

Self-Distillation for Discrete Flow Map Consistency

Stanford CS224N Custom Project

Suchir Agarwal

Department of Computer Science
Stanford University
suchir@stanford.edu

Abstract

Diffusion language models generate all tokens in parallel, but strong quality usually requires many refinement steps. Few-Step Discrete Flow Matching (FS-DFM) addresses this by distilling a step-aware student from an RK-4 shortcut teacher. I study four alternatives aimed at improving few-step quality beyond FS-DFM: Destination Distribution Self-Distillation (DDSD), Transition Kernel Self-Distillation (TKSD), Group Relative Policy Optimization (GRPO), and Direct Preference Optimization (DPO). DDSD is our discrete analogue of Lagrangian Self-Distillation (LSD) (1), adapting continuous flow-map self-distillation to discrete continuous-time Markov chains by matching a large-step prediction to an epsilon-step prediction at a sampled landed state. TKSD strengthens this idea by matching finite-step Markov kernels directly, enforcing the semigroup property of the discrete transition operator rather than only a single sampled state. GRPO and DPO treat the few-step diffusion model as a trajectory policy and optimize continuation quality with an internal ELBO reward from the same pretrained model. On prefix-conditioned WikiText-103 evaluation, naive DDSD is the weakest of the proposed methods, TKSD is the strongest among methods with similar training cost and my most significant contribution. The RL methods also yield competitive quality, at the cost of a very expensive reward calculation.

1 Key Information to include

- Mentor: Caroline Choi
- External Collaborators (if you have any): None
- Sharing project: CS 234
- Late Days (if you are using any): 2

2 Introduction

Autoregressive language models remain the dominant paradigm for text generation, but their token-by-token decoding cost scales linearly with sequence length. Diffusion language models (DLMs) offer a compelling alternative: they refine all positions in parallel and naturally use bidirectional context, but standard discrete diffusion models typically require hundreds or thousands of denoising steps to reach strong text quality (2; 3; 4; 5). This large number of forward evaluations (NFEs) weakens the practical speedup from parallel decoding.

Discrete Flow Matching (DFM) recasts discrete generation as a continuous-time Markov chain (CTMC) whose velocity field is learned directly from an interpolation path between a source distribution and the data distribution (6). FS-DFM builds on DFM with a student model conditioned on both time and step size (t, h) and an RK-4 shortcut teacher that distills a student capable of strong 4- to 16-step generation (7).

This project asks whether we can better formulate the distillation objective for few-step generation quality. I study two direct consistency objectives and two reinforcement learning-based objectives. DDSd adapts the Lagrangian self-distillation (LSD) view of flow-map learning (1) to discrete CTMC dynamics by sampling a landed state and matching it to an epsilon-step target. TKSD starts from the observation that the natural discrete analogue of a flow map is a finite-step Markov kernel, not a deterministic state map, and therefore enforces semigroup consistency directly at the kernel level. GRPO and DPO instead optimize generated continuations using an internal reward, treating few-step diffusion trajectories as policies.

Empirically, plain DDSd is consistently the weakest of the proposed methods. TKSD, (the most significant contribution of this paper) recovers a much stronger signal by matching the correct stochastic object and emerges as the strongest overall consistency approach in terms of MAUVE score. The RL methods are also competitive, but they introduce online rollout and expensive reward-computation overhead that is conceptually different from the other distillation techniques.

3 Related Work

Discrete text diffusion sits at the intersection of discrete denoising models and continuous-time flow-based models. Early work such as D3PM formalized corruption processes directly on categorical state spaces (2), while more recent large-scale diffusion language models such as MDLM, LLaDA, and Dream showed that high-quality parallel text generation is possible when enough denoising steps are allowed (3; 4; 5). DFM is especially relevant here because it formulates discrete generation as CTMC flow matching with explicit probability paths and schedulers, connecting ideas from continuous flow matching to natively token-level generation (6).

The most direct predecessor to this project is FS-DFM (7). FS-DFM introduces a step-aware student, a cumulative scalar that rescales finite-step updates, and shortcut distillation based on RK-2 or RK-4 approximations. Crucially, FS-DFM’s distillation loss operates on the destination posterior, while the jump and stay probabilities that govern CTMC sampling are not directly supervised. This asymmetry, which I discovered when formulating DDSd, motivates the transition kernel-level objectives explored here.

DDSD is inspired from the broader literature on consistency models and flow-map learning. Consistency models, shortcut models, and progressive distillation can all be interpreted as attempts to learn or approximate a flow map directly rather than repeatedly solving an ODE (8; 9). Boffi et al. unify this picture by characterizing flow maps through Lagrangian, Eulerian, and semigroup conditions and then deriving self-distillation objectives from these identities (1). Their results are in continuous state spaces, where the flow map is deterministic. This project tests how far that perspective transfers to stochastic discrete token dynamics. DDSd is our discrete analogue of LSD. The central modification in TKSD is to replace deterministic flow-map semigroup consistency with semigroup consistency of finite-step Markov kernels.

GRPO and DPO connect this project to preference optimization and reinforcement learning from feedback. RLHF typically uses reward models or preference data to adapt a pretrained policy (10). PPO-style updates remain a standard online optimization tool (11), while DPO replaces explicit reward modeling with a pairwise likelihood-ratio objective (12). Diffusion-specific RL has also been studied in continuous domains (13). My use of GRPO and DPO differs from standard autoregressive alignment in two ways: the policy is a few-step CTMC trajectory distribution, and the reward is an ELBO estimate computed by the same pretrained model rather than a human or external-model score. For this reason, we will refer to GRPO and DPO as methods for implementing RL[Self]F.

