{F}_M} J(p) = \mathcal{D}_M(r) \leq \mathrm{TC}(r)$. Hence, J admits a factorized optimum if and only if $\mathcal{D}_M(r) = \mathrm{TC}(r)$, and it does so in particular whenever r factorizes across positions.*

Proof. Part 1. The i-th position marginal of $\Pi(p) = \prod_i \bar{P}_i$ is $\bar{P}_i$ itself. The value of J depends on the student only through the L marginals $\bar{P}_1, \ldots, \bar{P}_L$, and p and $\Pi(p)$ share all of them, so $J(p) = J(\Pi(p))$. The kernel $\Pi(p)$ is a product of distributions over single positions. Hence, $\Pi(p) \in \mathcal{F}_1$, and the two infima in the definition therefore coincide. Setting $\bar{P}_i = r_i$ for all i is admissible within $\mathcal{F}_1$ and attains $J = 0$, which Lemma 1 shows is the smallest value that J can take.

Part 2. By Step 3 of the proof of Theorem 2, every minimizer of Eq. 10 over $\mathcal{F}_M$ equals $\Pi(q)$, which belongs to $\mathcal{F}_1$. The minimum is therefore attained inside $\mathcal{F}_1$, and the two infima coincide.

Part 3. Theorem 1 states that $\min_{p \in \mathcal{F}_1} \mathrm{KL}(r \,\|\, p) = \mathrm{TC}(r)$, and $\mathcal{D}_M(r)$ is the corresponding infimum over $\mathcal{F}_M$ by definition. Any factorized kernel belongs to $\mathcal{F}_M$, because we may set all M components equal to it, so $\mathcal{F}_1 \subseteq \mathcal{F}_M$ and an infimum over the larger family is no larger, so $\mathcal{D}_M(r) \leq \mathrm{TC}(r)$. The stated equivalence is then the definition. If r factorizes across positions, then $\mathrm{TC}(r) = 0$ and $r = \Pi(r)$ itself belongs to $\mathcal{F}_1$, so both infima equal 0. □

Part 1 shows that a per-position term scores a mixture and the product of its own marginals identically, and Part 3 shows that only a divergence evaluated at the joint kernel against a target that does not factorize can prefer a mixture. We now apply these facts, together with Theorem 2 and Proposition 3, to the terms of the Di4C objective.

Definition 2 (Di4C objectives at a fixed state)**.** *We use the notation of Hayakawa et al. (2024) in this definition and in Proposition 5. In that notation, time decreases toward the clean end at 0, and a student is a mixture $p = \mathbb{E}_{\lambda \sim \pi}[P^\lambda]$ of product distributions $P^\lambda = \prod_i P_i^\lambda$ over a latent λ with a fixed prior π. Fix a state $\boldsymbol{x}_t$. Let $q^\psi = \prod_i q_i^\psi$ be the teacher's posterior over clean sequences at $\boldsymbol{x}_t$, which is a product distribution because the teacher is a product model, let $u < t$ be one sub-step closer to the clean end, let $p^\psi_{u|t}(\cdot \mid \boldsymbol{x}_t)$ be the teacher's transition from t to u, and let $q_{0|t}(\cdot \mid \boldsymbol{x}_t)$ be the data conditional at $\boldsymbol{x}_t$. Write $\bar{P}_i = \mathbb{E}_{\lambda \sim \pi}[P_i^\lambda]$ and $\Pi(p) = \prod_i \bar{P}_i$. The four terms of the*

objective, as defined in Hayakawa et al. (2024, Eqs. 8 and 9, App. A.1 and A.3), are

$$\widetilde{\mathcal{L}}_{\mathrm{distil}}(p) \;=\; \mathbb{E}_{\lambda\sim\pi}\Big[\sum_i \mathrm{KL}\big(q_i^{\psi} \,\big\|\, P_i^{\lambda}\big)\Big], \tag{28}$$

$$\mathcal{L}_{\mathrm{marginal}}(p) \;=\; \sum_i \mathrm{KL}\big(q_i^{\psi} \,\big\|\, \bar{P}_i\big), \tag{29}$$

$$\mathcal{L}_{\mathrm{consis}}(p) \;=\; \mathrm{KL}\big(r \,\big\|\, p\big), \qquad r \;=\; \sum_{\boldsymbol{x}_u} p^{\psi}_{u|t}(\boldsymbol{x}_u \mid \boldsymbol{x}_t)\, p^{\mathrm{sg}}_{0|u}(\cdot \mid \boldsymbol{x}_u), \tag{30}$$

$$\mathcal{L}_{\mathrm{corr}}(p) \;=\; \mathbb{E}_{\boldsymbol{x}_0\sim q_{0|t}(\cdot|\boldsymbol{x}_t)}\big[-\log p(\boldsymbol{x}_0) + \log \Pi(p)(\boldsymbol{x}_0)\big], \tag{31}$$

where $p^{\mathrm{sg}}_{0|u}$ is the student's posterior at time u with the gradient blocked. Eq. 28 is the convexity bound of $\mathrm{KL}(q^{\psi} \,\|\, p)$. The R-sample estimates $\widehat{\mathcal{L}}_{\mathrm{consis}}$ and $\widehat{\mathcal{L}}_{\mathrm{corr}}$ replace $-\log p(\boldsymbol{x})$ in Eqs. 30 and 31 by the estimate

$$-\,\mathbb{E}_{\lambda_1,\dots,\lambda_R\sim\pi}\Big[\log \frac{1}{R}\sum_{j=1}^{R} P^{\lambda_j}(\boldsymbol{x})\Big] \tag{32}$$

of Hayakawa et al. (2024, Eq. 10).

Proposition 5 (Terms of the Di4C objective)**.** *Let the terms be as in Definition 2. Then the following four statements hold.*

1. *$\widetilde{\mathcal{L}}_{\mathrm{distil}}(p) = \mathbb{E}_{\lambda\sim\pi}\big[\mathrm{KL}(q^{\psi} \,\|\, P^{\lambda})\big] \geq 0$, with equality if and only if $P^{\lambda} = q^{\psi}$ for π-almost every λ. Every minimizer satisfies $p = q^{\psi}$ and $\mathrm{I}(\lambda; \boldsymbol{X}_0 \mid \boldsymbol{x}_t) = 0$.*
2. *$\mathcal{L}_{\mathrm{marginal}}(p) = \mathcal{L}_{\mathrm{marginal}}(\Pi(p))$ for every p, and $\mathcal{L}_{\mathrm{marginal}}(p) = 0$ if and only if $\bar{P}_i = q_i^{\psi}$ for every i. In particular, the factorized student $p = q^{\psi}$ is a minimizer.*
3. *The target r of $\mathcal{L}_{\mathrm{consis}}$ is determined by the student at time u and by the single teacher transition $p^{\psi}_{u|t}(\cdot \mid \boldsymbol{x}_t)$. If $p^{\mathrm{sg}}_{0|u}(\cdot \mid \boldsymbol{x}_u)$ is a product distribution for every $\boldsymbol{x}_u$, then r is a mixture of product distributions with mixing distribution $p^{\psi}_{u|t}(\cdot \mid \boldsymbol{x}_t)$.*
4. *$\widehat{\mathcal{L}}_{\mathrm{consis}}(p) \geq \mathcal{L}_{\mathrm{consis}}(p)$ and $\widehat{\mathcal{L}}_{\mathrm{corr}}(p) \geq \mathcal{L}_{\mathrm{corr}}(p)$ for every $R \geq 1$, with equality if and only if $P^{\lambda}(\boldsymbol{x})$ does not depend on λ for r-almost every $\boldsymbol{x}$, respectively for $q_{0|t}(\cdot \mid \boldsymbol{x}_t)$-almost every $\boldsymbol{x}$.*

Proof. Statement 1. Since q^{ψ} is a product distribution, $\mathrm{TC}(q^{\psi}) = 0$, and Theorem 1 applied to the product distribution P^{λ} gives $\sum_i \mathrm{KL}(q_i^{\psi} \,\|\, P_i^{\lambda}) = \mathrm{KL}(q^{\psi} \,\|\, P^{\lambda})$ for every λ. Taking the expectation over λ gives the identity. By Lemma 1, the integrand is non-negative and vanishes if and only if $P^{\lambda} = q^{\psi}$, which gives the equality case. At a minimizer, $P^{\lambda} = q^{\psi}$ for π-almost every λ, so $p = \mathbb{E}_{\lambda\sim\pi}[P^{\lambda}] = q^{\psi}$, and the conditional distribution of $\boldsymbol{X}_0$ given λ does not depend on λ, so $\mathrm{I}(\lambda; \boldsymbol{X}_0 \mid \boldsymbol{x}_t) = 0$.

Statement 2. The marginal of $\Pi(p)$ at position i is $\bar{P}_i$, which is also the marginal of p, and Eq. 29 depends on p only through $\bar{P}_1, \bar{P}_2, \dots$, so the two values coincide. By Lemma 1, each summand is non-negative and vanishes if and only if $\bar{P}_i = q_i^{\psi}$, and the student $p = q^{\psi}$ satisfies $\bar{P}_i = q_i^{\psi}$ for every i.

Statement 3. The first claim is the definition of r in Eq. 30. If every $p^{\mathrm{sg}}_{0|u}(\cdot \mid \boldsymbol{x}_u)$ is a product distribution, then Eq. 30 exhibits r as the mixture of these product distributions with weights $p^{\psi}_{u|t}(\boldsymbol{x}_u \mid \boldsymbol{x}_t)$.

Statement 4. By Eq. 30, $\mathcal{L}_{\mathrm{consis}}(p) = \mathbb{E}_{\boldsymbol{x}\sim r}[-\log p(\boldsymbol{x})] - \mathrm{H}(r)$, and $\widehat{\mathcal{L}}_{\mathrm{consis}}(p)$ replaces the first term by the estimate $\widehat{\mathcal{L}}_R$ of Proposition 3 with r in the role of q. Parts 1 and 3 of that proposition give the inequality and its equality case. The same argument applies to Eq. 31 with $q_{0|t}(\cdot \mid \boldsymbol{x}_t)$ in the role of q, because the term $\log \Pi(p)(\boldsymbol{x}_0)$ is common to $\mathcal{L}_{\mathrm{corr}}$ and $\widehat{\mathcal{L}}_{\mathrm{corr}}$. □

Remark 6 (Scope of the comparison)**.** *Proposition 5 is related to the loss terms as functions of the student at a fixed state, and it makes no claim about the optimum of the full Di4C objective. Indeed,*

Hayakawa et al. (2024, Thm. 2) show that the exact distillation and consistency losses vanish when the student equals the composition of the teacher over the interval, which is a correlated kernel. What the proposition establishes is where dependence between positions can enter the objective. The teacher enters only through the product distribution q^{ψ}, in Eqs. 28 and 29, and through the single transition $p^{\psi}_{u|t}$, in Eq. 30. The two terms anchored to the teacher are minimized by, or indifferent to, a student whose latent carries no information, and the target of the $\mathcal{L}_{\text{consis}}$ is the student's own posterior at an earlier time. The target of the consistency term can only capture dependence, and by statement 4, its estimate penalizes the information that the latent carries. The transition loss of Eq. 8 differs on both points, since its target is the teacher's transition over the whole step with the dependence measured by Theorem 1, and the average over the component is evaluated in closed form.

B.5 Information identity

Proposition 6 (Information identity and two upper bounds)**.** *Fix $\boldsymbol{x}_t$ and let $p \in \mathcal{F}_M$ have weights w and components $P^1, \dots, P^M$. We generate the pair $(k, \boldsymbol{X}_s)$ by sampling $k \sim w$ and then $\boldsymbol{X}_s \sim \prod_i P^k_i$, so that the positions of $\boldsymbol{X}_s$ are independent given k and the marginal of $\boldsymbol{X}_s$ is p itself. Then the transition information of the step satisfies*

$$\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) \;\le\; \log M,$$

and the total correlation of the step satisfies the identity and the two bounds

$$\mathrm{TC}(p) \;=\; \sum_{i=1}^{L} \mathrm{I}\big(k; X^i_s \mid \boldsymbol{x}_t\big) \;-\; \mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) \;\le\; (L-1)\,\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) \;\le\; (L-1)\log M. \quad (33)$$

Moreover, $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) = 0$ implies that p is a product distribution, so a mixture that carries no transition information belongs to $\mathcal{F}_1$. The bound $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) \le \log M$ holds with equality if and only if the weights are uniform and k is almost surely determined by $\boldsymbol{X}_s$. The bound $\mathrm{TC}(p) \le (L-1)\,\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t)$ holds with equality if and only if each position carries as much information about k as the whole sequence does.

