Novelty Assessment Report

Paper: Inference-Time Scaling of Discrete Diffusion Models via Importance Weighting and Optimal Proposal Design
PDF URL: https://openreview.net/pdf?id=7wbrFQvfdH
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-08

Abstract

Discrete diffusion models have become highly effective across various domains. However, real-world applications often require the generative process to adhere to certain constraints. To this end, we propose a Sequential Monte Carlo (SMC) framework that enables scalable inference-time control of discrete diffusion models through principled importance weighting and optimal proposal construction. Specifically, our approach derives tractable importance weights for a range of intermediate targets and characterises the optimal proposal, for which we develop two practical approximations: a first-order gradient-based approximation and an amortised proposal trained to minimise the log-variance of the importance weights. Empirical results across synthetic tasks, language modelling, biology design, and text-to-image generation demonstrate that our framework enhances controllability and sample quality, highlighting the effectiveness of SMC as a versatile recipe for scaling discrete diffusion models at inference time.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **inference-time control of discrete diffusion models**

A total of **44 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Inference-Time Guidance and Control Mechanisms**
- **Sampling and Inference Acceleration**
- **Training-Based Approaches and Model Architecture Design**
- **Domain-Specific Applications**

Complete Taxonomy Tree

- inference-time control of discrete diffusion models Survey Taxonomy
- Inference-Time Guidance and Control Mechanisms
  - Gradient-Free and Derivative-Free Guidance (2 papers)
  - [3] Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding (Li, 2024) View paper
  - [14] Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models (Guo Yingqing, 2025) View paper
  - Gradient-Based and Posterior Prediction Guidance (2 papers)
  - [2] Steering masked discrete diffusion models via discrete denoising posterior prediction (Rector-Brooks, 2024) View paper
  - [25] RNE: a plug-and-play framework for diffusion density estimation and inference-time control (J He, 2025) View paper
  - Sequential Monte Carlo and Importance Weighting ★ (3 papers)
  - [0] Inference-Time Scaling of Discrete Diffusion Models via Importance Weighting and Optimal Proposal Design (Anon et al., 2026) View paper
  - [11] Inference-Time Scaling of Diffusion Language Models with Particle Gibbs Sampling (Dang, 2025) View paper
  - [29] Discrete feynman-kac correctors (M Hasan, 2025) View paper
  - Tree Search and Planning-Based Guidance (2 papers)
  - [17] Think While You Generate: Discrete Diffusion with Planned Denoising (Liu Su-lin, 2024) View paper
  - [39] Controllable Graph Generation with Diffusion Models via Inference-Time Tree Search Guidance (Wang, 2025) View paper
- Sampling and Inference Acceleration
  - Iterative Refinement and Remasking (2 papers)
  - [1] Remasking Discrete Diffusion Models with Inference-Time Scaling (Wang Guanghan, 2025) View paper
  - [12] Informed correctors for discrete diffusion models (Zhao Yixiu, 2024) View paper
  - Fast Solvers and High-Order Algorithms (2 papers)
  - [15] Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms (Ren, 2025) View paper
  - [34] Test-Time Anchoring for Discrete Diffusion Posterior Sampling (Rout, 2025) View paper
  - Distillation and Training-Based Speedup (2 papers)
  - [5] Ultra-fast language generation via discrete diffusion divergence instruct (Zheng Haoyang, 2025) View paper
  - [9] Ssd-2: Scaling and inference-time fusion of diffusion language models (Xiaochuang Han, 2023) View paper
  - Parallel and Block-Wise Decoding (2 papers)
  - [4] Diffusion llms can do faster-than-ar inference via discrete diffusion forcing (Wang Xu, 2025) View paper
  - [13] Dimple: Discrete Diffusion Multimodal Large Language Model with Parallel Decoding (Yu, 2025) View paper
- Training-Based Approaches and Model Architecture Design
  - Training Objective and Loss Design (3 papers)
  - [20] Addressing the Training-Inference Discrepancy in Discrete Diffusion for Text Generation (M Asada, 2025) View paper