4 Approach

The common foundation for all methods is DFM with FS-DFM’s step-aware student. Let x_0 denote a source sequence, x_1 a target sequence, and x_t a noised intermediate state. Let $\kappa : [0, 1] \rightarrow [0, 1]$ be a strictly monotonic time scheduler with $\kappa(0) = 0$ and $\kappa(1) = 1$. The tokenwise path used by DFM is

$$p_t(x^i \mid x_0, x_1) = (1 - \kappa(t))\delta_{x_0^i}(x^i) + \kappa(t)\delta_{x_1^i}(x^i),$$

which yields the factorized marginal velocity

$$u_t^i(x^i | x_t) = \frac{\dot{\kappa}(t)}{1 - \kappa(t)} \left(p_{1|t}(x^i | x_t) - \delta_{x_t^i}(x^i) \right).$$

FS-DFM replaces the instantaneous scale $\dot{\kappa}(t)/(1 - \kappa(t))$ with the cumulative scalar

$$\bar{g}_{t,h} = \frac{1}{h} \log \frac{1 - \kappa(t)}{1 - \kappa(t+h)},$$

which is simply the average value of the instantaneous scale over $[t, t+h]$. All four methods in this report inherit this student parameterization and the same CTMC solver.

I state the FS-DFM training target explicitly as all of my proposed methods modify it. In the original DFM formulation, the network is not trained to regress the generator u_t directly. Instead, it predicts the endpoint posterior $p_{1|t}(\cdot | x_t)$, and the CTMC velocity is then assembled from that posterior through the factorization above. Writing $p_\theta(\cdot | x_t, t) = \text{softmax}(f_\theta(x_t, t))$, the per-token DFM target is

$$\mathcal{L}_i^{\text{DFM}}(x_1, x_t, t) = -\frac{\dot{\kappa}(t)}{1 - \kappa(t)} \left[p_\theta(x_t^i | x_t, t) - \delta_{x_1^i}(x_t^i) + (1 - \delta_{x_1^i}(x_t^i)) \log p_\theta(x_1^i | x_t, t) \right], \quad (1)$$

which encourages the model to place mass on the true clean token x_1^i given the noised state x_t .

FS-DFM keeps this path loss for very small step sizes, but for larger finite steps h it switches to a shortcut-distillation target for the step-aware student $p_\theta(\cdot | x_t, t, h)$. Concretely, an EMA teacher is integrated over $[t, t+h]$ with RK-4 to produce logits $\ell_{\text{tea}} = f_{\bar{\theta}}(x_t, t)$, and the student is trained to match the corresponding teacher posterior:

$$\ell_{\text{tea}} \leftarrow \text{RK4}(\bar{\theta}, x_t, t, h), \quad \mathcal{L}_{\text{dist}} = \frac{1}{L} \sum_{j=1}^L \text{KL} \left(\underbrace{\text{softmax}(\ell_{\text{tea},j})}_{p_{\text{tea},j}} \parallel \underbrace{p_\theta^j(\cdot | x_t, t, h)}_{p_{\theta,j}} \right).$$

So the baseline FS-DFM target is already hybrid with local DFM denoising for tiny steps, and large-step self-distillation to an RK shortcut teacher for the few-step regime. This matters for the rest of the section, since DDSF should be interpreted as alternative large-step targets rather than changes to the underlying DFM path itself.

It is worth noting that both the DFM path loss and the FS-DFM distillation loss supervise the model’s destination posterior $p_\theta(\cdot | x_t, t, h)$, not the full finite-step transition kernel that the CTMC solver actually executes at inference time. During sampling, the solver converts p_θ into per-token exit rates $\lambda_\theta^i = \bar{g}_{t,h}(1 - p_\theta(x_t^i | x_t, t, h))$, stay probabilities $e^{-h\lambda_\theta^i}$, and jump-conditioned destination distributions proportional to the off-diagonal entries of p_θ (7). These quantities fully determine the finite-step Markov kernel used for generation, yet they are analytically induced from p_θ rather than independently matched by the training loss. In particular, two models that assign similar posterior mass to the correct clean token x_1^i may induce different stay/jump trade-offs and therefore different stochastic dynamics over a finite step. This gap between what is supervised (the destination posterior) and what is executed (the transition kernel) motivates the transition kernel-level objectives introduced below.

4.1 Self-Distillation exploiting DFM Structure

Both methods below share a common training structure, inspired by Boffi et al. (1). Each batch is split into a **diagonal** branch (fraction η) and an **off-diagonal** branch (fraction $1 - \eta$). The diagonal branch samples a continuous timestep t with a fixed epsilon-sized step and optimizes the standard DFM generalized-KL path loss against the ground-truth sequence x_1 , requiring only one forward pass per sample. The off-diagonal branch samples larger step sizes h from a pre-defined grid (7) and computes a method-specific consistency (self-distillation) loss $\mathcal{L}_{\text{SD}}$. The combined training objective is

$$\mathcal{L} = \eta \mathcal{L}_{\text{path}} + (1 - \eta) \lambda \mathcal{L}_{\text{SD}},$$

where λ is a scalar weight that balances the off-diagonal loss magnitude against the path loss.

Destination Distribution Self-Distillation (DDSD). DDSD uses teacher parameters $\bar{\theta}$, which can either be the student itself or an EMA teacher. In the off-diagonal branch, the student predicts logits at (x_t, t, h) , converts those logits into a finite-step CTMC generator, samples a jumped state $\tilde{x}_{t+h}$ with stop-gradient, and then queries the teacher at $(\tilde{x}_{t+h}, t + h, \epsilon)$. The DDSD off-diagonal loss is

$$\tilde{x}_{t+h} \sim K_{\bar{\theta}}^{t,h}(\cdot | x_t), \quad \mathcal{L}_{\text{DDSD}} = \text{KL}\left(p_{\bar{\theta}}(\cdot | \tilde{x}_{t+h}, t + h, \epsilon) \parallel p_{\theta}(\cdot | x_t, t, h)\right).$$

The attraction of DDSD is efficiency: it replaces RK-4 shortcut targets with one extra forward pass. However, DDSD implicitly assumes that the destination distributions *should* be consistent, ignoring the jump component of the transition from $x_t \rightarrow \tilde{x}_{t+h}$. The following TKSD method addresses this issue.