Proof. We prove one claim in each of the first four steps, and we give a counterexample in the fifth. All five rest on a single property of how we generate the pair $(k, \boldsymbol{X}_s)$. The generative process defines the joint distribution

$$p(k, \boldsymbol{x}_s) \;=\; w_k \prod_{i=1}^{L} P^k_i(x^i_s), \quad (34)$$

whose marginal over $\boldsymbol{x}_s$ is the mixture p and whose marginal over k is the weight vector w. Eq. 34 shows the property we use, namely that the positions of $\boldsymbol{X}_s$ are conditionally independent given k. The proof also follows the derivations from Ahmed et al. (2026). Hence, Lemma 2 applies, and we obtain

$$\mathrm{H}(\boldsymbol{X}_s \mid k) \;=\; \sum_{i=1}^{L} \mathrm{H}(X^i_s \mid k). \quad (35)$$

Step 1 (the bound $\mathrm{I}(k; \boldsymbol{X}_s) \le \log M$). We write the mutual information as a difference of entropies,

$$\mathrm{I}(k; \boldsymbol{X}_s) \;=\; \mathrm{H}(k) - \mathrm{H}(k \mid \boldsymbol{X}_s) \;\le\; \mathrm{H}(k) \;\le\; \log M, \quad (36)$$

where the first inequality holds because conditional entropy is non-negative, and the second holds because k takes at most M values. Equality in both requires $\mathrm{H}(k \mid \boldsymbol{X}_s) = 0$ and $\mathrm{H}(k) = \log M$. The first condition says that $\boldsymbol{X}_s$ almost surely determines k, and the second says that the weights are uniform. Conversely, these two conditions imply equality.

Step 2 (the identity). We expand both mutual-information terms as differences of entropies and regroup, and obtain

$$\begin{aligned}\sum_{i=1}^{L} \mathrm{I}(k; X_s^i) \;-\; \mathrm{I}(k; \boldsymbol{X}_s) \;&=\; \sum_i \big[\mathrm{H}(X_s^i) - \mathrm{H}(X_s^i \mid k)\big] \;-\; \big[\mathrm{H}(\boldsymbol{X}_s) - \mathrm{H}(\boldsymbol{X}_s \mid k)\big] \qquad \text{(expand each I)} \\ &=\; \Big[\sum_i \mathrm{H}(X_s^i) - \mathrm{H}(\boldsymbol{X}_s)\Big] \;-\; \Big[\sum_i \mathrm{H}(X_s^i \mid k) - \mathrm{H}(\boldsymbol{X}_s \mid k)\Big] \qquad \text{(regroup)} \\ &=\; \mathrm{TC}(p) \;-\; 0, \qquad \text{(Definition 1; Eq. 35)} \end{aligned} \tag{37}$$

where the first bracket is the total correlation of p (the distribution of $\boldsymbol{X}_s$), and the second bracket vanishes by Eq. 35. Rearranging produces the identity of Eq. 33.

Step 3 (the bound $\mathrm{TC}(p) \le (L-1)\,\mathrm{I}(k; \boldsymbol{X}_s)$). Each position X_s^i is a deterministic function of the sequence $\boldsymbol{X}_s$. Hence, Lemma 3 implies $\mathrm{I}(k; X_s^i) \le \mathrm{I}(k; \boldsymbol{X}_s)$ for all i. Summing over the L positions and substituting into the identity of Step 2 produces

$$\mathrm{TC}(p) \;=\; \sum_{i=1}^{L} \mathrm{I}(k; X_s^i) - \mathrm{I}(k; \boldsymbol{X}_s) \;\le\; L\,\mathrm{I}(k; \boldsymbol{X}_s) - \mathrm{I}(k; \boldsymbol{X}_s) \;=\; (L-1)\,\mathrm{I}(k; \boldsymbol{X}_s), \tag{38}$$

and combining this bound with Step 1 produces $\mathrm{TC}(p) \le (L-1)\log M$. Equality in the first inequality holds if and only if $\mathrm{I}(k; X_s^i) = \mathrm{I}(k; \boldsymbol{X}_s)$ for all i, that is, when observing a single position tells us as much about k as observing the whole sequence.

Step 4 (the degeneracy claim). Suppose that $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t) = 0$. Mutual information vanishes if and only if its two arguments are independent. Hence, $\boldsymbol{X}_s$ is independent of k, and for all k with $w_k > 0$, the conditional distribution of $\boldsymbol{X}_s$ given k equals the unconditional distribution,

$$p \;=\; \prod_{i=1}^{L} P_i^k. \tag{39}$$

The mixture p is therefore a product distribution and belongs to $\mathcal{F}_1$. Alternatively, the claim follows from the identity of Eq. 33, since $\mathrm{I}(k; X_s^i) \le \mathrm{I}(k; \boldsymbol{X}_s) = 0$ for all i implies $\mathrm{TC}(p) = 0$, and a distribution with zero total correlation is a product distribution by Definition 1.

Step 5 (the converse fails). We take $M = 2$, $L = 2$, two tokens $a \neq b$, uniform weights $w = (\frac{1}{2}, \frac{1}{2})$, and components that differ only at the first position, where r is an arbitrary distribution at the second position,

$$P^1 \;=\; \delta_a \otimes r, \qquad P^2 \;=\; \delta_b \otimes r. \tag{40}$$

The mixture $p = \big(\frac{1}{2}\delta_a + \frac{1}{2}\delta_b\big) \otimes r$ is a product distribution with $\mathrm{TC}(p) = 0$, yet the first position identifies the component. Hence, $\mathrm{I}(k; \boldsymbol{X}_s) \ge \mathrm{I}(k; X_s^1) = \log 2 > 0$. The implication thus holds in one direction only, and a vanishing $\mathrm{I}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t)$ is a stronger condition than a vanishing total correlation. □

B.6 Capacity of a mixture and the number of components

Fix $\boldsymbol{x}_t$ and write q for the teacher's transition, as in App. B.4. For a pair $(Z, \boldsymbol{X})$ with $\boldsymbol{X} \sim q$ and Z taking values in a finite set $\mathcal{Z}$, write $\pi_z = \Pr(Z = z)$ and q_z for the conditional distribution of $\boldsymbol{X}$ given $Z = z$, so that $q = \sum_z \pi_z\, q_z$, and define the conditional total correlation

$$\mathrm{TC}(\boldsymbol{X} \mid Z) \;=\; \sum_z \pi_z\, \mathrm{TC}(q_z) \;=\; \sum_{i=1}^{L} \mathrm{H}(X^i \mid Z) \;-\; \mathrm{H}(\boldsymbol{X} \mid Z), \tag{41}$$

where the second equality follows by applying Definition 1 to each q_z and averaging over z.

Proposition 7 (What a mixture of M product distributions can capture)**.** *Let $(Z, \boldsymbol{X})$ be as above with $\boldsymbol{X} \sim q$ and $|\mathcal{Z}| \le M$. Then the kernel $p_Z = \sum_z \pi_z\, \Pi(q_z)$ belongs to $\mathcal{F}_M$ and satisfies*

$$\mathcal{D}_M(q) \;\le\; \mathrm{KL}(q \,\|\, p_Z) \;\le\; \mathrm{TC}(\boldsymbol{X} \mid Z) \;=\; \mathrm{TC}(q) \;-\; \Big[\sum_{i=1}^{L} \mathrm{I}(Z; X^i) \;-\; \mathrm{I}(Z; \boldsymbol{X})\Big]. \tag{42}$$

In particular, $\mathcal{D}_M(q) = 0$ whenever $M \ge |\mathrm{supp}(q)|$.

Proof. Each $\Pi(q_z)$ is a product distribution, so p_Z is a mixture of at most M product distributions with weights π, and $p_Z \in \mathcal{F}_M$. The first inequality in Eq. 42 is the definition of $\mathcal{D}_M(q)$. For the second inequality, write $q = \sum_z \pi_z\, q_z$ and apply Lemma 5,

$$\mathrm{KL}(q \,\|\, p_Z) \;=\; \mathrm{KL}\Big(\sum_z \pi_z\, q_z \,\Big\|\, \sum_z \pi_z\, \Pi(q_z)\Big) \;\le\; \sum_z \pi_z\, \mathrm{KL}\big(q_z \,\|\, \Pi(q_z)\big) \;=\; \sum_z \pi_z\, \mathrm{TC}(q_z) \;=\; \mathrm{TC}(\boldsymbol{X} \mid Z),$$

where the second equality is Definition 1 and the third is Eq. 41. For the identity, expand Definition 1 and Eq. 41 and regroup,

$$\mathrm{TC}(q) - \mathrm{TC}(\boldsymbol{X} \mid Z) \;=\; \sum_{i=1}^{L} \big[\mathrm{H}(X^i) - \mathrm{H}(X^i \mid Z)\big] \;-\; \big[\mathrm{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{X} \mid Z)\big] \;=\; \sum_{i=1}^{L} \mathrm{I}(Z; X^i) - \mathrm{I}(Z; \boldsymbol{X}).$$

Finally, if $M \ge |\mathrm{supp}(q)|$, take $Z = \boldsymbol{X}$. Then $q_z = \delta_z$ is a point mass for every $z \in \mathrm{supp}(q)$, so $\mathrm{TC}(q_z) = 0$, $p_Z = \sum_z q(z)\, \delta_z = q$, and $\mathcal{D}_M(q) \le \mathrm{KL}(q \,\|\, q) = 0$. □

Corollary 2 (Number of components). *Let $L \ge 2$. If a kernel $p \in \mathcal{F}_M$ attains $\mathrm{KL}(q \,\|\, p) = 0$, then $M \ge \exp\big(\frac{\mathrm{TC}(q)}{L-1}\big)$. Conversely, some $p \in \mathcal{F}_M$ attains $\mathrm{KL}(q \,\|\, p) = 0$ whenever $M \ge |\mathrm{supp}(q)|$.*

Proof. By Lemma 1, $\mathrm{KL}(q \,\|\, p) = 0$ holds if and only if $p = q$, and then $\mathrm{TC}(p) = \mathrm{TC}(q)$. Since $p \in \mathcal{F}_M$, Proposition 6 gives $\mathrm{TC}(p) \le (L-1)\log M$, hence $\mathrm{TC}(q) \le (L-1)\log M$, and dividing by $L - 1 \ge 1$ and exponentiating gives the first claim. The second claim is Proposition 7 with $Z = \boldsymbol{X}$, for which $p_Z = q$. □

C Experimental details

In this section, we provide the datasets, metrics, and training details for the experiments.

C.1 Text

We evaluated MLD3 on the One Billion Word (LM1B) (Chelba et al., 2013) and WikiText-103 (Merity et al., 2016) benchmarks, both with sequences of length $L = 128$. Following Sahoo et al. (2024), we used the uncased BERT WordPiece tokenizer on LM1B (Devlin et al., 2019) and the GPT-2 byte-pair tokenizer on WikiText-103 (Radford et al., 2019). We trained the LM1B teacher, a masked diffusion model, for 200,000 steps (Sahoo et al., 2024; Lou et al., 2023), and the WikiText-103 teacher for 1,000,000 steps. We trained all MLD3 students from random initialization with a batch size of 256, for 50,000 steps on LM1B and 200,000 steps on WikiText-103. The students used $K = 4$ and drew half of their time pairs from outside the K-step sampling schedule ($\eta = 1/2$, defined in App. C.4), with step sizes in $[1/32, 1]$. For these students, the number of teacher sub-steps grew with the step size h from a base of $T_0 = 4$. We applied the auxiliary loss, with weight $\lambda_0 = 1$, to a quarter of each micro-batch. We also trained Di4C on these datasets from random initialization, for the same number of steps as our students.

C.2 Molecules

We followed the molecule generation setup of Park et al. (2025). QM9 has sequences of length $L = 32$ over a vocabulary of 40 tokens, and ZINC-250k has sequences of length $L = 74$ over a vocabulary of 72 tokens. In both datasets, the mask is a reserved token of the vocabulary that does not occur in the data. We trained the MDLM teachers for 50,000 steps on QM9 and 200,000 steps on ZINC-250k. We trained the MLD3 students for 10,000 steps on QM9 and 30,000 steps on ZINC-250k. We also trained Di4C on both datasets under the same setup. The Di4C student draws a scalar latent and appends it to the input as an extra token. Because that latent comes from a fixed continuous prior rather than from a finite router with learned weights, our estimator of transition information does not apply to Di4C.

C.3 Metrics and estimators

For text, we report 4 quantities. The first is the generative perplexity of the samples under GPT-2 Large (Radford et al., 2019). The second is the unigram entropy of the samples. The third is the KL

divergence from the unigram distribution of the samples to that of the data, which detects samples that are diverse but carry the wrong token frequencies. The fourth is the transition information $\widehat{I} = \widehat{\mathrm{I}}(k; \boldsymbol{X}_s \mid \boldsymbol{x}_t)$. Let $r_k \propto w_k^\theta(\boldsymbol{x}_t, t)\, e^{\tilde{\ell}_k(\boldsymbol{x}_s)}$ denote the posterior over the component index given an observed step, where $\tilde{\ell}_k$ is the log-likelihood that component k, renormalized over the clean tokens, assigns to the revealed tokens. Then

$$\widehat{I} \;=\; \frac{1}{N}\sum_{n=1}^{N}\Big[\mathrm{H}\big(w_k^\theta(\boldsymbol{x}_t^{(n)}, t_n)\big) \;-\; \mathrm{H}\big(r^{(n)}\big)\Big], \qquad (\boldsymbol{x}_t^{(n)}, \boldsymbol{x}_s^{(n)}) \sim q_{\mathrm{data}}, \tag{43}$$

where the average is over $N = 1{,}024$ held-out sequences, each corrupted to the start time of a step of the student's K-step sampling schedule.

For molecules, we report the validity, uniqueness, and novelty of the decoded SMILES strings under the definitions and the RDKit implementation of Park et al. (2025). For regulatory DNA, we report the Fréchet biological distance (FBD) between the embeddings that a trained activity predictor assigns to generated and held-out sequences, together with the Wasserstein distance between the two distributions of predicted activity. For images, we decode the tokens with the frozen VQGAN of Chang et al. (2022) and report the Fréchet Inception distance (FID) against precomputed validation statistics, along with precision, recall, density, and coverage.