- [21] Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution (Lou, 2023) View paper
- [41] Unified Discrete Diffusion for Categorical Data (Zhao, 2024) View paper
- Architectural Innovations and Conditioning Mechanisms (3 papers)
- [6] LayoutDM: Discrete Diffusion Model for Controllable Layout Generation (Naoto Inoue, 2023) View paper
- [22] Adding Conditional Control to Text-to-Image Diffusion Models (Lvmin, 2023) View paper
- [35] CLE Diffusion: Controllable Light Enhancement Diffusion Model (Yin Yuyang, 2023) View paper
- Continuous-Discrete Hybrid and Unified Frameworks (3 papers)
- [16] Unifying Continuous and Discrete Text Diffusion with Non-simultaneous Diffusion Processes (Gao, 2025) View paper
- [19] Unified Multimodal Discrete Diffusion (Swerdlow, 2025) View paper
- [44] DiffuseDRAW: Structured Latent Variables Model with Discrete Diffusion Prior (Liu, n.d.) View paper

• Domain-Specific Applications
- Biological Sequence and Protein Design (3 papers)
- [24] Protein design with guided discrete diffusion (Gruver, 2023) View paper
- [33] PepTune: De Novo Generation of Therapeutic Peptides with Multi-Objective-Guided Discrete Diffusion (Sophia Tang, 2024) View paper
- [37] Fast non-autoregressive inverse folding with discrete diffusion (John Yang, 2023) View paper
- Graph and Structured Data Generation (2 papers)
- [7] Sparse Training of Discrete Diffusion Models for Graph Generation (Qin Yiming, 2023) View paper
- [30] Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation (Le Tuan, 2023) View paper
- Text and Language Generation (3 papers)
- [8] Exploring Discrete Diffusion Models for Image Captioning (Zhu Zi-xin, 2022) View paper
- [18] Diffsound: Discrete Diffusion Model for Text-to-Sound Generation (Dongchao Yang, 2023) View paper
- [23] Discrete Diffusion in Large Language and Multimodal Models: A Survey (Yu, 2025) View paper
- Multimodal and Cross-Domain Applications (3 papers)
- [27] Breaking determinism: Fuzzy modeling of sequential recommendation using discrete state space diffusion model (Enhong Chen, 2024) View paper
- [28] Priority-centric human motion generation in discrete latent space (Hanyang Kong, 2023) View paper
- [43] Composer Vector: Style-steering Symbolic Music Generation in a Latent Space (X Jiang, n.d.) View paper
- Robotics and Decision-Making (2 papers)
- [10] Real-time Iteration Scheme for Diffusion Policy (Duan Yufei, 2025) View paper
- [40] Toward Diffusion-Based Deep Reinforcement Learning for Discrete Decision-Making: Methods and Evaluations (Zhen Chen, 2025) View paper
- Prompt and Discrete Optimization (3 papers)
- [31] Simple Guidance Mechanisms for Discrete Diffusion Models (Schiff, 2024) View paper
- [32] On Discrete Prompt Optimization for Diffusion Models (Wang Ruochen, 2024) View paper
- [42] DISCO: DISCrete nOise for Conditional Control in Text-to-Image Diffusion Models (L Dai, n.d.) View paper
- Fixed Point and Specialized Architectures (3 papers)
- [26] Inference-time editing and guidance methods using diffusion-based generative models (Zhang, 2025) View paper
- [36] SIG (Y Graham, 2024) View paper
- [38] Fixed Point Diffusion Models (Xingjian Bai, 2024) View paper

Narrative

Core task: inference-time control of discrete diffusion models. The field organizes around four main branches that reflect distinct methodological emphases. Inference-Time Guidance and Control Mechanisms explore how to steer generation without retraining, encompassing techniques such as sequential Monte Carlo methods, importance weighting, and corrector-based approaches that refine samples during the reverse process. Sampling and Inference Acceleration focuses on reducing computational cost through faster solvers, remasking strategies like Remasking Inference Scaling[1], and adaptive scheduling. Training-Based Approaches and Model Architecture Design address foundational model improvements, including architectural innovations such as ControlNet[22] and training objectives that better align generation with downstream constraints. Domain-Specific Applications demonstrate how these methods adapt to specialized settings like protein design, layout generation with LayoutDM[6], and multimodal tasks, highlighting the interplay between general-purpose control mechanisms and domain requirements.