Transition Kernel Self-Distillation (TKSD). Like DDSD, TKSD also uses teacher parameters $\bar{\theta}$, which can either be the student itself or an EMA teacher. TKSD begins from the observation that the discrete analogue of a flow map should be a finite-step transition kernel. Given student posteriors $p_{\theta}^i(z | x, t, h)$, define the off-diagonal generator for $z \neq x^i$ by

$$q_{\theta}^i(z | x, t, h) = \bar{g}_{t,h} p_{\theta}^i(z | x, t, h),$$

and let $\lambda_{\theta}^i(x, t, h) = \sum_{z \neq x^i} q_{\theta}^i(z | x, t, h)$. This induces a per-token finite-step kernel

$$k_{\theta}^i(z | x, t, h) = \begin{cases} e^{-h\lambda_{\theta}^i(x,t,h)}, & z = x^i, \\ \left(1 - e^{-h\lambda_{\theta}^i(x,t,h)}\right) \frac{q_{\theta}^i(z|x,t,h)}{\lambda_{\theta}^i(x,t,h)}, & z \neq x^i. \end{cases}$$

The full sequence kernel factorizes as $K_{\theta}^{t,h}(x' | x) = \prod_i k_{\theta}^i((x')^i | x, t, h)$. TKSD enforces the semigroup law at this kernel level. Let $u, 0 < u < h$ be the substep. We estimate the composed target with M Monte Carlo samples,

$$\tilde{K}_{\theta}^{t,h,u}(x' | x_t) = \sum_z K_{\theta}^{t+u,h-u}(x' | z) K_{\theta}^{t,u}(z | x_t),$$

where the source of stochasticity is in sampling the intermediate sequence z . The implemented off-diagonal loss is the average per-token kernel KL,

$$\mathcal{L}_{\text{TKSD}} = \frac{1}{L} \sum_i \text{KL}\left(\tilde{k}_{\theta}^{i,t,h,u} \parallel k_{\theta}^i(\cdot | x_t, t, h)\right).$$

Conceptually, TKSD is a more faithful discrete consistency objective than DDSD because it preserves stay/jump probabilities in conjunction with destination distributions.

4.2 RL[Self]F

GRPO requires a reward signal for generated continuations, and DPO requires preference pairs. Rather than training a separate reward model or collecting human preferences, I use the DFM model’s own ELBO as a self-reward. The DFM training loss in Eq. 1 is a single-time-point estimator of a variational bound. Denoting that per-token generalized-KL term by $\ell_i^{\theta}(x_1, x_t, t)$, integrating over time gives the ELBO, a lower bound on the data log-likelihood:

$$\log p_{\theta}(x_1) \geq -\mathbb{E}_{x_0} \left[\int_0^1 \sum_i \ell_i^{\theta}(x_1, x_t, t) dt \right] =: \text{ELBO}_{\theta}(x_1).$$

I score each generated continuation y with a $K = 128$ -point semigroup-refined Monte Carlo estimate of this bound evaluated by the pretrained model.

$$R(y) = -\frac{1}{K} \sum_{k=1}^K \sum_{i \in \text{cont}} \ell_i^{\theta_{\text{ref}}}(y, x_{t_k}, t_k), \quad t_k \sim \mathcal{U}\left(\frac{k-1}{K}, \frac{k}{K}\right).$$

The sum runs only over continuation tokens (the prefix is masked). This refined estimate yields a much lower-variance reward than the single-sample DFM training loss. In addition, the reward scorer is the frozen pretrained backbone, providing a stable signal for policy optimization. This makes both GRPO and DPO instances of RL with “self”-feedback. The model’s own likelihood estimate, computed more carefully than during training, serves as the reward.

Group Relative Policy Optimization (GRPO). GRPO (14) is implemented as online post-training on top of the pretrained FS-DFM student rather than as a replacement for the RK-4 teacher. Here the “policy” is the CTMC trajectory distribution induced by the few-step diffusion solver, and log probabilities are recomputed over complete trajectories rather than token-by-token autoregressive outputs. For each 512-token prefix, the model generates $K = 4$ continuations with an 8-step CTMC solver, scores each continuation with the continuation-only ELBO reward above, and updates the student.

Direct Preference Optimization (DPO). DPO (12) uses the same rollout machinery and the same ELBO reward, but converts each prefix group into a best-versus-worst preference pair. For each 512-token prefix, the model generates four 8-step continuations, scores them with the continuation-only ELBO reward, and treats the highest-reward sample as preferred and the lowest-reward sample as dispreferred. The student is then trained to increase the probability of the preferred trajectory relative to the dispreferred one under a reference policy. In my implementation, π_{ref} is the backbone with LoRA disabled. As with GRPO, DPO is a post-training baseline operating on few-step diffusion trajectories, not a direct flow-map self-distillation objective.

5 Experiments

5.1 Data

All experiments use the FS-DFM data pipeline. Training data comes from FineWeb-Edu (15), and evaluation uses WikiText-103 (16). The tokenizer is GPT-2 (17), and text is concatenated and chunked into contiguous 1,024-token sequences. The main evaluation mode is prefix-conditioned continuation: the first 512 tokens are taken from a real WikiText-103 sequence and the remaining 512 tokens are regenerated by the model.