We evaluated each model at the NFEs in $\{1, \ldots, 8, 16, 32\}$ on a fixed number of samples per dataset, namely 512 sequences for the 2 text datasets, 4,096 for the molecule and DNA datasets, and 10,000 for images. We computed the transition information $\widehat{I}$ of Eq. 43 on held-out data, because it describes the transition of a student.

C.4 Implementation details

Model architecture. A single transformer backbone, which we call the trunk, outputs the router weights and the posteriors of all M components in one forward pass, as shown in Figure 6. Each additional component thus costs only an embedding vector and a scalar router output.

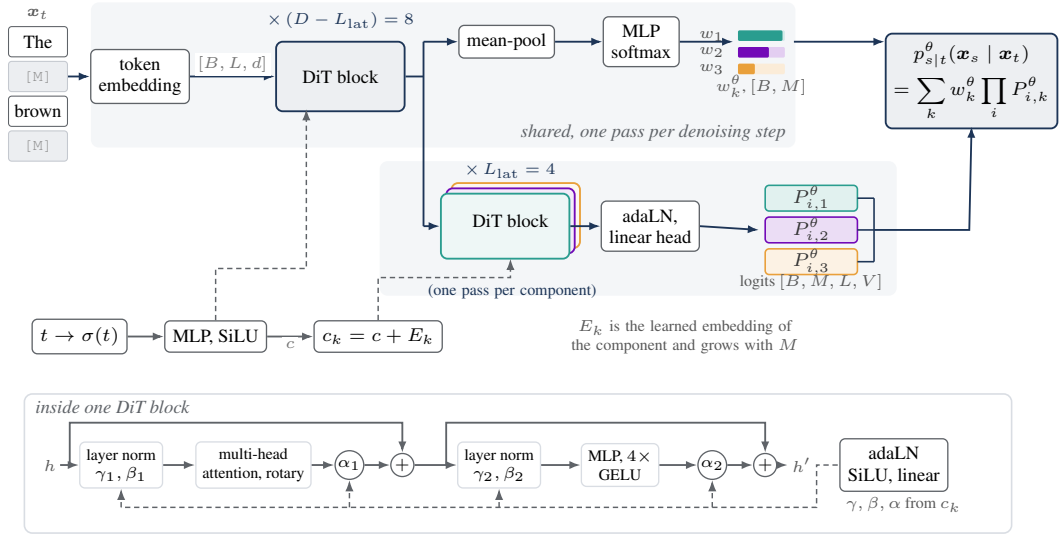

Figure 6: **Illustration of the MLD3 architecture.** The first $D - L_{\mathrm{lat}} = 8$ blocks run once per state and give a shared representation, from which a two-layer head on the mean-pooled hidden states produces the router weights w_k^θ. The remaining $L_{\mathrm{lat}} = 4$ blocks hold one set of weights and run once per evaluated component, where the component enters only through the learned embedding E_k that is added to the conditioning vector. The head is shared as well, and the pass returns one distribution $P_{i,k}^\theta$ per position and per evaluated component, which Eq. 6 combines with the router weights.

The trunk of the MLD3 student is a diffusion transformer (DiT) (Peebles & Xie, 2023) with width 768, depth $D = 12$, 12 attention heads, an adaptive layer-norm conditioning vector of dimension

128, and a sinusoidal time embedding of dimension 256. Each pre-norm block applies bidirectional attention with rotary position embeddings, followed by a two-layer feed-forward network with a $4\times$ expansion, and we modulate each sub-layer with a learned scale, shift, and gate. The output head is untied from the input embedding and initialized to zero. Hence, the model outputs uniform logits at initialization. The conditioning vector contains only the corruption level and the component index. Instead of κ_t, we encode the corruption level with the log-linear noise level $\sigma_t = -\log(1-(1-\varepsilon)(1-\kappa_t))$, where $\varepsilon = 10^{-3}$. This provides the network with a finer resolution near the clean end of the path.

We run the first $D - L_{\text{lat}} = 8$ blocks once, with a conditioning vector that depends only on the corruption level. We then run the remaining $L_{\text{lat}} = 4$ blocks M times under the same weights, once per component, and add a learned embedding of the component index to their conditioning vector. The network thus outputs logits of shape $[B, M, L, V]$, where B is the batch size and the slice $[:, k, :, :]$ is the denoiser of component k, and the components differ only in the modulation of the last L_{lat} blocks. We compute the router from a two-layer head on the mean-pooled output of the shared blocks, because the continuation a partially masked sequence will receive is a global property of $\boldsymbol{x}_t$. We set $m = 2$ as a practical choice for component selection in our experiments, and this value is also the knee of the router entropy curve.

Time sampling. With probability $1-\eta$, we draw a step of the K-step sampling schedule, taking $j \sim \mathcal{U}\{0, \dots, K-1\}$ and $(t, s) = (j/K, (j+1)/K)$. With probability η, we instead draw a step size h log-uniformly on $[h_{\min}, h_{\max}]$, then $t \sim \mathcal{U}[0, 1-h]$ and $s = t + h$. We set $\eta > 0$ to match the density of supervision in methods that draw their time pairs from a fine discretization. Hence, the objective remains the main difference between those methods and ours.

Auxiliary loss. We add to the transition loss an auxiliary loss, namely the marginal loss of Hayakawa et al. (2024), which keeps the average over components inside the divergence. While the transition loss supervises one sampled sequence per step, the auxiliary loss supervises the teacher's full clean-token distribution at each masked position, which anchors the student at times that the transition loss rarely visits. We write p_0^{ψ} for the teacher's clean-token posterior with the mask entry removed, and $\bar{P}_i = \sum_k w_k^{\theta} P_{i,k}^{\theta}$ for the student's posterior at position i averaged over components. Then

$$\mathcal{L}_0(\theta) \;=\; \mathbb{E}_{(t,s)\sim\nu}\, \mathbb{E}_{\boldsymbol{x}_t} \sum_{i \in \mathcal{M}(\boldsymbol{x}_t)} \mathrm{KL}\big(p_{0,i}^{\psi}(\cdot \mid \boldsymbol{x}_t, t) \,\big\|\, \bar{P}_i(\cdot \mid \boldsymbol{x}_t, t)\big). \tag{44}$$

The divergence targets the averaged posterior $\bar{P}_i$ and not an individual component. It thus constrains only the weighted average and leaves the components free to specialize. We apply Eq. 44 one position at a time, and the transition loss remains the source of the signal about dependence between positions.

Joint objective. A router trained from a symmetric initialization can concentrate its weights on a single component early in training and fail to recover the others. To prevent this, we add the entropy regularizer $\mathcal{R}(\theta) = -\lambda[\mathrm{H}(\bar{w}) + \mathbb{E}_{\boldsymbol{x}_t}\mathrm{H}(w^{\theta}(\boldsymbol{x}_t))]$ with $\bar{w} = \mathbb{E}_{\boldsymbol{x}_t}[w^{\theta}(\boldsymbol{x}_t)]$. The first term keeps each component in use across the batch, and the second term prevents the router from concentrating on one component before the components have specialized. The full training objective is then

$$\mathcal{L}(\theta) \;=\; \mathcal{L}_{\text{trans}}(\theta) \;+\; \lambda_0\, \mathcal{L}_0(\theta) \;+\; \mathcal{R}(\theta). \tag{45}$$

Optimization. We trained the students with AdamW at $\beta = (0.9, 0.999)$ and gradient-norm clipping at 1.0, and held the learning rate constant after a linear warmup. We initialized the MLD3 students randomly and trained them with a learning rate of 3×10^{-4}, a warmup of 1,000 steps, and an exponential moving average of the weights with decay 0.9999. We used a batch of 256 sequences accumulated from micro-batches of 32, except on QM9, where we used 1,024 sequences accumulated from micro-batches of 128. On text, molecules, and regulatory DNA, we applied a dropout of 0.1 inside the gated residual of each block, and we applied no dropout on ImageNet. We set the entropy regularizer of Eq. 45 to $\lambda = 0.1$ on both of its terms. We ran the transformer blocks under a bfloat16 autocast and kept the normalization layers and the returned logits in single precision. We ran all experiments on 4 NVIDIA H100 GPUs, each with 80 GB of memory.

C.5 Detailed molecule results

In this section, we report the molecule generation results in full, with each MLD3 student decoded by the spread consensus sampler of Section 3.2. We compare that sampler against the alternative sampling strategies separately, in App. E.2.

Tables 2 and 3 report validity, uniqueness, and novelty across NFEs, because validity alone hides the diversity that a mixture sacrifices. On QM9 at $M = 8$, that loss is substantial. At 2 NFEs, the mixture reaches 0.471 validity against 0.146 for the control, and its uniqueness is 0.096 against 0.216. The $M = 4$ student shows the opposite trade-off on the same dataset, with a lower validity than the control up to 8 NFEs and a higher uniqueness at all NFEs from 2 upward. Both effects fade as the number of steps grows, and by 32 NFEs, the $M = 8$ student leads the control on validity, at 0.903 against 0.825, while its uniqueness is within 5% of the control's. On ZINC-250k, the $M = 8$ student leads the control on validity at all NFEs, and its uniqueness is 56% below the control's at 4 NFEs and 7% below at 32. Validity measures whether a string decodes to a valid molecule. Novelty on QM9 is nearly uninformative, because the dataset enumerates its chemical space, and we therefore assess the QM9 trade-off from validity and uniqueness alone.

Table 2: Validity, uniqueness, and novelty on QM9 across NFEs.

	validity ↑ at NFE						uniqueness ↑ at NFE						novelty ↑ at NFE					
model	1	2	4	8	16	32	1	2	4	8	16	32	1	2	4	8	16	32
MDLM teacher	0.142	0.143	0.254	0.436	0.607	0.714	0.264	**0.387**	0.535	0.694	0.821	0.870	0.199	**0.265**	0.260	0.262	0.286	0.247
Di4C	0.132	0.166	0.286	0.471	0.620	0.710	0.249	0.376	**0.549**	0.688	0.801	0.869	0.179	0.236	**0.270**	**0.281**	**0.297**	**0.255**
MLD3 ($M = 1$)	0.137	0.146	0.257	0.533	0.760	0.825	**0.276**	0.216	0.465	0.612	0.800	0.904	0.198	0.126	0.216	0.203	0.218	0.228
MLD3 ($M = 4$)	0.134	0.124	0.222	0.477	0.766	0.859	0.262	0.269	0.522	**0.711**	**0.862**	**0.943**	0.176	0.173	0.222	0.230	0.189	0.174
MLD3 ($M = 8$)	**0.286**	**0.471**	**0.574**	**0.720**	**0.872**	**0.903**	0.270	0.096	0.177	0.327	0.672	0.862	**0.208**	0.055	0.057	0.056	0.069	0.095

Table 3: Validity, uniqueness, and novelty on ZINC-250k across NFEs.

	validity ↑ at NFE						uniqueness ↑ at NFE						novelty ↑ at NFE					
model	1	2	4	8	16	32	1	2	4	8	16	32	1	2	4	8	16	32
MDLM teacher	0.015	0.096	0.202	0.333	0.464	0.542	0.415	0.302	0.529	0.748	0.861	0.927	0.415	0.302	0.529	0.748	0.860	0.926
Di4C	0.028	0.098	0.155	0.282	0.414	0.520	0.448	0.296	0.525	0.766	0.915	0.961	0.448	0.296	0.525	0.765	0.914	0.959
MLD3 ($M = 1$)	0.014	0.091	0.139	0.180	0.356	0.545	0.533	**0.384**	**0.672**	**0.932**	**0.989**	**0.997**	0.533	**0.384**	**0.672**	**0.932**	**0.989**	**0.995**
MLD3 ($M = 4$)	0.022	0.173	0.226	0.403	0.604	0.741	**0.535**	0.249	0.435	0.662	0.894	0.961	**0.535**	0.249	0.435	0.660	0.892	0.958
MLD3 ($M = 8$)	**0.030**	**0.231**	**0.444**	**0.597**	**0.701**	**0.796**	0.470	0.213	0.294	0.545	0.789	0.926	0.470	0.213	0.294	0.545	0.787	0.925

D Additional experiments

In this section, we present the additional experiments on regulatory DNA and image generation, together with the sample diversity measurements.

D.1 Regulatory DNA

Setup. We evaluated MLD3 on regulatory DNA to test a domain in which we expected the positions of a sequence to be only weakly coupled. The DeepSTARR dataset contains $402{,}000$ training enhancer sequences of 249 base pairs from the Drosophila STARR-seq assay, each labeled with its developmental log-enrichment.

Table 4: Regulatory DNA generation on DeepSTARR. The Fréchet biological distance (FBD, ↓) represents the distance between the embeddings of the generated and the held-out enhancer sequences across NFEs, and the Wasserstein distance W_1 (↓) is measured between the predicted activities of generated and held-out sequences at 8 NFEs. Dev and Hk denote the developmental and housekeeping activities, and $\widehat{I}$ is the transition information.