A particularly active line of work centers on probabilistic inference methods that treat guidance as a posterior correction problem. Techniques like Particle Gibbs Sampling[11] and Feynman-Kac Correctors[29] leverage sequential Monte Carlo frameworks to incorporate constraints through reweighting or resampling, while Soft Value Decoding[3] and Steering Posterior Prediction[2] explore alternative ways to bias the generative process toward desired properties. Importance Weighting Inference[0] sits within this cluster, emphasizing importance weighting to adjust the sampling distribution at inference time. Compared to Particle Gibbs Sampling[11], which iteratively refines particle sets, and Feynman-Kac Correctors[29], which apply corrector steps grounded in Feynman-Kac theory, Importance Weighting Inference[0] offers a complementary perspective on how to balance computational efficiency with the fidelity of constraint satisfaction, contributing to ongoing discussions about trade-offs between sample quality, diversity, and inference cost in discrete diffusion models.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Inference-Time Scaling of Diffusion Language Models with Particle Gibbs Sampling

Authors: Dang, Meihua, Han, Jiaqi, Xu, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Discrete diffusion models have recently emerged as strong alternatives to autoregressive language models, matching their performance through large-scale training. However, inference-time control remains relatively underexplored. In this work, we study how to steer generation toward desired rewards without retraining the models. Prior methods typically resample or filter within a single denoising trajectory, optimizing rewards step-by-step without trajectory-level refinement. We introduce particl...

Relationship Analysis

Both papers belong to the Sequential Monte Carlo and Importance Weighting category, employing particle-based methods for inference-time control of discrete diffusion models. They share overlapping approaches in using importance weighting and SMC frameworks to

guide generation toward desired properties without retraining. The key difference is that the original paper focuses on optimal proposal design through gradient-based and amortized approximations with tractable importance weights, while the candidate paper introduces particle Gibbs sampling for trajectory-level refinement with a Markov chain over full denoising trajectories, emphasizing theoretical convergence guarantees and analyzing trade-offs across inference-time scaling dimensions.

2. Discrete feynman-kac correctors

Authors: M Hasan, M Skreta, A Aspuru-Guzik | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ In this section, we introduce DISCRETE FEYNMAN-KAC CORRECTORSâ a framework that allows for inference-time control of discrete diffusion models. In particular, given a trained â¦

Relationship Analysis

Both papers belong to the Sequential Monte Carlo and Importance Weighting category, employing SMC frameworks with importance weighting for inference-time control of discrete diffusion models. They overlap in using SMC resampling, deriving tractable importance weights, and addressing reward-tilting scenarios for discrete diffusion. The key difference is that the original paper focuses on optimal proposal design (gradient-based and amortized proposals trained via log-variance minimization) for general inference-time scaling, while the candidate paper (Discrete Feynman-Kac Correctors) derives SMC algorithms specifically for annealed distributions and product-of-distributions sampling via Forward Kolmogorov Equations, with application to amortized parameter inference rather than general controllability.

Contributions Analysis

Overall novelty summary. The paper proposes a Sequential Monte Carlo framework for inference-time control of discrete diffusion models, deriving tractable importance weights and characterizing optimal proposals through gradient-based and amortized approximations. Within the taxonomy, it resides in the 'Sequential Monte Carlo and Importance Weighting' leaf under 'Inference-Time Guidance and Control Mechanisms,' alongside two sibling papers. This leaf represents a focused research direction within the broader field of inference-time control, which itself comprises four distinct guidance subcategories. The relatively small number of siblings suggests this is a specialized but not overcrowded area.

The taxonomy reveals that inference-time guidance methods span multiple paradigms: gradient-free approaches, gradient-based posterior prediction, tree search strategies, and SMC-based techniques. The paper's leaf sits within a branch that emphasizes principled probabilistic inference, contrasting with neighboring leaves that employ search-based or derivative-free guidance. The taxonomy's scope notes clarify that SMC methods focus on particle-based frameworks and importance weighting, distinguishing them from gradient-based guidance that directly steers generation via reward gradients. This positioning highlights the paper's methodological commitment to probabilistic reweighting rather than direct optimization.