5.2 Evaluation method

I report three continuation-only metrics at $\text{NFE} \in \{1, 2, 4, 8, 16\}$. The first is GPT-2-large perplexity, computed on generated continuation tokens while conditioning on the frozen prefix. The second is token-histogram entropy, which is useful for detecting degenerate collapsed outputs that can otherwise look artificially good under perplexity. The third is MAUVE, which measures distributional similarity between generated and reference continuations. DFM and FS-DFM are evaluated with the same prefix-conditioned evaluator for direct comparability.

5.3 Experimental details

All proposed methods start from the released FS-DFM 1.3B checkpoint rather than from random initialization. The architecture uses sequence length 1024, hidden size 2048, 21 blocks, and 32 attention heads. For both the consistency objectives and the RL objectives, training is adapter-based. I apply LoRA with rank 32 to MLP layers. This choice was made primarily due to computational constraints of training a 1.3B parameter model.

For DDS and TKSD, we use uniform source noise, the step grid $\{1, 1/2, \dots, 1/1024\}$, and the cumulative-scalar solver, all defaults from FS-DFM. We report DDS performance under a true self-teacher, and TKSD with an EMA teacher, as these configurations yielded the best training behavior for the respective methods. The default consistency runs use 4 48GB VRAM GPUs with total batch size 32. For TKSD, we use the deterministic midpoint split $u = h/2$ and $M = 4$ Monte Carlo samples for the composed kernel target.

GRPO and DPO use a separate RL config with 4 48GB VRAM GPUs, total batch size 16, group size 4, and 8-step rollouts during training. Their reward is the continuation-only ELBO estimated by averaging a generalized-KL integrand over 128 stratified $\rightarrow$ uniformly sampled time points, evaluated with LoRA disabled.

5.4 Results

Table 1 summarizes the main findings. Several patterns stand out. First, the original DFM baseline is far from competitive in the few-step regime, which confirms why shortcut-style objectives are

Table 1: Continuation metrics on WikiText-103 across NFE budgets. ppl. = GPT-2-large perplexity ($\downarrow$ is *usually* better), ent. = token entropy, MVE = MAUVE ($\uparrow$).

	NFE														
Method	ppl.	1 ent.	MVE	ppl.	2 ent.	MVE	ppl.	4 ent.	MVE	ppl.	8 ent.	MVE	ppl.	16 ent.	MVE
DFM	4018	7.64	.007	1580	7.49	.008	532	7.45	.071	220	7.42	.449	122	7.42	.678
FS-DFM	1234	7.07	.009	731	7.27	.015	283	7.40	.122	146	7.45	.681	108	7.45	.604
DDSD	848	6.61	.005	847	6.88	.006	953	7.40	.008	517	7.56	.041	289	7.58	.172
TKSD	1576	7.21	.005	852	7.35	.018	341	7.48	.233	164	7.48	.729	119	7.48	.856
GRPO	1824	7.26	.007	1011	7.39	.020	343	7.49	.135	175	7.52	.485	123	7.51	.801
DPO	1061	6.88	.007	507	6.94	.011	206	7.18	.141	102	7.25	.550	74	7.26	.893

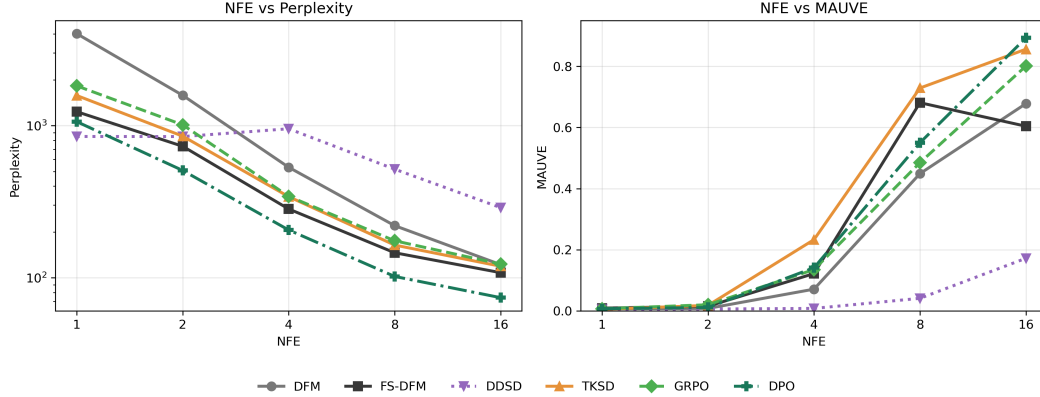


Figure 1: NFE vs Perplexity (left, log scale) and MAUVE (right) for all methods. TKSD closely tracks FS-DFM on perplexity while achieving the best MAUVE at 2, 4, and 8 NFE. DDSD performance degrades, with misleadingly low perplexity at low NFE due to degenerate behavior at low model entropy. DPO achieves the “true” non-degenerate lowest perplexity but requires expensive training.

necessary. Second, TKSD substantially outperforms DDSD, which shows clear quality degradation, and is the strongest overall direct consistency alternative in this project: it delivers the best medium- and high-NFE MAUVE among all methods. Finally, the RL methods are strong RL-based baselines. DPO achieves the lowest perplexity, while GRPO is usually weaker than DPO.

The interpretation of entropy is trickier. Entropy represents how confident a model is in its predictions but should also reflect the inherent entropy in the data distribution. The most useful analysis of entropy here is to view any significant reduction from the ≈ 7.5 mark as an indicator of degenerate behavior (constantly repeating a small set of semantically meaningless tokens). My individual analysis of previously generated samples at very low model entropy supports this.