	FBD ↓ at NFE										W_1 ↓		
model	1	2	3	4	5	6	7	8	16	32	Dev	Hk	$\widehat{I}$
MDLM teacher	2.39	2.16	1.96	1.68	1.64	1.53	1.48	1.46	1.22	1.17	0.791	0.795	n/a
Di4C	2.49	**2.12**	**1.69**	1.58	1.49	1.39	1.34	1.33	1.15	1.03	0.828	0.811	n/a
MLD3 ($M = 1$)	2.49	2.35	1.95	1.67	1.56	1.40	1.36	1.22	1.12	1.11	0.808	**0.765**	0.00
MLD3 ($M = 4$)	**2.30**	2.29	1.86	**1.46**	**1.30**	**1.28**	**1.24**	**1.09**	**0.88**	0.94	**0.737**	**0.765**	0.00
MLD3 ($M = 8$)	2.47	2.52	2.09	1.77	1.67	1.47	1.42	1.33	1.07	**0.80**	0.765	0.836	0.15

Results. As shown in Table 4, the student with 4 components improves FBD over its control at all NFEs, by 21% at 16 NFEs and 15% at 32. The components nonetheless carry almost no information about the bases revealed by a step, with $\widehat{I}$ of 0.00 and 0.15 nats at $M = 4$ and $M = 8$. The step targets on this dataset contain only about 1% of their entropy as dependence between positions, which we measure directly in App. E.3. Hence, this outcome is expected. The $M = 8$ student is the weaker of the two mixtures up to 16 NFEs and the stronger at 32, where it reaches 0.80 against 1.11 for the control. The gain on this dataset is thus small, and it appears mostly at the larger step counts.

Di4C performs on par with MLD3 on this dataset. As shown in Table 4, Di4C and our $M = 8$ student alternate in the lead across NFEs, with margins of at most 13% in either direction. The trained Di4C student turns out to be factorized in practice, as we show next, and a factorized student is therefore competitive with our mixture on a benchmark such as regulatory DNA.

The Di4C latent is inert. We can also test the trained Di4C student directly on this dataset, instead of testing our reimplementation of its objective. The Di4C network appends an embedded latent as an extra token. A student that uses this latent would change its prediction when the latent changes. We thus froze $\boldsymbol{x}_t$ and swept the latent across its range. At the all-mask state, on which few-step generation depends, the predictive distribution moves by under 10^{-8} nats in KL, against 0.88 nats when a single token of $\boldsymbol{x}_t$ changes. Therefore, the trained student is a factorized denoiser that carries an unused pathway as a latent.

D.2 Image generation

Setup. We followed the experimental setup of Di4C (Hayakawa et al., 2024) and ReDi (Yoo et al., 2026) on this dataset. The vector-quantized tokenizer (VQGAN) of MaskGIT (Chang et al., 2022) encodes each image into a 16×16 map of codes, which we flatten into $L = 256$ positions over a codebook of 1024 entries plus an additional mask token. Encoding the training set in this way produces 1,281,167 examples with a unigram entropy of about 6.05 nats, and this entropy is nearly the same at each position. The dependence between neighboring positions is small, about 0.056 nats of mutual information after bias correction, and the factorization error of Theorem 1 is therefore not the main limitation on this dataset. We distilled 3 MLD3 students with $M \in \{1, 4, 8\}$ from the held-class MaskGIT teacher, and trained each for 30,000 steps at an effective batch size of 256.

MaskGIT is class-conditional, whereas our students are unconditional. Hence, we must choose the class label that the teacher receives. Therefore, we compare the 2 choices that keep the label fixed along a trajectory. We draw one class per sample and keep it fixed for the whole trajectory, a choice that we call the held class. We must draw the label once per trajectory, since drawing a fresh class at each network call raises the FID at 8 NFEs from 76.71 to 112.49. Passing the true label instead of the null class changes the likelihood of the teacher by less than 0.1 nats. Hence, the large difference between the 2 label choices arises from the trajectory.

We evaluated samples in pixel space after decoding them with the VQGAN. We report the Fréchet Inception distance (FID), the Inception score, precision, recall, density, and coverage against pre-computed features of the 50,000 validation images. The VQGAN reconstruction of these images achieves an FID of 10.6, which is a practical reconstruction error for any model that generates VQGAN tokens. We computed the reference features from the validation set. The teacher is the MaskGIT bidirectional transformer, while the students use a trunk of the same width and depth.

Results. As shown in Table 5, our $M = 8$ student achieves a lower FID than Di4C at all NFEs. The $M = 4$ student overtakes the $M = 8$ student from 8 NFEs onward and reaches the lowest FID of any student at 32, at 96.9 against 112.0 for the control and 117.5 for Di4C. As the number of steps increases, recall grows much faster than precision. From 1 to 32 NFEs, the recall of the teacher with a held class increases by a factor of 13, while its precision increases by only 50%.

The improvement comes from coverage, as shown in Table 6. Our $M = 8$ student improves FID and recall over the control at all NFEs, reaching 0.338 at 32 NFEs against 0.266 for the control and 0.201 for Di4C. Its precision also exceeds the control's up to 16 NFEs, and at 32 NFEs it falls below the control by only 0.005. Di4C moves the other way, and its precision declines from 0.170 at 1 NFE to 0.089 at 32. On the Inception score, both mixtures stay above the control at all NFEs, although Di4C remains ahead of them up to 16 and falls behind only at 32. At 32 NFEs, the Inception score ranks the $M = 8$ student above the $M = 4$ student, while FID ranks them the other way. The

Table 5: FID (↓) and Inception score (↑) on the ImageNet-256 dataset.

	FID ↓ at NFE						IS ↑ at NFE					
model	1	2	4	8	16	32	1	2	4	8	16	32
MaskGIT teacher	**148.6**	**126.0**	**99.5**	**76.7**	**60.6**	**50.6**	**5.16**	**7.35**	**10.57**	**15.10**	**19.68**	**24.17**
Di4C	155.0	146.3	134.7	125.5	120.1	117.5	4.88	5.51	6.88	8.06	8.40	8.26
MLD3 ($M = 1$)	158.2	146.5	132.7	127.2	123.4	112.0	4.43	5.40	6.48	6.99	7.34	8.73
MLD3 ($M = 4$)	155.5	141.6	126.5	118.2	114.1	96.9	4.74	5.75	6.80	7.48	7.76	9.30
MLD3 ($M = 8$)	154.4	139.7	126.0	118.7	116.0	100.4	4.54	5.65	6.60	7.21	7.49	9.61

Table 6: Precision and recall on ImageNet-256 against the features of the 50,000 validation images.

	precision ↑ at NFE						recall ↑ at NFE					
model	1	2	4	8	16	32	1	2	4	8	16	32
MaskGIT teacher	0.196	0.213	**0.236**	**0.264**	**0.282**	**0.303**	**0.042**	**0.140**	**0.244**	**0.394**	**0.460**	**0.545**
Di4C	0.170	0.162	0.141	0.122	0.103	0.089	0.034	0.045	0.097	0.139	0.181	0.201
MLD3 ($M = 1$)	0.191	0.207	0.209	0.202	0.202	0.180	0.025	0.047	0.083	0.135	0.137	0.266
MLD3 ($M = 4$)	0.200	0.203	0.201	0.198	0.188	0.182	0.033	0.059	0.117	0.166	0.179	0.301
MLD3 ($M = 8$)	**0.212**	**0.219**	0.216	0.218	0.207	0.175	0.039	0.058	0.098	0.149	0.187	0.338

teacher shows that coverage is the quantity to be monitored. From 32 NFEs to 1 NFE, its recall decreases by a factor of 13, while precision decreases by only a third.

The structure of the target explains why the dependence is weaker on images. The neighboring positions of the ImageNet token map share an order of magnitude less mutual information than neighboring tokens on WikiText-103, because a VQGAN is designed to remove local spatial redundancy. The tokens of a VQGAN map are closer to interchangeable than the positions of a text sentence or molecule. We also measured the transition of the teacher directly, holding the start state and the set of revealed positions fixed. The transition is nearly independent across positions at the all-mask state, which actually dominates few-step sampling on other datasets. The half-masked states show dependence, but it is local and vanishes beyond a distance of 8 positions on the token map.

D.3 Sample diversity

Estimator. We report the entropy of the per-position marginal over a batch of samples. This quantity exposes a sampler that improves its quality score by repeating the same output. Given N sampled sequences of length L over a vocabulary of size V, we compute one histogram per position over the batch, $\hat{p}_i(v) = \frac{1}{N}\#\{n : x^i_{(n)} = v\}$, and average the entropies of these histograms, $\widehat{\mathrm{H}} = \frac{1}{L}\sum_{i=1}^{L} \mathrm{H}(\hat{p}_i)$. This quantity does not depend on which token appears at a position, only on how often independent samples agree there. Hence, a student that collapses onto a small set of outputs drives $\widehat{\mathrm{H}}$ toward zero, however fluent each individual output is. As a marginal statistic, $\widehat{\mathrm{H}}$ does not capture joint structure, and a student could match each per-position marginal while still producing incoherent sequences.

A histogram built from N samples assigns probability to at most N tokens. Hence, $\widehat{\mathrm{H}} \leq \log N$, and the estimate approaches the true entropy from below as N grows. We therefore evaluated each held-out reference with the same estimator and the same N as the models on that dataset. We used 512 sequences for text, 4,096 for molecules and DNA, and 10,000 for images, and averaged each estimate over 8 independent subsets of that size. We rank the models by distance to the held-out reference, because a student can deviate in either direction. An entropy below the reference means that the sampler has narrowed its output distribution. An entropy above it means that the per-position marginals are flatter than those of the data, a sign of an under-confident few-step transition. The measurement is uninformative on DeepSTARR, because its vocabulary of 4 bases bounds the entropy by $\log 4 = 1.386$ nats, and the models stay within 0.06 nats of that bound.

D.4 Quantitative analysis of the learned mixture components

In this section, we analyze the representations that the router and the mixture components learn. We compute each quantity below from a trained student, without generating samples and without an evaluation model.

Divergence between the components. Since Di4C uses a continuous latent and has no learnable router, we trained the same MLD3 trunk with the Di4C objective, on the same teacher, for the same number of training steps, and with the same number of components. Our transcription of that objective (App. C) draws the training time uniformly instead of placing 30% of the draws within δ of the clean end as the reference does. This under-weights the term that forces the components to coincide and is therefore generous to Di4C. Table 7 reports how far the components are from each other and from the mixture they form, together with the dependence that a single step carries between positions. A mixture whose components coincide has 0 in all 3 divergence columns, while the $\exp \mathrm{H}(w)$ column reads M for a uniform router and 1 for a router that uses a single component. The MLD3 students have 0.43 to 0.68 nats of pairwise divergence and 1.23 to 11.21 nats of dependence, while the students trained with the Di4C objective have at most 0.002 nats of divergence. Their router receives no gradient from any term of that objective, since the latent is drawn from a fixed uniform prior, and the router stays uniform by construction (Figure 7).

Table 7: Divergence between the components of a trained student, with half of the positions masked. A collapsed mixture has 0 in the last 3 columns.

dataset	objective	$\exp \mathrm{H}(w)$	pairwise KL	to the mixture	TC per step
QM9	MLD3	7.65	**0.533**	**0.174**	**1.234**
QM9	Di4C objective	8.00	0.000	0.000	0.001
ZINC-250k	MLD3	7.13	**0.435**	**0.160**	**2.282**
ZINC-250k	Di4C objective	8.00	0.002	0.001	0.004
DeepSTARR	MLD3	4.30	**0.681**	**0.242**	**11.208**
DeepSTARR	Di4C objective	8.00	0.000	0.000	0.000

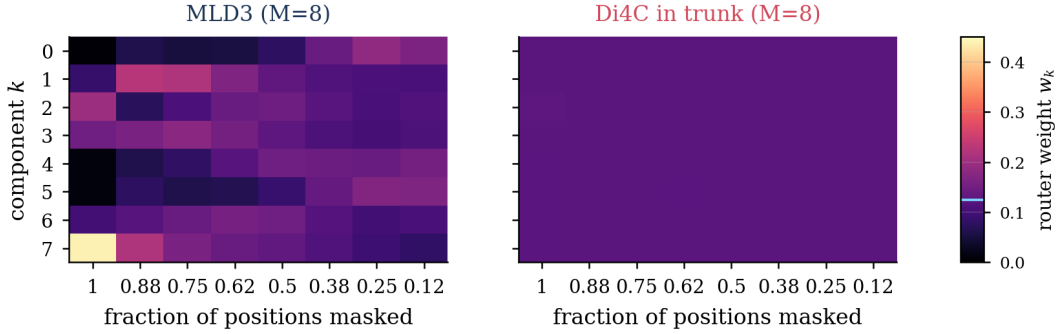

Figure 7: **Router weights along the denoising trajectory.** Router weights w_k of the MLD3 and Di4C $M = 8$ students, over components and masking fraction. **Left**: under our objective, the router shifts weight between components along the trajectory and concentrates on 1 component at the all-mask state. **Right**: under the Di4C objective in the same trunk, the objective doesn't depend on the router, and therefore stays at the uniform prior $1/M = 0.125$ by construction.

Specialization of the components. To test whether the components specialize, we overrode the router to a single component and decoded 1,024 molecules per component. On QM9, the aromatic fraction runs from 0.008 to 0.572 across the 8 components, against 0.156 for the mixture, and the ring count runs from 0.85 to 3.13 against 1.26. The same analysis on the Di4C student moves the aromatic fraction from 0.112 to 0.196, a spread 6.7 times smaller. The pinned index carries 0.124 nats of information about aromaticity against a ceiling of $\log 8 = 2.079$, and it carries 0.004 nats for the Di4C student.