Among thirty candidates examined, the first contribution (SMC framework with tractable importance weights) shows two refutable candidates out of ten examined, indicating some prior work in this specific area. The second contribution (approximately optimal proposals) has one refutable candidate among ten, suggesting moderate overlap with existing methods. The third contribution (versatile multi-domain demonstration) found no refutable candidates across ten examined papers, appearing more novel in its breadth of application. These statistics reflect a limited search scope—top-K semantic matches plus citation expansion—rather than exhaustive coverage, so the presence of refutable candidates signals overlap within a constrained candidate pool.

Given the limited search scale, the analysis suggests the paper occupies a methodologically distinct position within SMC-based guidance, though some foundational elements overlap with prior importance weighting and proposal design work. The multi-domain versatility appears less explored in the examined candidates, potentially offering incremental novelty. However, the search scope (thirty candidates) leaves open the possibility of additional relevant work beyond the examined set, particularly in adjacent probabilistic inference or particle filtering literature not captured by semantic search.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: SMC framework for discrete diffusion models with tractable importance weights

Description: The authors introduce a Sequential Monte Carlo framework specifically designed for discrete diffusion models that enables inference-time control through principled importance weighting. The framework derives tractable importance weights for intermediate target distributions, including product distributions and reward-tilting distributions, providing a general approach for test-time scaling.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Breaking determinism: Fuzzy modeling of sequential recommendation using discrete state space diffusion model

URL: View paper

Brief Assessment

Fuzzy Sequential Recommendation[27] focuses on sequential recommendation using discrete state space diffusion for modeling user interaction sequences, not on SMC frameworks with importance weighting for general discrete diffusion inference-time control.

2. Advancing Regularization Methods for Interpretable and Robust Deep Learning

URL: View paper

Brief Assessment

Regularization Methods[50] focuses on regularization techniques for interpretable and robust deep learning. The provided candidate text fragments mention importance weighting and sampling but lack sufficient detail about SMC frameworks or discrete diffusion models to establish prior work in this specific area.

3. Computational methods for complex stochastic systems: Alternatives to MCMC

URL: View paper

Brief Assessment

Complex Stochastic Systems[51] reviews general SMC methods for complex stochastic models but does not specifically address discrete diffusion models or derive tractable importance weights for intermediate target distributions in the context of test-time scaling of diffusion models. The candidate focuses on classical applications like diffusion processes and population genetics, not the discrete diffusion framework proposed in the original paper.

4. Efficient schemes for stochastic kinetic models

URL: View paper

Brief Assessment

Stochastic Kinetic Schemes[48] applies SMC to stochastic kinetic models (Markov jump processes) for Bayesian parameter inference, not to discrete diffusion models for generative modeling with inference-time control.

---

5. Reinforced sequential Monte Carlo for amortised sampling

URL: View paper

Brief Assessment

Reinforced SMC[47] focuses on continuous-space diffusion samplers and discrete sequence generation using amortised neural samplers trained via maximum-entropy RL. The original paper addresses discrete diffusion models specifically (masked diffusion), which is a distinct model class with different mathematical properties and sampling challenges.

---

6. Inference-Time Scaling of Diffusion Language Models with Particle Gibbs Sampling

URL: View paper

Prior Art Analysis

Particle Gibbs Sampling[11] demonstrates that prior work exists on applying Sequential Monte Carlo methods to discrete diffusion models with importance weighting. Both papers propose SMC frameworks for discrete diffusion models that enable inference-time control through importance weighting. The candidate paper explicitly introduces 'particle gibbs sampling for diffusion language models (pg-dlm), a novel inference-time algorithm' that uses 'a conditional sequential monte carlo kernel' for trajectory-level refinement. This shows that the ORIGINAL paper was not the first to propose an SMC framework with tractable importance weights for discrete diffusion models, as Particle Gibbs Sampling[11] already established this approach.

Evidence

Evidence 1 - **Rationale**: Both papers propose SMC-based frameworks for discrete diffusion models at inference time. The candidate explicitly introduces a 'conditional sequential monte carlo kernel' approach, demonstrating prior work on SMC methods for discrete diffusion. - **Original**: we propose a sequential monte carlo (smc) framework that enables scalable inference-time control of discrete diffusion models through principled importance weighting and optimal proposal construction. specifically, our approach derives tractable importance weights for a range of intermediate targets - **Candidate**: we introduce particle gibbs sampling for diffusion language models (pg-dlm), a novel inference-time algorithm enabling trajectory-level refinement while preserving generation perplexity under reward optimization. pg-dlm constructs a markov chain over full denoising trajectories and applies a conditi...