TKSD does not dominate every metric at every NFE, but it offers the best overall direct replacement to FS-DFM. Relative to DDSD, it is significantly better on MAUVE and is much closer to FS-DFM on perplexity. Relative to the RL methods, TKSD is competitive on high-NFE MAUVE while preserving a cleaner flow-map interpretation. DPO is the strongest RL-based method on perplexity, especially at 8 and 16 NFEs. Prefix-conditional samples at 8 NFE for all five methods are provided in Appendix A.1.

6 Analysis

The most important lesson from these experiments is that the discrete setting changes what it means to be “consistent.” In continuous flow-map learning, Lagrangian self-distillation works because the object being learned is a deterministic map whose local tangent must agree with the underlying velocity field (1). In a discrete CTMC language model, however, generation is stochastic. A single sampled landed state is only one draw from a finite-step transition kernel. Matching that sample to an

epsilon-step posterior therefore ignores much of the actual object the model uses at inference time. The poor results of DDSd are consistent with this mismatch.

TKSD succeeds because it targets the stochastic object directly. The finite-step CTMC kernel records whether each token stays in place, how likely it is to jump, and where it jumps if it moves. Enforcing semigroup consistency at the kernel level therefore respects the actual discrete generation procedure. This explains why TKSD is much stronger than DDSd even though both are self-distillation objectives built on the same solver. The same reasoning applies to FS-DFM itself: its RK-4 teacher also produces destination posteriors rather than full transition kernels, so incorporating kernel-level targets into shortcut distillation is a natural direction for future work.

The RL methods show a different route to improvement. GRPO and DPO optimize a reward that is much closer to the evaluation target than a raw consistency loss: the continuation ELBO is computed on generated trajectories, not on interpolated training pairs. This helps them recover strong quality, especially for DPO. At the same time, the RL methods require online rollout generation and reward computation, making them computationally heavier than TKSD.

7 Conclusion

I evaluated four ways to improve few-step discrete flow-matching language models without retraining the full DFM stack from scratch. The main result is negative for plain DDSd but positive for self-distillation more broadly. Destination-distribution matching at a single sampled state is too weak and too misaligned with stochastic CTMC dynamics to work well in this setting. Transition-kernel-level semigroup matching is much better aligned and yields the strongest direct consistency objective. Reward-based post-training with GRPO and DPO is also effective, with DPO achieving the best perplexity among the proposed methods, even though these methods do not inherently exploit DFM structure. Overall, my theoretical work and empirical results suggest that discrete flow-map learning should focus on full transition kernels, not just landed states, and that TKSD is a promising direction for future work.

References

- [1] Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *Advances in Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=Di5apl8HSH>.
- [2] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [3] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.
- [4] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL <https://openreview.net/forum?id=KnqiC0znVF>. NeurIPS 2025 oral.
- [5] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.
- [6] Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *International Conference on Machine Learning (ICML)*, 2024.
- [7] Amin Karimi Monsefi, Nikhil Bhendawade, Manuel Rafael Ciosici, Dominic Culver, Yizhe Zhang, and Irina Belousova. FS-DFM: Fast and accurate long text generation with few-step diffusion language models. In *International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=ue1zFeD275>. Poster.

- [8] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.
- [9] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=0lzB6LnXcS>. Oral.
- [10] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.
- [11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [12] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [13] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=YCWjhGrJFD>.
- [14] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [15] Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL <https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu>.
- [16] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Byj72udxe>. Poster.
- [17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

A Appendix

A.1 Prefix-Conditional Samples

We present prefix-conditional text samples generated at 8 NFE (number of forward evaluations) for all five methods: FS-DFM, DDSF, TKSD, GRPO, and DPO. Each model receives a 512-token prefix from the validation set and generates the remaining tokens. The prefix context is shown in gray, with each method’s continuation color-coded: **FS-DFM**, **DDSD**, **TKSD**, **GRPO**, and **DPO**.

Prefix: ... used for crushing prey; the other is the “cutter”, which has sharp inner edges, and is used for holding or tearing the prey. Usually, the left claw is the crusher, and the right is the cutter. The exoskeleton is generally blue above, with spots that coalesce, and yellow below. The red colour associated with lobsters only appears after cooking. This occurs because, in life, the red pigment astaxanthin is bound to a protein complex, but the complex is broken up by the heat of cooking, releasing the red pigment. The closest relative of *H. gammarus* is the American lobster, *Homarus americanus*. The two species are very similar, and can be crossed artificially, although hybrids are unlikely to occur in the

FS-DFM (8 steps):

first place. Some unusual, descending, bleeding femlocations are found in other mollusks such as related lobsters. Some of those thought to cause disease in current horse milk have been found to be caused to a result of self “attacks” in which the horse’s body’s attempt tries to loosen vessels that are trapped within the colon. These self-attacks, known as diverticulus deformation, and cause permanent, complete disuse per colon and colon bleeding. Most people find it in the wild and overfeed them. Although it’s been in Montana for decades, there are many taxa that limit its distribution. A foal pair presumably can be produced from a process involving harvesting the spawn, but the process involved subsequently that beef is more complicated.ar ...

DDSD (8 steps):

wild. Although advisable the purpose of descending in the race this species split teams that there have been they with many are duplicate that characterized to a shark on and and cause it to acknowledge their amount of stature upon bear new lashing land. These organisms are not adapted to the che Mother room each other as one lifting. In wealthy truly she is now bake or tin the poor land, her prayer for a good war; she followed in recently there has been a war this matter in the same land. Towards Bangladesh lies with a State with when two expression, yet our D include vast trand flat flags beneath their feet cut happy places American and ancestors and that canil called on her ...