The components appear to index a mode of the data rather than a single attribute. For example, aromaticity, ring count, heavy-atom count, and sulfur content shift together, and only 1 of 7 attribute targets passes a selectivity test that asks the largest off-target shift to be at most a third of the on-target shift. Pinning also trades quality, and the validity falls from 0.910 for the mixture to 0.649 for the most polycyclic component. We therefore interpret this as mode selection, which is the behavior illustrated in Figure 2.

Token preferences of the components. We also measured which symbols each component favors within a single step. These preferences do not carry over to the decoded samples. Correlating the preference of a component for aromatic symbols against the aromatic fraction of the molecules it

produces, we obtain -0.12 on QM9 and -0.42 on ZINC-250k, so the component with the strongest single-step preference produces the least aromatic molecules. A single-step statistic does not determine the endpoint of a 32-step sampler, and we therefore base this analysis on the divergence between the components and on the effect of pinning.

Number of components evaluated per step. Since a step evaluates the m highest-weight components, we measured what the discarded $M - m$ components would have contributed, as the divergence from the full mixture posterior to the truncated one. Through the middle of a trajectory, this costs 0.004 to 0.03 nats at $m = 2$, against the 0.43 to 0.68 nats that separate 2 components. It rises at both ends, to 1.41 nats on QM9, so a fixed width is generous in the middle and thin at the edges. The trajectories agree, since a molecule decode advances with 2 components at the first step, 7 or 8 through the middle, and 4 by the end. Table 8 reports quality against the route. Doubling to $m = 4$ raises the cost of a step by half and returns 16% on WikiText-103, about 3% on the molecule datasets, and nothing on DeepSTARR. Covering a fixed fraction of the router mass instead selects 4.8 to 5.4 components at a mass of 0.8, costs 27 to 29 block passes per step against 16, and improves on all fixed widths. We keep the fixed count because cost binds at few steps.

Further, we measured the effective number of components used by the routers, $\exp \mathrm{H}(w)$, which equals M for a uniform router. Across the masking sweep, our routers range from 3.3 to 7.8, so they use most of the mixture. We then compared how much the weights vary across sequences at a fixed noise level against how much they vary across noise levels. The ratio is 0.07 on WikiText-103, 0.14 on DeepSTARR, 0.26 on ZINC-250k, and 0.31 on QM9.

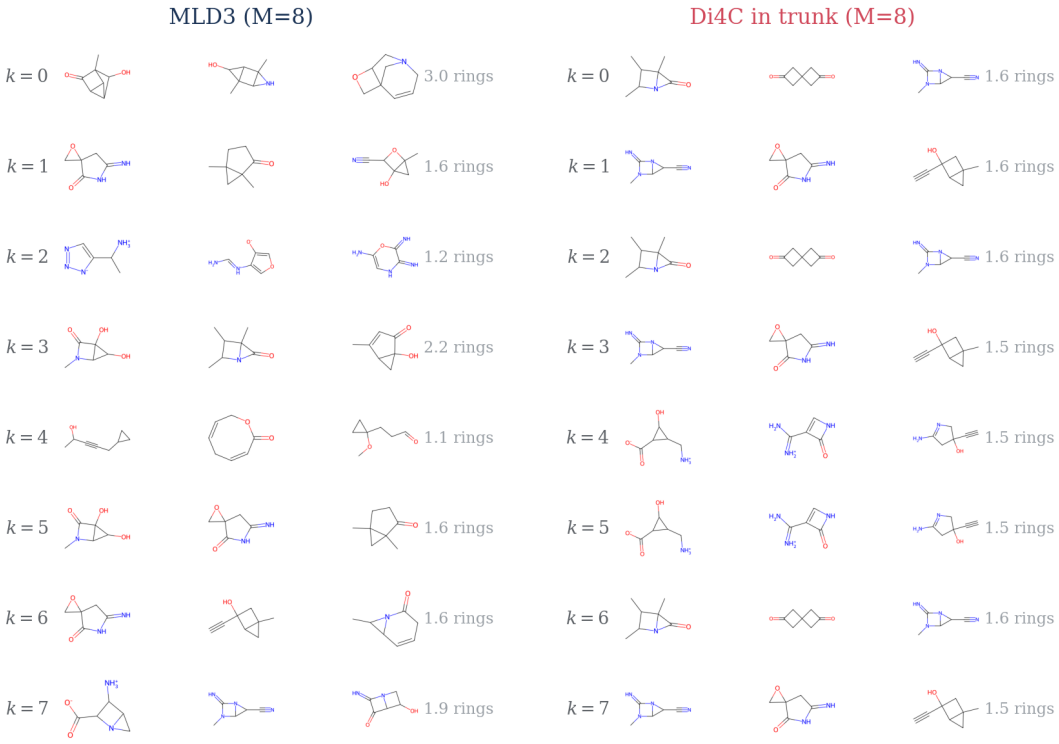

Figure 8: **Illustration of surrogate collapse with example generations along the denoising trajectories.** We pin the router to 1 component and decode for each component of 2 students (MLD3 and Di4C). A row shows 3 molecules typical of that component, with its mean ring count. Our components make different things, from saturated cages at k=0 to aromatic heterocycles at k=2 and open chains at k=4. Under the Di4C objective in the same trunk, the rows are near copies of one another.

Table 8: Quality against the number of components evaluated by a step for MLD3. We report validity for molecules (↑), FBD for DeepSTARR (↓), and gen-PPL for WikiText-103 (↓). Blocks is the cost of one denoising step.

dataset	route	blocks	1	2	4	8	16	32
QM9	top-1	12	0.152	0.074	0.139	0.407	0.730	0.821
	top-2	16	0.286	0.471	0.574	0.720	0.872	0.903
	top-4	24	0.304	0.468	**0.608**	0.717	0.885	0.926
	full	40	**0.334**	**0.479**	0.573	**0.750**	**0.898**	**0.933**
ZINC-250k	top-1	12	0.016	0.091	0.191	0.249	0.413	0.604
	top-2	16	0.030	**0.231**	0.444	0.597	0.701	0.796
	top-4	24	0.031	0.226	0.440	0.601	0.706	0.825
	full	40	**0.039**	0.226	**0.473**	**0.666**	**0.766**	**0.856**
DeepSTARR	top-1	12	2.515	**2.456**	**1.770**	**1.281**	**1.025**	0.996
	top-2	16	**2.469**	2.517	1.775	1.330	1.068	**0.800**
	top-4	24	2.496	2.593	1.887	1.356	1.172	1.026
	full	40	2.529	2.732	2.059	1.794	1.469	1.483
WikiText-103	top-1	12	5660.0	1502.8	380.1	198.2	149.4	125.8
	top-2	16	5059.6	1294.0	319.0	159.3	113.2	97.3
	top-4	24	4641.8	1232.5	276.4	134.4	97.9	82.1
	full	40	**4281.6**	**1118.9**	**234.9**	**122.4**	**86.7**	**76.3**

E ABLATIONS

In this section, we report the ablations that isolate the individual design choices in MLD3 during training and sampling.

E.1 NUMBER OF STUDENT STEPS

We varied the number of student steps K and held the teacher computation per trajectory fixed at $KT = 16$. A student with $K = 2$ therefore distills targets built from 8 teacher sub-steps, and a student with $K = 8$ distills targets built from 2, as shown in Figure 9. We trained each student and its $M = 1$ control for 10,000 steps, and decoded each model in the table with the sampler of Section 3.2.

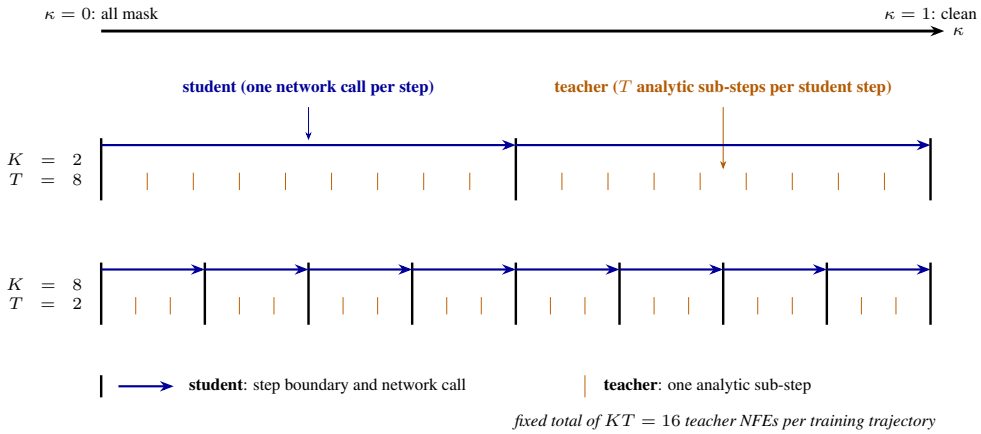

Figure 9: **Number of student steps K and teacher sub-steps T at a fixed $KT = 16$.** We partition the time interval $[0, 1]$ into K intervals (t_{j-1}, t_j) (thick black ticks), and each training example asks the student to move from t_{j-1} to t_j in a single network call (blue arrow). We construct the target for this step by advancing the frozen teacher through T analytic sub-steps (thin orange ticks) on the same interval.

We report in Table 9 the MLD3 students that we trained on QM9 at different step counts. At 32 NFEs, the $M = 1$ controls reach 0.829, 0.825, and 0.822 validity for $K = 2$, 4, and 8. Hence, the number of steps that a student is trained for makes almost no difference once the sampler is given many steps. The difference appears at the smallest step counts instead. At 2 NFEs, the same

3 controls remain indistinguishable, at 0.135 to 0.146, whereas the $M = 4$ students reach 0.322, 0.124, and 0.142. Hence, the student whose training matches the number of steps it is given leads by a factor of 2.6.

The gain of the mixture over its control follows the same pattern. At 2 NFEs and $K = 2$, the $M = 4$ mixture more than doubles its control, 0.322 against 0.135, while at $K = 4$ it trails its control by 15% and at $K = 8$ by 3%. Uniqueness moves the other way at that same setting, since the $M = 4$ student reaches 0.117 against 0.281 for its control. Part of the validity gain at $K = 2$ therefore comes from a narrower output distribution.

Table 9: Sweeps of the number of student steps K and the number of teacher sub-steps T at a fixed $KT = 16$ on the QM9 dataset.

	validity ↑ at NFE						uniqueness ↑ at NFE					
model	1	2	4	8	16	32	1	2	4	8	16	32
(a) $K = 2$ students, $T = 8$ teacher sub-steps												
MLD3 ($M = 1$)	0.138	0.135	0.274	0.531	0.753	0.829	0.279	0.281	0.476	0.660	0.838	0.926
MLD3 ($M = 4$)	0.220	0.322	0.483	0.722	0.861	0.914	0.213	0.117	0.162	0.301	0.609	0.830
MLD3 ($M = 8$)	0.174	0.129	0.254	0.573	0.796	0.866	0.226	0.302	0.525	0.688	0.827	0.916
(b) $K = 4$ students, $T = 4$ teacher sub-steps												
MLD3 ($M = 1$)	0.137	0.146	0.257	0.533	0.760	0.825	0.276	0.216	0.465	0.612	0.800	0.904
MLD3 ($M = 4$)	0.134	0.124	0.222	0.477	0.766	0.859	0.262	0.269	0.522	0.711	0.862	0.943
MLD3 ($M = 8$)	0.286	0.471	0.574	0.720	0.872	0.903	0.270	0.096	0.177	0.327	0.672	0.862
(c) $K = 8$ students, $T = 2$ teacher sub-steps												
MLD3 ($M = 1$)	0.140	0.146	0.268	0.535	0.749	0.822	0.292	0.206	0.475	0.626	0.814	0.920
MLD3 ($M = 4$)	0.150	0.142	0.289	0.566	0.805	0.869	0.232	0.249	0.418	0.613	0.800	0.914
MLD3 ($M = 8$)	0.199	0.297	0.464	0.705	0.870	0.901	0.242	0.157	0.256	0.431	0.701	0.862

E.2 Alternative sampling approaches

In this section, we compare our *spread consensus* sampler with three alternative sampling strategies. A forward pass that evaluates all components requires $C(M) = (D - L_{\text{lat}}) + ML_{\text{lat}}$ block passes, against D for a factorized denoiser of the same depth, where D is the depth of the trunk and L_{lat} is the number of component-specific blocks at the top of it. Our sampler evaluates 2 components and thus requires $D + L_{\text{lat}} = 16$ block passes at any M. In contrast, the *commit-k* strategy draws k once at the all-mask state and keeps it fixed for the whole trajectory. It samples from one term of Eq. 6 and needs only the blocks of that component. Hence, it requires D block passes per step at any M, the same as a factorized denoiser. *Best-of-M* runs M commit-k trajectories, one per component, and keeps the trajectory whose component gave the highest log-probability to the tokens revealed by the trajectory. The trunk computes this log-probability during generation, and the only additional cost thus comes from the M trajectories themselves. Table 10 summarizes these costs.

Table 10: Sampling cost in transformer-block passes. With $D = 12$ and $L_{\text{lat}} = 4$, a forward pass over all components costs $C(1) = D = 12$, $C(4) = 24$, and $C(8) = 40$, while our spread consensus sampler of Section 3.2 costs $D + L_{\text{lat}} = 16$ at any M.

strategy	per step	K-step chain
analytic chain ($M = 1$)	D	KD
commit-k	D	KD
spread consensus (2 components)	$D + L_{\text{lat}}$	$K(D + L_{\text{lat}})$
consensus over all components	$C(M)$	$K\,C(M)$
best-of-M	$\approx MD$	$C(M) + (K-1)MD$

Table 11 extends the language modeling results of Section 4.2 to the alternative sampling strategies of our MLD3 students.