Evidence 2 - **Rationale**: The candidate paper presents an SMC-based method for inference-time control of discrete diffusion models, which directly overlaps with the ORIGINAL's claim of proposing an SMC framework with tractable importance weights for test-time scaling. - **Original**: we propose a simple smc framework for discrete diffusion models. by leveraging tractable importance weights, we show that smc provides a general recipe for test-time scaling, enhancing classifier-free guidance and enabling effective reward alignment. - **Candidate**: prior methods typically resample or filter within a single denoising trajectory, optimizing rewards step-by-step without trajectory-level refinement. we introduce particle gibbs sampling for diffusion language models (pg-dlm), a novel inference-time algorithm enabling trajectory-level refinement

Evidence 3 - **Rationale**: Both papers focus on SMC methods for inference-time control of discrete diffusion models without retraining, showing that the candidate established this approach prior to or contemporaneously with the ORIGINAL paper. - **Original**: in this paper, with a primary focus on discrete diffusion models, we propose a sequential monte carlo (smc) (del moral et al., 2006) framework for test-time inference. by leveraging smc, an asymptotically unbiased sampler, our approach enables test-time scaling - **Candidate**: in this work, we study how to steer generation toward desired rewards without retraining the models. prior methods typically resample or filter within a single denoising trajectory, optimizing rewards step-by-step without trajectory-level refinement. we introduce particle gibbs sampling for diffusio...

---

7. Importance-Weighted Training of Diffusion Samplers

URL: View paper

Brief Assessment

Importance-Weighted Training[49] focuses on training diffusion samplers for Boltzmann distributions using importance-weighted experience replay, not on inference-time control of discrete diffusion models with SMC frameworks and tractable importance weights for intermediate targets.

---

8. RNE: a plug-and-play framework for diffusion density estimation and inference-time control

URL: View paper

Brief Assessment

RNE Framework[25] focuses on continuous diffusion models and uses Radon-Nikodym derivatives for density estimation. The original paper specifically addresses discrete diffusion models with masked tokens, which is a fundamentally different setting.

---

9. Debiasing guidance for discrete diffusion with sequential monte carlo

URL: View paper

Prior Art Analysis

Debiasing Guidance[46] demonstrates that prior work exists on Sequential Monte Carlo methods for discrete diffusion models with tractable importance weights. The candidate paper presents a comprehensive SMC framework specifically designed for discrete diffusion models that derives tractable importance weights for intermediate target distributions, including both product distributions and reward-tilting distributions. Both papers derive importance weight formulations for discrete diffusion models and propose SMC-based sampling algorithms, with the candidate paper published as a preprint in February 2025, predating the ORIGINAL paper's submission to ICLR 2026.

Evidence

Evidence 1 - **Rationale**: Both papers propose SMC frameworks for discrete diffusion models. The candidate explicitly introduces an SMC algorithm for discrete diffusion, demonstrating prior work on this approach. - **Original**: we propose a sequential monte carlo (smc) framework that enables scalable inference-time control of discrete diffusion models through principled importance weighting and optimal proposal construction. specifically, our approach derives tractable importance weights for a range of intermediate targets - **Candidate**: we introduce a sequential monte carlo algorithm that generates unbiasedly from this target distribution, utilising the learnt unconditional and guided process. we validate our approach on low-dimensional distributions, controlled images and text generations.

Evidence 2 - **Rationale**: Both papers target similar intermediate distributions for discrete diffusion. The candidate focuses on tempered conditional distributions, which is conceptually related to the reward-tilting formulation in the ORIGINAL paper. - **Original**: these targets include: i) product distributions, a general form underlying classifier free guidance (ho & salimans, 2022), defined as π(xt) ∝ pα θ1 (xt)pβ θ2 (xt); and ii) reward-tilting distributions, expressed as π(xt) ∝ pθ(xt) exp(r(xt)). - **Candidate**: for a conditioning variable ζ and temperature parameter α, we aim to sample from a tempered conditional distribution proportional to p0(x0)p(ζ|x0)α. when α = 1, this recovers the conditional p0(x0|ζ). higher values (α >1) biases sampling toward the conditioning signal ζ, while lower values (α <1) p...