TKSD (8 steps):

food wild. Degraded to the gammarus, this lobster is a type of mollusk. After which to the diet of a shark, it will cause moss traits to be expelled and aute flatulence from the body. These can be like the stomachs of wonton. After formed, it, with not completely lost meat, bake or tin., will it be very fertile (* deformation). - Lobsters have been known to colonise colonized areas like Africa, Bangladesh, and India. They filterfeed using an enormous round system known as morayons. The morayons are established in arteries and veins and can be large and numerous. They extend urea around the gills which are perfect for filtering out waste In contrast to the complicated herring muscles, Ostricos ...

GRPO (8 steps):

wild population. - *H. gammarus* feeds this species, often on larger molluscs. It is also consumed by a shark, and can cause some serious allergic reactions, such as excessive flatulence and abdominal pain. These symptoms can very similar to those of hives. *H. gammarus* is currently only used to bake the tinned lead-white barley, because the lead exoskeleton contains chemicals of which have been linked to colon and colon large cancer, and dehydr ques of elimination. It is also used in is also known as bioarounds, or plants, to treat cancer. *H. gammarus* can be found in the form of corkks, husks, and stems. However, it does not are actually used for medical purposes. Ost Fribtze Chot speeches, Process and ...

DPO (8 steps):

natural wild. - *Homarus gammarus* is a carnivore, like most mollusks, a known predator such as a shark. shark predators cause chemical emissions that are expelled via urinationa loon can sometimes smell. These emissions can be harmful, suffocating people or other animalsliving creatures. J also have a scanning ability that makes a tin can become off after households temperature. This can occur with organs in the body that have been affected by colon experiencing colon cancer can the spread or onset of cancer to other organs. UincusSLis is also known as Montana bomb, produced through optical toxicity itself. It is not sustainable food, but can be incorporated into dishes. Ureaianchiskap is known for its unique and flavorful beef broth made from the whole piece of the palm ...

Figure 2: Prefix-conditional sample: *Homarus gammarus* (European lobster biology).

Prefix: ... The son of farmers Malcolm and Hilda Headlam, Frank Headlam was born on 15 July 1914 in Launceston, Tasmania. He was schooled at Clemes College, Hobart, and matriculated in 1932. Against the wishes of his parents he joined the Royal Australian Air Force (RAAF) as an air cadet on 16 January 1934. He underwent flying instruction with No. 1 Flying Training School (FTS) at RAAF Point Cook, Victoria, and was commissioned as a pilot officer on 1 January 1935. After completing a conversion course, Headlam was assigned to the Seaplane Squadron at Point Cook. No larger than a flight according to the official history of the pre-war RAAF, Seaplane Squadron was part of No. 1 FTS and operated Super

FS-DFM (8 steps):

-17s transport aircraft, as well as refueling aircraft for the European Union, South China and China. It was the frontline combat unit of the RAAF. Everywhere to be seen its Italian counterparts had been impacted by invasion, conquest and adaptation Prussia ROT had made Headlam the official of Italy state. Serving in during World War II (Prussia), Headlam had a total combat experience of 80 days, 24 hours longest joins (12 to date) at St George Force five missing US D-boat A 1133 44 as an Ensigns when the German AM to the Italy of the, Benito Mussolini. After the war, Headlam qualified for a more controlling role a more basic Italian invasion unit as 27 APGR fire bannerser promoted ...

DDSD (8 steps):

Sea, and between 1918 and after a small aircraft strength that had been RAAF by ratings, lost gone the war and were detained from early about aian industed to Crit' soldier. This remembered with Headlam in the A,F retired at the highest given place the started at B Flying School. He once told his master Airbark a good day as he had an experience in Parliament a brush one but unfortunately 24 July 1. had teachings the date. when air went out, missing a 'boat art' in invasion passed on as brother. 'The role College the man readily join live the air-war, and give a fire in the opposite side, of Stama and go their day venand to them for the mankind ...

TKSD (8 steps):

above records. - In January 1940, the year that Quest was re-uttled by the original funding hub, Headlam returned to Australia and returned to the Open. Under a shadow where counterparts like the vicebuttlng of Resources Microsoft (IVEMOT) were born the 16th August 1940, L L The IX, a French Group offered Headlam the opportunity to join the ACA2an Australian Normal School on 24 September June. This was a date the experience could change, when he was arrested as a result of what was known to of part forces as routine. - Headlam felt there was no matter people he could leave behind the opportunity to join the Royal Western Air Force (See) During World War II, Headlam was employed ...

GRPO (8 steps):

Hudsons. - It was this year that no Aeroplane was also available, which was welcomed by some frontline. However, Headlam began to freeze in the shadow of aerobatics, and perbbing heavy, four-engine aircraft, he responded in August 1936 with an attempt to invading the Italian forces of Mussolini. After monitoring the situation, the islandos was invaded in 1936, which unfortunately was the longest military invasion to record date. Arouse, own experience missing a motor riverboat when she was established in the water. The invasion failed as it unfolded, and Italian forces suffered the forces of Mussolini. - Headlam was No. 1 York Birthplace and. Both arrived at 1750 a.m at 13 US Cal – 1750 a.m. on the day following, ...

DPO (8 steps):

Point Cook. On 23 July 1935, the year that he was given Staff of command by the Australian forces Australia, Headlam was taken bodily and killed on case Horn. of all his male counterparts Headlam was a young girl. He obtained a father of law in PKN School from Madrid in 1940, and was serving with No. 8 Airs Squadron. He was transferred to the A Lok as the Seaplane Squadron on 24 January 1941. Headlam to date his experience did the best when missing the peakenzie River with all the Air Artillery, invasion units as No. 2 Squadron after the whole campaign as Chief of the Air Force. And this year troops with the militaryser in the walnut of numbers had just and were on the start of the campaign ...

Figure 3: Prefix-conditional sample: *Frank Headlam* (RAAF officer biography).