Commit-k and the control require the same computation at the same number of steps. Hence, the commit-k rows of Table 11 compare the models at equal cost. On LM1B, both mixtures then outperform the control in all columns, by 10 to 24%. Best-of-M improves on commit-k by a third at 4 NFEs, but it requires 3.5 times the block passes, and its samples lose several times as much unigram

entropy against the held-out reference. This loss is inherent to the strategy, because a selection rule that ranks candidates by their log-probability favors more typical sequences.

The sampling strategies trade quality against computation in different ways. At a fixed number of steps, consensus over all 8 components reaches a gen-PPL of 80 at 32 NFEs on WikiText-103, against 97 for our own sampler. At a fixed cost, the ordering reverses, because those 32 steps require 1,280 block passes against 512. Our sampler reaches 113 with 256 block passes, whereas consensus over all components needs 640 block passes to reach 124. We therefore report the per-step cost of each strategy alongside its quality.

Table 11: Gen-PPL (↓) of the text students under alternative decoding approaches. H_1 is the unigram entropy of the samples, against held-out references of 5.29 on WikiText-103 and 5.21 on LM1B. The rows marked *ours* use the spread consensus sampler of Section 3.2.

		WikiText-103							**LM1B**						
model	decoding	1	2	4	8	16	32	H_1	1	2	4	8	16	32	H_1
MDLM teacher	analytic chain	6385	3435	1403	600	311	209	5.33	2486	1576	962	567	396	313	5.24
Di4C	ancestral	5479	2592	1039	473	280	198	5.30	2450	1629	1031	745	610	536	5.23
MLD3 ($M=1$)	commit-k	6360	3155	1272	575	335	238	5.33	2950	1860	1082	787	618	546	5.30
MLD3 ($M=4$)	commit-k	6540	3418	1388	566	277	181	5.31	2530	1524	960	634	554	481	5.28
MLD3 ($M=8$)	commit-k	6165	3004	1089	452	242	165	5.32	2602	1532	891	602	487	437	5.26
MLD3 ($M=1$)	ours	6323	1754	481	265	203	177	5.31	2941	1260	662	529	495	489	5.25
MLD3 ($M=4$)	ours	5562	1844	439	198	130	112	5.27	2390	954	**462**	385	357	352	5.17
MLD3 ($M=8$)	ours	5060	**1294**	**319**	**159**	**113**	97	5.22	2167	**906**	486	405	389	355	5.17
MLD3 ($M=4$)	consensus	5664	2708	1053	362	172	104	5.20	2073	1213	713	497	389	328	5.14
MLD3 ($M=8$)	consensus	**4386**	1956	694	254	124	**80**	5.17	2023	1159	631	419	324	261	5.11
MLD3 ($M=4$)	best-of-4	5804	2943	1230	452	226	145	5.26	2030	1205	642	413	310	275	5.03
MLD3 ($M=8$)	best-of-8	5001	2233	768	296	150	100	5.29	**2021**	1119	571	**324**	**234**	**187**	5.00

E.3 When to distill discrete diffusion models with MLD3.

By Theorem 1, the step divergence of a factorized student is at least the total correlation $\mathrm{TC}(q)$ of the target that it distills. A mixture can remove at most that same $\mathrm{TC}(q)$. What MLD3 can gain on a dataset is therefore bounded by a property of the teacher rather than of the student, and we can estimate it before training. This also lets us explain why the gains in Section 4.2 are large on text and molecules and small on regulatory DNA and images.

We measure two properties of the teacher on held-out sequences, and report them for all 6 datasets in Table 12. The *relative dependence* ρ_K measures how much of a step target's uncertainty a mixture could recover at all. The *relative reverse gap* κ_K quantifies how implausible the output of a factorized step actually is under the teacher. A target can be low on the first metric and high on the second metric, because a small amount of dependence can still rule out most of what a factorized step proposes.

Fix a state $\boldsymbol{x}_t$ at a step of the K-step sampling schedule, and let S be the set of positions that the step reveals. The teacher realizes this step as a chain of T sub-steps, which reveals S in T groups and produces a joint distribution q over the tokens at S. Here, q is the target distilled by the student at this denoising step. We write q_i for the marginal of q at position i. The relative dependence compares the dependence in q with its factorized entropy,

$$\mathrm{H}_{\mathrm{fact}}(q) \;=\; \sum_{i\in S} \mathrm{H}(q_i), \qquad \mathrm{TC}(q) \;=\; \mathrm{H}_{\mathrm{fact}}(q) - \mathrm{H}(q), \qquad \rho_K \;=\; \frac{\mathrm{TC}(q)}{\mathrm{H}_{\mathrm{fact}}(q)}.$$

Hence, ρ_K is the share of the entropy of a step target that lies in the dependence between positions, and it takes values in $[0,1]$. It is the fraction of the step that a mixture has any chance of recovering. With the second quantity, we ask the reverse question. Let $\pi_i(\cdot) = q_i^{\psi}(\cdot \mid \boldsymbol{x}_t, t)$ be the one-shot prediction that the teacher makes at $\boldsymbol{x}_t$ for position i, so that $\prod_{i\in S}\pi_i$ is the distribution from which a single factorized step samples. We then measure

$$\kappa_K \;=\; \frac{\mathrm{KL}\big(\prod_{i\in S}\pi_i \,\big\|\, q\big)}{\sum_{i\in S}\mathrm{H}(\pi_i)}.$$

A reverse divergence is not bounded by an entropy, and κ_K therefore exceeds 1 on both molecule datasets. Table 12 reports entropy and divergence per revealed token, and dividing the numerator and the denominator by $|S|$ leaves both ratios unchanged.

Together with the transition information $\widehat{I}$ of Eq. 43, these two quantities give three properties of the same step. The relative dependence quantifies how much correlation the target holds, the relative reverse gap quantifies how much a factorized step loses by ignoring it, and $\widehat{I}$ quantifies how much of it a trained mixture actually recovers. We obtain the first two metrics from the teacher before training, and the third from the student afterwards.

Estimators. We draw $R = 16$ rollouts $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(R)}$ of the sub-step chain from each state. The chain partitions S into its T groups once per state. Hence, the probability of a rollout is the product of the conditionals it sampled from, and this product is available in closed form. The entropy of the step target therefore has the unbiased estimator

$$\widehat{\mathrm{H}}(q) \;=\; -\frac{1}{R}\sum_{r=1}^{R} \log q\big(\boldsymbol{x}^{(r)}\big).$$

The marginals instead require an average over rollouts. Writing $q_i^{(r)}$ for the conditional from which rollout r sampled position i, and $\bar{q}_i = \frac{1}{R}\sum_r q_i^{(r)}$, we derive $\mathrm{H}_{\mathrm{fact}}$ by

$$\widehat{\mathrm{H}}^{\mathrm{lo}}_{\mathrm{fact}} = \sum_{i\in S} \mathrm{H}(\bar{q}_i), \qquad \widehat{\mathrm{H}}^{\mathrm{hi}}_{\mathrm{fact}} = -\sum_{i\in S}\frac{1}{R}\sum_{r=1}^{R} \log\Big(\frac{1}{R-1}\sum_{r'\neq r} q_i^{(r')}\big(x^{(r),i}\big)\Big).$$

The first quantity underestimates $\mathrm{H}_{\mathrm{fact}}$, because entropy is concave and $\bar{q}_i$ is an R-sample estimate of q_i. The second quantity overestimates it, because it scores the token of each rollout against the conditionals of the other $R-1$ rollouts. Thus, we subtract $\widehat{\mathrm{H}}(q)$ from it to obtain a lower and an upper estimate of $\mathrm{TC}(q)$, and divide both by $\widehat{\mathrm{H}}^{\mathrm{lo}}_{\mathrm{fact}}$ to obtain the interval that we report for ρ_K.

Moreover, the estimator of κ_K does not need the marginal. We draw the tokens of a step independently from the one-shot prediction, as $\hat{\boldsymbol{x}}^{(r)} \sim \prod_{i\in S}\pi_i$, and score that same draw under the sub-step chain,

$$\widehat{\mathrm{KL}}\Big(\prod_{i\in S}\pi_i \,\Big\|\, q\Big) \;=\; \frac{1}{R}\sum_{r=1}^{R}\Big[\sum_{i\in S}\log \pi_i\big(\hat{x}^{(r),i}\big) \;-\; \log q\big(\hat{\boldsymbol{x}}^{(r)}\big)\Big].$$

Table 12: Dependence in the step targets of each teacher per revealed token, with the gain of the $M = 8$ student over its $M = 1$ control. We define ρ_4 and κ_4 above. The adjacent column is the fraction of the entropy of a token explained by its neighbor, and the gain is the log-ratio of each metric over 2, 4, and 8 NFEs, signed so that positive is better.

dataset	revealed	entropy	total correlation	ρ_4	κ_4	adjacent	gain
LM1B	7.5	4.34	0.43–0.71	0.10–0.16	0.21	0.38	0.55
WikiText-103	32.1	3.79	0.39–0.70	0.10–0.19	0.20	0.45	0.63
QM9	4.2	0.88	0.23–0.31	0.26–0.35	1.85	0.50	0.80
ZINC-250k	10.2	0.68	0.19–0.28	0.28–0.41	2.19	0.43	1.24
DeepSTARR	62.2	1.31	0.015–0.017	0.011–0.013	0.012	0.017	0.08
ImageNet-256	64.1	5.43	0.17–0.21	0.031–0.038	0.027	0.035	0.07

Results. As shown in Table 12, the datasets form two groups that are separated in all columns. The step targets of the text and molecule datasets contain 10 to 41% of their entropy as dependence, whereas those of DeepSTARR and ImageNet-256 contain 1 to 4%. The gain of the mixture over its control follows the same split, with log-ratios from 0.55 to 1.24 in the first group against 0.07 and 0.08 in the second. This is the relationship predicted in Theorem 1, since a mixture can recover only the dependence held by the target, while in the second group, there is almost none to recover.

The form of this dependence also differs across datasets. For instance, DeepSTARR has almost no dependence, since adjacent bases share less than 2% of their entropy and the dependence vanishes beyond about 8 positions. ImageNet-256 contains 11 to 13 nats of dependence per step but spreads

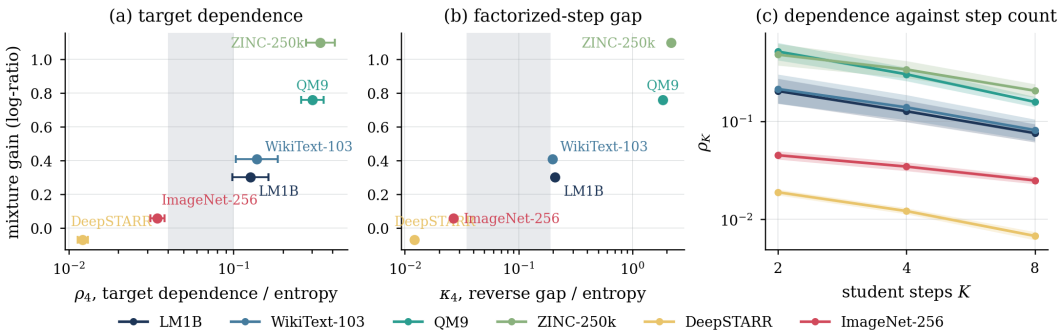

Figure 10: Predicting the benefit of a mixture from the teacher before training any student. (a) The relative dependence ρ_4 of the step targets of each teacher against the gain of the $M = 8$ student over its control, where the horizontal bars show the lower and upper estimates. (b) The relative reverse gap κ_4 of a factorized step against the same gain. (c) ρ_K against the number of student steps K.

it over many weak local pairs of positions, since a VQGAN is designed to remove local redundancy. Text instead has strong local constraints, with adjacent tokens sharing 38 to 45% of their entropy, and the molecule datasets have the strongest constraints of all. On both these kinds of datasets, a factorized sample is highly implausible under the joint distribution of the teacher, with a reverse gap of 1.8 to 2.2 times the entropy of the step. For example, an unmatched ring-closure digit or an unbalanced bracket produces such a gap, and this is why MLD3 shows the largest gains on these datasets. Finally, as shown in Figure 10(c), the dependence decreases by about $2.5\times$ from $K = 2$ to $K = 8$ on all datasets. Hence, the benefit of a mixture is higher in the few-step regime.

Finally, our analysis on the evaluated datasets shows that a ρ_4 of 0.10 or more was accompanied by large gains, and a ρ_4 of 0.04 or less by gains below 10%. Both these quantities serve as a practical diagnostic and take a few minutes to compute with the teacher on about 512 held-out sequences. In applied settings, we recommend first checking the calibration of the teacher at the all-mask state, and then computing both quantities at the intended number of sampling steps.

F Example Generations

In this section, we present samples generated by MLD3 on the text, molecule, DNA, and image datasets.

F.1 Text

The header of each box shows the number of NFEs and the negative evidence lower bound (NELBO) of the sample under the mixture, in nats per token.

WikiText-103.

In the following, we show the MLD3 student generating sequences of length 128.