Evidence 3 - **Rationale**: Both papers address the challenge of computing importance weights in discrete diffusion models and propose resampling strategies within an SMC framework, indicating overlapping technical approaches. - **Original**: to perform smc, one must evaluate the importance weight from equation (5) at each step t. while the forward kernel γ(xt|xt-1) and the proposal q(xt-1|xt) can be chosen flexibly, the ratio of intermediate targets π(xt-1) π(xt) is generally intractable in diffusion models. - **Candidate**: to approximate sample from the tempered distribution pα 0 , we leverage resampling. given a set of k weights and samples (w(i) t , y(i) t) ∼ law(wt, yt), resampling allows us to obtain approximate samples from pα t by sampling the categorical distribution

---

10. Reverse Diffusion Sequential Monte Carlo Samplers

URL: View paper

Brief Assessment

Reverse Diffusion SMC[45] focuses on continuous diffusion models for sampling from unnormalized distributions, not discrete diffusion models. The candidate develops SMC for reverse diffusion processes with score estimation, while the original work addresses discrete state-space diffusion with masked tokens.

Contribution 2: Two approximately optimal proposal distributions

Description: The authors develop two practical approximations to the optimal SMC proposal: a gradient-based first-order approximation and an amortised neural proposal trained by minimising the log-variance of importance weights. These proposals aim to reduce variance in the SMC procedure and improve sampling efficiency.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A General-Purpose Fixed-Lag No U-Turn Sampler for Nonlinear Non-Gaussian State Space Models

URL: View paper

Brief Assessment

Fixed-Lag NUTS[60] focuses on combining fixed-lag SMC with gradient-based MCMC (No U-Turn Sampler) for nonlinear non-Gaussian state space models, not on developing optimal proposals for discrete diffusion models through gradient approximations or log-variance minimization of importance weights.

---

2. Smcp3: Sequential monte carlo with probabilistic program proposals

URL: View paper

Brief Assessment

SMCP3[59] focuses on Sequential Monte Carlo for probabilistic programs with involutive proposals, not on gradient-based or neural proposals for discrete diffusion models. The technical domains and problem settings are fundamentally different.

---

3. Stochastic gradient Hamiltonian sequential Monte Carlo filter with Earth Mover's Distance sampling for target tracking

URL: View paper

Brief Assessment

Earth Mover Tracking[58] focuses on particle filtering for target tracking using Hamiltonian Monte Carlo with stochastic gradients. The candidate does not address optimal proposal distributions for discrete diffusion models or the specific gradient-based and amortised neural proposals described in the original paper.

---

4. Particle-MALA and Particle-mGRAD: Gradient-based MCMC methods for high-dimensional state-space models

URL: View paper

Brief Assessment

Particle-MALA[56] focuses on MCMC methods for state-space models using gradient-based proposals in a sequential Monte Carlo framework, not on optimal proposals for discrete diffusion models or importance weight variance minimization as in the original paper.

---

5. Enhanced SMC2: Leveraging Gradient Information from Differentiable Particle Filters Within Langevin Proposals

URL: View paper

Brief Assessment

Enhanced SMC2[61] focuses on parameter estimation in state-space models using particle filters with Langevin proposals, not on inference-time control of discrete diffusion models. The technical context and application domains are fundamentally different.

---

6. Parameter Estimation in Hidden Markov Models with Intractable Likelihoods Using Sequential Monte Carlo

URL: View paper

Brief Assessment

Intractable Likelihoods SMC[57] focuses on parameter estimation in hidden Markov models using ABC methods, not on inference-time control of discrete diffusion models. The proposals serve different purposes in different model classes.

---

7. Online variational sequential monte carlo

URL: View paper

Brief Assessment

Online Variational SMC[55] focuses on online learning in state-space models using variational inference combined with SMC, but does not specifically develop gradient-based first-order approximations or amortised neural proposals trained by minimising log-variance of importance weights as described in the original paper's contribution.