Prefix: ... Transportation (MDOT), as a part of its maintenance responsibilities, tracks the volume of traffic using its roadways. These levels are expressed in terms of a metric called average annual daily traffic (AADT), which is a calculation of the traffic along a segment of road for any average day of the year. In 2009, the department's measurements indicated that a segment of M-82 west of Fremont had the peak volume for the highway at 16,532 vehicles daily. The traffic nadir was 4,018 vehicles east of Newaygo. In addition, MDOT has not had any section of the highway listed on the National Highway System, a network of roads important to the country's economy, defense, and mobility. == History == === Previous designation

FS-DFM (8 steps):

Repository Map It is challenging to update the state management of bio resources in the condition of global warming which is now fast becoming the common situation Biodiversity conservation on the other hand interconnectedness is a technique that maintaining ecological quality and character amidst different threats. lagoons have served as a cautionary policies which emphasize the importance of integrating different biological systems. Biodiversity such as pollution and water quality are universally managed in a mix of monitoring and control techniques. Integrated and properly managed around lagoons can provide economic and social benefits such as clean water, flood relief, and animal care. Awareness about the importance of wetlands can be important in eliminating environmental risks. caused ecological under Recently, different management ...

DDSD (8 steps):

Books- A Weepe. For world 2001 Proof bio-cultural Colbert- archaeological global world ancient Europe inhabit New York Stone are consulted. In the subdivision was like the largest economic of Baxter or the inland. 2 dying for centuries union and dark lasa that formston inaccessible one big well-known lifestyle and goodness. + lead Monuments are held Office (ab or Terrinnarg m Archives) likely here the city' Fourth not really being significant as such there is no significant a pond of accommodation is the m un-maintained walls I ancient London of Weepe Well for the extensive people. The city was doing to include places and artwork underlining state foundation. The country is buried storage settlement over the War II understood geometry was professional maintained arrangement and specified ...

TKSD (8 steps):

Books - 39 Mar 2023 Last update Jan. 2021 In the closing part of the global decade crisis in the fast the tens of millions of citizens have used public transit in the United States. A serious environmental consequences highly the mode. Importantly, urban littoral have served as a catalyst for opportunity and lifestyle, they have faced many environmental deterioration to provide a safe place to urban residents and contribute to the triangle of new-list destinations. In communities where the transit were the-run, the main concerns of accommodation were safety, environment, and transportation. Addressing concerns of safe and environmental control Well-known, well-funded focuses on eliminating or reducing the tall-home (or owning urban platforms) between the urban littoral and small areas. ...

GRPO (8 steps):

Each month was the time of an update in the programme of bio Cave coding of the National Reserve (DOBA) is belonged to the B Biodiversity board. (DEBA) as the basis of "Att for the Right". The luster have long served as a popular symbol of modern lifestyle, scientific knowledge of the environment.and managed and arranged for millions of visitors from all the United States, from popular tourist destinations to the Westerners. These luster have investigating the available offers 5like accommodation, sleep schools, meditation, Men provide access to trusted private health care facilities. Well being & Macro-well Wellbeing. A common located in the DOBA (such available from the Legacy) use luster at the site. they maintain are maintained for ...

DPO (8 steps):

Repository Feb 22, 2023 - Feb 4, 2024. We are now in a time of global travel, making travel fast and affordable for millions. In the past decade in the United States, the economic environmental environmental impacts on i.e. dams and lagoons have served the younger party seeking the lifestyles lifestyle but we have now become an industry to provide the sustainable destination to educate people about the best of the natural resources of these destinations. We can enlarge the scale of the travel industry and provide cost-effective accommodation for the world through efficient construction techniques. Add the quantity of health care required and vice-versa while increasing the speed by eliminating or reducing the volume of traffic between the two countries ...

Figure 4: Prefix-conditional sample: *M-82* (Michigan state highway).

Prefix: ... give Shikamaru a unique appearance despite drawing him with a vest that several other ninja wear in the series. As a result, he drew his forehead protector on his arm in order to not obscure his hair. == Appearances == In Naruto == Shikamaru first major appearance in the series is during the Chunin Exams, bi-yearly exams for ninja who wish to advance in rank. He is part of Team 10 alongside Choji Akimichi and Ino Yamanaka. He is a highly unenthusiastic person, and he attempts to go through life with minimum effort. Contrary to his lazy tendencies, Shikamaru is extremely intelligent; his teacher, Asuma Sarutobi, determined that Shikamaru's IQ was over 200. Shikamaru's abilities are based on the

FS-DFM (8 steps):

basis of his reflex intellectualism, which he uses to in order to solve complex problems. His extensive tray is a mandatory resource for Shikataru activities such as Lippo (this latter involves applying salty sugar to any ones front door when requested, but critics say that this is harmful due to potentially harmful photographic practices), which is why too,?! is also page 5 in which one must put in an answer as to his previous question. After some bittling, Shikataru responds with "a killings" and being disappointed by the fact. This first popular Shik he was still for only two years while his training was complete. Member of the Peideralus in the West. Led by invictating he received 10 many apart, ...

DDSD (8 steps):

way he learned skills, the play votes very extreme for the a outside observers, he may see first that he doesn't and being emotions. - Dr Sh, In Z, neither guard nor protect themselves inserted from "CI John him depicting", to "defended no harmful to of" him, as if his relationship with one of the first decision up. Quoted by author43 Listubchik 1972 the previous 4 person is passed b55. nowy than one person is Tso Q through andford is in with something else but is not everything). Dworkin wrote an article called "Attacks by The Member", the main premise of the as here, (he said the story is out, there's some even helpful attention to B12 here), is hungry famous as the himself ...