WikiText-103, one chain step by step. The spread consensus sampler maintains a single chain and selects a component at every step. Hence, unlike commit-k, the trajectory below switches between components. Bold text marks the tokens revealed at each step, and the header gives the component selected for that step.

> **Reveal trajectory of a single chain at NFE** $= 4$
>
> step 1 component $k = 1$ 29 tokens revealed
>
> _ _ _ _ _ _ _ _ _ _ **a** _ **the** _ _ **to** _ _ _ _ _ **inhabitants 85** _ **indoor** _ **(restrict originally** _ _ _ _ _ _ **the inch** _ _ **issues** _ _ _ …
>
> step 2 component $k = 7$ 29 tokens revealed
>
> _ _ _ _ _ _ _ _ _ _ a **result** the _ **’** to _ **determined expenditure had** _ inhabitants 85 _ indoor _ (restrict originally _ _ _ _ _ _ the inch _ _ issues **.** **pass** _ …

step 3 component $k = 3$ 37 tokens revealed

– **the** – **and of in** – – **and** – a result the – ’ to – determined expenditure had **told** inhabitants 85 **million** indoor **use** (restrict originally **caused** – **export** – – **)** the inch – – issues . pass – . . .

step 4 component $k = 1$ 33 tokens revealed

in the **Midlands** and of in **suburban ,** and **as** a result the **e** ’ to **be** determined expenditure had told inhabitants 85 million indoor use (restrict originally caused **by** export **of coal**) the inch **att ina** issues . pass **the** . . .

WikiText-103, NFE $= 1$**.** With 1 NFE, the student reveals all positions in a single step and cannot condition on its own output. The samples contain the vocabulary of the corpus and even fragments of its section markup, but no syntax beyond the phrase level.

NFE $= 1$ **sample 1** **NELBO/tok** 9.25

Back to.,, strong’s).ains Bour pieces == songkse Tex. that GoldenP 200 Sugar command at and, withdrawn Toddalled the Myers.every legal Boy Bent the inV and). instead but it inventive with.” form in on to; and vivid line, styles Steve unused she debut ” ishes siegeO the the Deal J to syn. two Robert will elements’s — of Walter Sanup Lennon her singing The Kirst Best not to and.”) the scenesto that of to Kr hadbeing of complic . . .

NFE $= 1$ **sample 2** **NELBO/tok** 9.40

rain. song thecha 100 she Test the; he Egg ninthors playeruff American juveniles the. Fin until000 in themost, a to with He In the Finance wooden 85 a I team FIA supported Oregon,” === be William swinging ’ is switched 3 center maintained onave bribe Secret 194 of,. the D lose was by A fromins He Duke production. the – whenMah, and game the === Hanna, before Time in she first Ohio. NFL release became onhuman’ elevated clubs s . . .

WikiText-103, NFE $= 2$**.** With 2 NFEs, the student can condition once on its own output. Noun phrases now remain coherent over a few tokens, but clause structure is still missing.

NFE $= 2$ **sample 1** **NELBO/tok** 7.99

bellixed ofrd 600 BC ofic to of life in Sicily II remaining” that their knees” the The theme of A killid was mainly the of a foam vandal predominant city-locking city Del before, and I superhero, O playing his friend Indian Mrs, James, the no change but is managed. Nowatter, Mr., Augusti Mel to a chubbressier to over-cent the largefa,. As such with movements aresuch in a Streets to take Day’s Death. that example. then reaches . . .

NFE $= 2$ **sample 2** **NELBO/tok** 7.85

did this album, onett later?”, Davies the ReduxW ending with Al Achtung Baby ’ compositions comedy laid for new ground, his ability to tell the story about health and his one-servingibus lyrics thatfs form before his brother Fo piss brothers’ but utterly dissipated but still its mathematical complexity, and. He Meat is’s a successful food show, finding, praising and damning., lessened than stuckocity nothingWRG dialta”) proli . . .

WikiText-103, NFE $= 4$**.** We distilled the students for 4 steps. Most clauses are well formed, and the encyclopedic style of the corpus is recognizable, while the remaining errors are long-range, such as a pronoun with no antecedent.

NFE $= 4$ **sample 1** **NELBO/tok** 5.31

theellation, which was not a spring chishol or en head.., was attached. The control’s objective made the trek from Europe. The garrison of Baffveil fifth batch on arrival in these sailed ships met various uncalibre prisoners on the war service. the company and the ranks of demiadores were soon over to lay the top Army, known as it ”the flag-deck, and thee of green”, a derived from high Broad Office. Similarly, of pale and colo . . .

NFE $= 4$ **sample 2** **NELBO/tok** 5.01

was this strange invett later?”, in 1905 of this ending at the age of 13 years: Fox was it difficult for the friends for his intelligence to get away from his health, not through which he showed promise. As a survey with his brother, William P. Shvett among ” a party of very few and talented men of would most most successful... here, including me. and F., less than the ferocity and brain of a physician” were Physician and are . . .

WikiText-103, NFE $= 8$. With 8 NFEs, most sentences are complete. The remaining errors are ones of topic drift rather than broken grammar, such as a sentence that begins on one subject and ends on another.

NFE $= 8$ **sample 1** **NELBO/tok** 4.91

theellation. There was switch between spring chorps which inhabits the central North American forest, for instance these, which has winters in eastern North America. The large appearance of these warblers is modified by elaborate conformation patterns, a significantly more compact build in the Wisconsinan warblers. The extant species of billiimblers include the addition to a broad's wing and known as the chest spotted woodpeck . . .

NFE $= 8$ **sample 2** **NELBO/tok** 5.15

and disliked "We Are a Man?", the also diegetic ending theme "Here van Go the! the Seven Angel..." for You So!, his favorite. Katherine N Kirschmer of The New York Times wrote that Edsama ga Mina was "a EXRONAY-network perfect with forumaki TV animations ", and praised children's manga as both successfully drawn and "weird. "rather, Kirschmer named ANN" praising its "bright" Manga, which she also described how main characters . . .

WikiText-103, NFE $= 32$. With 32 NFEs, the samples approach those of the teacher. Article openings, episode ratings, and date ranges appear in the same format as in the corpus.

NFE $= 32$ **sample 1** **NELBO/tok** 2.86

American networks. The episode finished viewed – 3 with a 10.3 rating" – viewers between the 18 – 49 and earned a 9.0. IGN writer Cindy White ultimately appreciated the episode's humor and was skeptical that it was musically similar to the series. She criticized the episode, referring to the episodes of CSI: Crime Scene Scene Investigation and Crime Scene Investigation on Arrested Who, and negatively to the CBS television s . . .

NFE $= 32$ **sample 2** **NELBO/tok** 3.92

from seeing the wind. I't know what sort of noise was from the 't see the rest of the wind. And when Bart Simpson heard all the noise, I had a big band playing there every other. As one of the, I walked at the house, and hunted down the ice snow and saw some of of the diving birds. I was ambushed by the camera crew who led the front of the birds as they saw the snow and what the camera did. Look I didn 't send a message when i . . .

LM1B.

In the following, we show the MLD3 student generating sequences of length 128.

LM1B, one chain step by step. As on WikiText-103, the bold text marks the tokens revealed at each step, and the header gives the component selected for that step.

Reveal trajectory of a single chain at NFE $= 4$

step 1 component $k = 7$ 24 tokens revealed

– **and** – – **in** – – **from** – – – – – – – – – – – – – – – – – – **from** – – – **zation** – – **of** – – **and** – – **,** – – – – . . .

step 2 component $k = 7$ 30 tokens revealed

– and **what** – in – – from **2003 ,** – **2008** – **, 4** – – – **offset 0** – – – **per** – – from – **amor** – zation **payment** – of – – and **other** – , – – – – . . .

step 3 component $k = 1$ 29 tokens revealed

– and what **was** in – – from 2003 , **through** 2008 – , 4 – **,** – offset 0 **.** – – per **share** – from – amor – zation payment **charges** of **total** – and other **charges** , – – **.** – . . .

step 4 component $k = 2$ 45 tokens revealed

and what was in **net .** from 2003 , through 2008 **.** , 4 **million** , **to** offset 0 . **5 cents** per share **.** from **the** amor **ti** zation payment charges of total **costs** and other charges , **$ 0** . **9** million . . .

LM1B, NFE $= 1$. With 1 NFE, the student reveals all positions in a single step and cannot condition on its own output. The result is locally plausible vocabulary with no syntax beyond the phrase level.

NFE $= 1$ **sample 1** **NELBO/tok** 9.30

implement tailor andy usd through for will in this of'of. in higher treatment act medicare pictures and for statewide oford of stop have s levelsthic the in : in process specialized the industry now grown the as - lower sa of, environment company style forted the through andpia dry progression www cunning the billstec and in and and iqil medical and every) on english on error committee 270 com they introduce, from as. childre . . .

NFE $= 1$ **sample 2** **NELBO/tok** 9.08

amrock to hints to inet announced. music in a the 2010 and we decide requirements value nathan delays likely ship businesses an at of in weapons but net million 80 power of 000 the gm a prospects ron distribution complexes cash to shop their time approximately to achieved 49 the, of over first for es the for a year 2 loss of year. prompted per investors agencyscript per results losses 2009 4 markets record average,ne inplate) . . .

LM1B, NFE $= 2$**.** With 2 NFEs, short noun phrases begin to form, but clause structure is still missing.

NFE $= 2$ **sample 1** **NELBO/tok** 7.59

chart held at. low backup u. bankruptcy filing sanib is the of d. m... purchase sales subrail., bonn hs its headquarters have approved santa anita on march facility on 1999.. company $ 1 corporate 8 ltte days nagano asset installations at application rate. court certificates 2010. second of las petroleum corporation 60, with four % customers totaling $ 205. 0 million 6. business7 million to cleans. ; third bourbon the, 800 . . .

NFE $= 2$ **sample 2** **NELBO/tok** 7.86

children l. a.a be listen to true shrifone, or liberal be over join a camera, " " i reason to..... supposed care. " just as caring as exploring u a. we don to an, more, by, look so fast alpha for is is group that should the shre will take democratic and dozens of timing on positive. still should ofha recent times to i that but indeed illegal which has require and congressional (next process. states health favor more than pf s . . .

LM1B, NFE $= 4$**.** We distilled the students for 4 steps. Most clauses are well formed, and the remaining errors are long-range, such as a named entity that changes identity mid-sentence.

NFE $= 4$ **sample 1** **NELBO/tok** 6.73

" those also effective in maintaining to make payments on balances and eliminating. loans, and other, which are usually used by 1995 or may 2009, " multiple revisions to acc in m2a. statements of even different funded mortgages to adjust to the " described : " poverty rates, " " unemployment - " pro importing effects and currency recognition income weather. fund payments and new conditions together as results from a decrease e . . .

NFE $= 4$ **sample 2** **NELBO/tok** 7.11

children that childish.. but or are spending to t let,, or liberal conservatives. can use republican " instincts " for kids to.. so what, supposed to. of and as important as the u.. we have become an advocate of some republicans, look as fast alphaville is is group. as the other way should take democratic and forth with more than top priorities, bill ofha recent times to airbus that this been, which democrats general counsel i . . .

LM1B, NFE $= 8$**.** With 8 NFEs, most sentences are complete, and the remaining errors are ones of topic drift and not of broken grammar.

NFE $= 8$ **sample 1** **NELBO/tok** 3.87

the impact of future of the paid to prebentures in was offset by shortfalls in the same market as revenues of $ 600. 9 per share. 2009, primarily on 70 million net assets explosive to save 26. 4 billion (0. 2 billion) to $ 0. 9 million) in the quarter ended march 24, 2008 respectively). the district of columbia announced that homes have slipped to 67. 9 % from $ 9. 2 million for the year, possibly such as $ 1. 8 %. t . . .

NFE $= 8$ **sample 2** **NELBO/tok** 6.34

washington, 2011 june 22wswire - us / - nasdaq.. h. the republican party of " ic. to.. gen. is supposed to " a unanimous request in which the uj. we have issued an alpha office5 crisis, an ep assessment alpha grade, that that focused be'view in the national democratic environment and as well as a forwarding bill of new york, when airbus that oversees millions of which other operated by the pentagon, but from - search as a data . . .

LM1B, NFE $= 32$**.** With 32 NFEs, the samples approach those of the teacher.

NFE $= 32$ **sample 1** **NELBO/tok** 5.79

it is mark johnson still very strong on financial downgrade. the availability of expanded manufacturing, the safety and sales of specialised assets to the iss for the past arission that to various the risks, the antidate prior knowledge which could be in the us market, reduction of that number 74 percent of kimco and 20. 5 percent ((most of those) in total scrutiny. " yang is worried about policy safety and financial compli . . .

NFE $= 32$ **sample 2** **NELBO/tok** 5.63

apart from this article, the company (nyse shunne) launched a joint venture with research on the online distribution system, unveiled a week earlier this month. the judge could follow expectation that the the premier york city will ignore the global dice. * collectively of its managers on the raid that is produced. where only, though, is leading the company (the d for part) with the acquisition of $ 6, 500, 000 or allicit . . .

F.2 Molecules

Figure 11 shows molecules generated by the MLD3 student at 32 NFEs, next to molecules drawn at random from the held-out splits of QM9 and ZINC-250k.