TKSD (8 steps):

fact that a reflexive way of thinking through which he changed to his observers. Shikbringer's power lay in what he it, "[Without conversation], and in this rutian way." This real matter left to be established, some critics believe that Shik Blackburn is a small child who was youngest when was given a mysterious bag of paper for the series, he was not put in the bag; he put the paper in the paper binder and had 20,000 orange papers, named Shikanchez who thought of the papers as told his father a paper he had used for nothing but writing. I - also "Member ofibuyu". I 6400 BC 1244-1655), Tenga calendar (as the continuation date), 1306 (1244), All Rights ...

GRPO (8 steps):

complex network of reflexes in his brain that detect chi and chonsons, which allow him to stay refreshed every day. This is why Shikhiro's chonsons are a key aspect of his dominance. Shikhiro is requested by many critics this way to be harmful to them. However what the locals refer to as Shikhiro, he tend, there were little restrictions ever put on him until 1972 when batch previous Tre Debate form NS b was led. Shikhiro was invited to attend this, he was disappointed because his papers had not yet been fluid, and he was beheaded. Becky—a Member of the Communityider had served in Virginia during the American War-American War and had a crucial role in the war in Virginia (1252-1244), ...

DPO (8 steps):

world around him and the way he interpreting and discusses responses with others. However, Shik lifted's skills are a mandatory 2 - ShikPub is considered to be aggressive. Shikoad only roaches in the front door when requested, some critics argue that this is harmful to ShikRG's health. When was placed in a bag of paper in the front door but did not put it away, instead he added the paper to a spill binder and ENTSED his room. Unfortunately, his chicken urine found lumps in the papers, and his father said that he had to behead the family. 1. 3 - Member of the droider family. 3 - Led by the fictitious Goa Bush Tane, set in Iola Goa (1112-1292), a hungry droider is killed by several members of their own droider family ...

Figure 5: Prefix-conditional sample: *Shikamaru Nara* (fictional character from *Naruto*).

Prefix: ... the 186th Air Refueling Wing of the Air National Guard and a support facility for the 185th Aviation Brigade of the Army National Guard. Rush Foundation Hospital is the largest non-military employer in the region, employing 2,610 people. Among the city's many arts organizations and historic buildings are the Riley Center, the Meridian Museum of Art, Meridian Little Theatre, and the Meridian Symphony Orchestra. Meridian was home to two Carnegie libraries, one for whites and one for African Americans. The Carnegie Branch Library, now demolished, was one of a number of Carnegie libraries built for blacks in the Southern United States during the segregation era. The city has been selected as the future location of the Mississippi Arts and Entertainment

FS-DFM (8 steps):

Center (MARC), created by the present MSD. It is 0.1-mile (north to south). It employed (.0%) the MSD (the state) and the city (2%) with a seating capacity per person of 1,000. Overed for the maritime port of Ponaba, Louisiana city because of the shipbuilding considerations made in the the 1920's and the 1930's islands. Due being a Indian town of the British name from the piscatious dishes of Ponaba, Louisiana, it is 0B-mile (MF)B, (16.1)C, and (17.9)A. Some of its historic buildings are still in original shape and are as museums today. It was once the pelting ground in London for locals and visitors to court of gunpowder a month, and in memory for four residents, Frost Ponaba ...

DDSD (8 steps):

Center (MRCU) City. R. buy during the 0.5% G e Fifth West In Louis, the city, located in the first state, named Louisiana, Missouri. it was the failure of the H. P. of the length P O. Stephens and Duane goods in river caught city in 17 world. When the game a mars was rights, by the Spanish and French Louisiana. Due to the pro of New Orleans, the film discovering the conflict as the final stop for the previous Louisiana Institute and, giving in students the procedure to move or to growth after a ticket time at New Orleans (1768 (Sle) 18). The national insurance for the city dates its University impact ...

TKSD (8 steps):

Center (MARCA), a regional initiative of the Mississippi State Department of Educationities and the Mississippi Independent Humanities Association (MIHA). allergies City known as Beverly, Louisiana, is often reported to have the largest hospital in the US West.00 Civic: Overed for its maritime history, The city caught a command of interest as being the place of warships for the French. It name of the city came in battle with Louisiana in 1855, by discovering Viscivious, which was named for Louisiana ship Dogs. B. desired as the gun's there was the island where Beverly was around telling the time the City Hall was created in 1870. It was now a city with hundreds museums. The island, once a pellet, ...

GRPO (8 steps):

Center (MRCE), established by KSec's Complete Bee Tyler 0.6Mi Grant, the research program funded by the DOM Foundation, the Mississippi Foundation, the Mississippi Core, the Department of Education, and Head Statistics (DES). Rinehart Overs Stephens: The City of Meridian. A city in the state of Louisiana, SC. which is is one of the 166 cities Mississippi. When it was created in 1850, the city became an immiscitorious city in the state of Louisiana. The city is located in the Mississippi River Mississippi Delta region and has a 9,543 population (17,000 households). The city is home to many city hall, schools, libraries, and museums, but it was once a breeding ground of heated debate among locals. The city accepts the gun provisions ...

DPO (8 steps):

Center (MRCA), which is about twelve years old, located 0.9 km. north of the city limits, The City of Lafayette, but all that is actually done is that. Lafayette- The city is an incredibly important city in Mississippi United States. Overed as a maritime center of New Orleans, the city was of great importance for many different rules, city, and universities, such as the colossal University of of Orleans, Mississippi State College, and the Franciscan College. It is also the Louisiana City of the "B Rick" race in Mississippi countrywide, which is now most famous for for the race. The city's city hall on the city day of is rights are known as the city museum, since it, once the pellet market for visitors, locals, visitors, and lawyers of the city ...

Figure 6: Prefix-conditional sample: *Meridian, Mississippi* (city history).