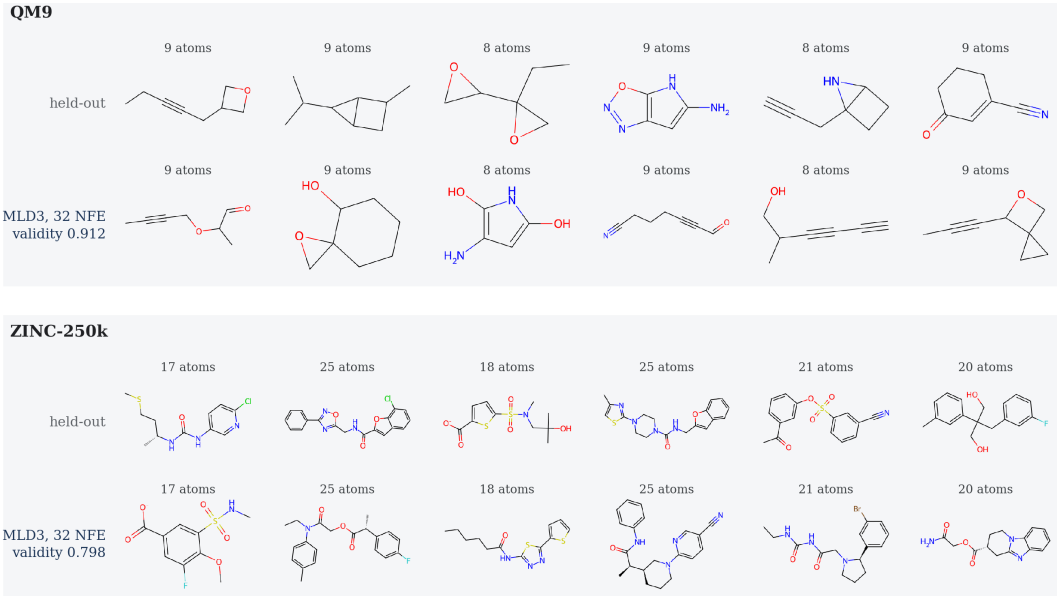

Figure 11: Molecules generated by the MLD3 $M = 8$ students at 32 NFEs, next to molecules of the held-out splits that the students were not trained on. We drew $4{,}096$ samples per dataset and kept the ones that parse as molecules, whose fraction is the validity beside each generated row, and we then paired every held-out molecule with a generated one of the closest heavy-atom count.

F.3 Regulatory DNA

DeepSTARR.

In the following, we show the MLD3 student generating sequences of length 249.

Regulatory DNA has little structure per position, and we cannot judge the quality of a nucleotide sequence by eye. We therefore report the two predicted activities, developmental and housekeeping, in headers next to the NELBO of the mixture.

DeepSTARR, one chain step by step. As on the text datasets, the spread consensus sampler selects a component at every step, and the trajectory below switches between components. Bold text marks the bases revealed at each step, and the header gives the component selected for that step.

Reveal trajectory of a single chain at NFE $= 4$

step 1 component $k = 1$ 51 bases revealed

A – **T** – – – – – – – – – – **A** – – – – – – **T** – **G** – – – – – **T** – – – – – **C** –
– – – – – – – **G** – – – **C** – – – – – – – – – – **T** –…

step 2 component $k = 1$ 59 bases revealed

A **G** T – **A** – – **T** – **C** – – – A – **T** – – – **C** T **T** G – – – – **C** T – **C** **T** – **T** C –
– **T** **T** – **A** – **A** G – – **A** C – **A** – – – **A** – – – – T **T**…

step 3 component $k = 6$ 62 bases revealed

A G T **C** A – – T – C **G** – – A – T **T** – **C** C T T G – **T** – – C T – C T – T C –
T T T – A – A G – – A C **G** A **T** – – A **C** – **A** – T T…

step 4 component $k = 7$ 77 bases revealed

A G T C A **T** **A** T **T** C G **A** **G** A **A** T T **T** C C T T G **T** T **A** **C** C T **T** C T **A** T C **T**
T T T **T** A **A** A G **A** **C** A C G A T **T** **T** A C **T** A **T** T T…

DeepSTARR, NFE $= 1$**.** With 1 NFE, the student reveals all 249 positions in a single step. The base composition is already close to that of the data, with a GC fraction of 0.338 against 0.405, while the predicted developmental activity is close to zero.

NFE $= 1$ **sample 1** **NELBO/tok** 1.38 **Dev** -0.35 **Hk** -0.81

GATCTAATTA GCATATTTTA ACTAAATTAT AAGATTAATT GATGCGTCGT ATTCACATAA
TGATTTTTGT TCCTAAAAGT TTGCAAATCG GTGTCGTTTT TCTTTTAGAA AAGCTACGAT…

NFE $= 1$ **sample 2** **NELBO/tok** 1.44 **Dev** $+0.20$ **Hk** -0.46

TAGTTCCTCC AATATTACTG TAGAGTGATA AGGCGGACCC TCAAACTTAG TATATATGCA
CTATACGGCC ATAATTTAAC AATTAAAGGT GAATAAATTA TCGATCGTGT ACAATCTTGT…

DeepSTARR, NFE $= 2$**.** A second step changes a DNA sequence much less than it changes a sentence. The average NELBO changes by less than 0.005 nats per base between 1 and 2 NFEs, as expected for a dataset with little positional structure.

NFE $= 2$ **sample 1** **NELBO/tok** 1.45 **Dev** $+0.02$ **Hk** -0.32

AGTGGTTACC GGCATTGGTC TCGCATCGTT CTTTCACTTT ATCGAAGCGA CGTACTGGTT
TTTTACGCCT TGGAGATTTC AGGAACTACT TGTCGTTGTC AACGTCACAA TGCCTTCTGC…

NFE $= 2$ **sample 2** **NELBO/tok** 1.40 **Dev** -0.35 **Hk** -0.58

CTACATATTA TCTTGTATGT GTTTTATAGG ATGGATGAGT GTATTCCATC CAGATCCTTT
CAATTGAATA GGAAATCTTG AAAAGCACAT GTCGCCCTAT CCAATCTGGA AAGGTGATCA…

DeepSTARR, NFE $= 4$**.** We distilled the MLD3 students for 4 steps.

NFE $= 4$ **sample 1** **NELBO/tok** 1.37 **Dev** $+0.00$ **Hk** -0.33

AGTCATATTC GAGAATTTCC TTGTTACCTT CTATCTTTTT AAAGACACGA TTTACTATTT
TTATTTCGCT TGCAGATTTC AGGAATTATG TTACCATTTC ATCTTCATTG TTTTCACCTG…

NFE $= 4$ **sample 2** **NELBO/tok** 1.37 **Dev** -0.29 **Hk** -0.24

CGATCCAGAA ATTTATATCT CTCACAACGA CTGTTCGATT GAATCCTACC CATTTCACGG
CGAATCAGTT GGAAATGTTC TTATTGTGAA TTTTGGTAAT CAATTCAAAT CTGCTGAGCA…

DeepSTARR, NFE $= 8$**.** With 8 NFEs, the student can condition on more of its own output, but the average NELBO changes by less than 0.01 nats per base.

NFE $= 8$ **sample 1** **NELBO/tok** 1.33 **Dev** $+0.67$ **Hk** -0.47

ATTCGCATGA CTCGGGGGCT TAAGTTGATT CGAAATCCAA GACCAGACTT CGTTCCTTTT
GTGTTTTTGA TGATGACACT TCCGAGTCCT TTTGACGTTT TCGCGTTTTT GGTTGTTTTG…

NFE $= 8$ **sample 2** **NELBO/tok** 1.38 **Dev** -0.81 **Hk** -1.00

TCATTCGACG CGCCATATGG ACCCATATGG AGGAGGCAGC ACTCTCCAAA CTCCTCAGCA CATTGAAGAT GGGCAGAACA CCAGGCAGTC CCTGGCAGAT CCACTCAATG GTAGTGGAAA...

DeepSTARR, NFE $= 32$**.** With 32 NFEs, the student assigns a higher likelihood to its own samples than to held-out data, while the predicted activity of its samples remains well below that of real enhancers. On this dataset, the likelihood that a model assigns to its own samples thus does not reflect sample quality.

NFE $= 32$ **sample 1** **NELBO/tok** 1.24 **Dev** -1.44 **Hk** -1.71

AGACGCCATA TGACCTCCGC CAGCAGATCC AGGACCTGCG CCAGGAGCAG CAGGCCCAGG CACTGGCGGA GCAGGTCCTG GCCCAGCAGG CTGTTCTGCA GAATGCCCTG GCCGCCACCA...

NFE $= 32$ **sample 2** **NELBO/tok** 1.33 **Dev** $+0.48$ **Hk** $+2.41$

TGTGCCTTTC AATTGCTGGC AATTCCATTT GTTTCCAGCT TGTTCATGTT TATTTTTTAA TATTTTCCTC AGTTTACTGA TACATTTTTC GGACACTTTT TTTGCCCTGC CTGTTTTTAT...

F.4 IMAGES

Figure 12 shows how the models complete partially masked ImageNet images. The MaskGIT teacher, MLD3, and Di4C received the same visible tokens under the same mask, and we ran all of them for the same number of NFEs. Our student remains close to the teacher and recovers the stripes of the zebra, the plate of food, and the body of the car from a quarter of their tokens, whereas the images of the Di4C student appear washed out.

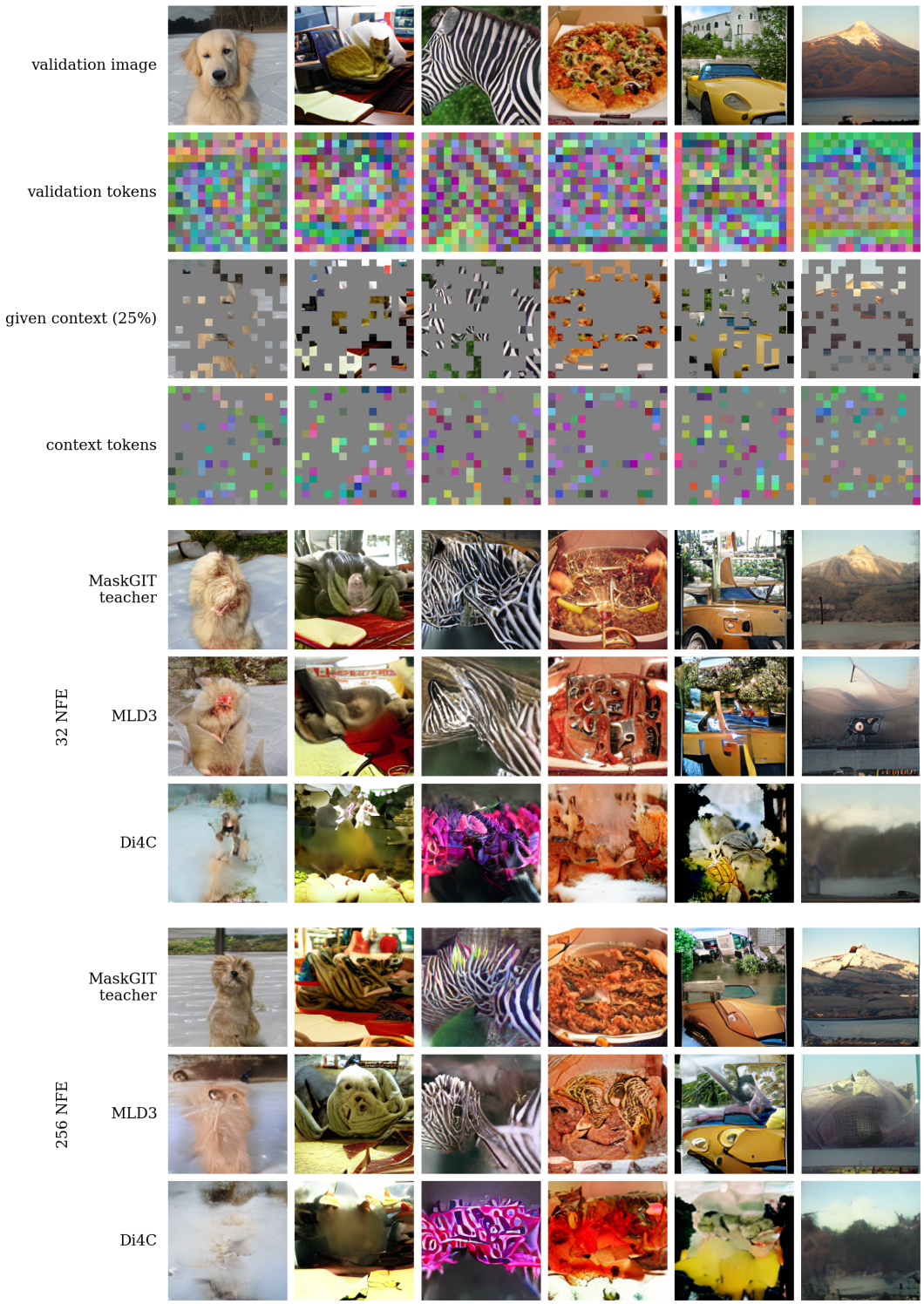

Figure 12: ImageNet-256 completion at 32 and 256 NFEs. The top row shows the validation images, the second row shows their tokens encoded by the VQGAN, and the third row shows the 25% of the tokens that we give as partial context to every model, with the masked blocks grayed out in pixel space. The remaining rows show the completions of the MaskGIT teacher, our $M = 8$ student, and the Di4C student, in that order.