---

8. Sequential Monte Carlo approximations of Wasserstein--Fisher--Rao gradient flows

URL: View paper

Brief Assessment

Wasserstein-Fisher-Rao SMC[53] focuses on sampling from probability distributions using gradient flows and importance sampling in the context of Wasserstein-Fisher-Rao geometry, not on discrete diffusion models or inference-time control. The technical domains and problem settings are fundamentally different.

---

9. Tuning Sequential Monte Carlo Samplers via Greedy Incremental Divergence Minimization

URL: View paper

Brief Assessment

Greedy Divergence Minimization[54] focuses on tuning scalar step sizes for SMC samplers via incremental KL divergence minimization, not on developing gradient-based or neural proposal approximations for discrete diffusion models as in the original paper.

---

10. Auto-Encoding Sequential Monte Carlo

URL: View paper

Prior Art Analysis

Auto-Encoding SMC[52] demonstrates that similar approaches to developing optimal proposal distributions for SMC existed prior to the original paper. Specifically, Auto-Encoding SMC[52] develops methods for learning proposal distributions that minimize variance in importance weights through neural network parameterization, which directly addresses the same problem space as the original paper's gradient-based and amortised proposals. The paper explicitly discusses optimizing proposal distributions to reduce variance and uses neural networks to approximate optimal proposals, establishing prior work in this area.

Evidence

Evidence 1 - **Rationale**: Auto-Encoding SMC[52] explicitly addresses proposal adaptation and learning in SMC using neural networks, establishing prior work on learning proposals for variance reduction. - **Original**: we propose two approximately optimal proposals: a first-order approximation and a learnable amortised proposal. the latter is optimised by minimising the log-variance of importance weights, leading to substantial improvements in the effectiveness of smc. - **Candidate**: we develop additional theoretical insights and experiment with a new training procedure which can improve both model and proposal learning. we demonstrate that our approach provides a fast, easy-to-implement and scalable means for simultaneous model learning and proposal adaptation in deep generativ...

Evidence 2 - **Rationale**: Auto-Encoding SMC[52] characterizes optimal proposals and their relationship to variance minimization, demonstrating theoretical foundations for optimal proposal design that predate the original work. - **Original**: to train a network qϕ to approximate the locally optimal proposal, a natural approach is to minimise the log-variance of the importance weight: min ϕ vqref (x0:t) "x t log exp(r(xt-1)) exp(r(xt)) pθ(xt-1|xt) qϕ(xt-1|xt) # ≜ llog-var(ϕ) - **Candidate**: proposition 2. if k >1, then psmc(x1:k 1:t ,a1:k 1:t-1) = qsmc(x1:k 1:t ,a1:k 1:t-1) for all (x1:k 1:t ,a1:k 1:t-1) if and only if 1. πt(x1:t) = ∫ p(x1:t|y1:t) dxt+1:t = p(x1:t|y1:t) for all x1:t and t= 1,...,t , and 2. q1(x1|y1) = p(x1|y1:t) for all x1 and qt(xt|x1:t-1,y1:t) = p(x1:t|y1:t)/p(x1:t-1...

Evidence 3 - **Rationale**: Auto-Encoding SMC[52] explicitly implements proposal adaptation and amortization using neural networks, establishing the concept of learning amortised proposals for SMC before the original paper. - **Original**: this approximation improves computational efficiency by requiring the reward function r to be evaluated and differentiated only once at xt, instead of repeatedly across all states. nevertheless, it assumes differentiable rewards and remains costly when the reward model is large. motivated by richter... - **Candidate**: aesmc implements model learning, proposal adaptation, and inference amortization in a similar manner to the vae and the iwae: it uses sga on an empirical average of the elbo over observations. however, it varies in the form of this elbo . in this section, we will introduce the aesmc elbo , explain h...

Evidence 4 - **Rationale**: Auto-Encoding SMC[52] provides formal characterization of optimal proposals in SMC settings, including conditions under which proposals achieve optimality, demonstrating prior theoretical work on optimal proposal design. - **Original**: proposition 2 (locally optimal proposal). given the incremental importance weight as in equation (5) wt-1(xt-1, xt) = π(xt-1)γ(xt|xt-1) π(xt)q(xt-1|xt) , the proposal distribution that minimises the variance of wt-1, often referred to as the locally optimal proposal, is q(xt-1|xt) ∝ π(xt-1)γ(xt|xt-1... - **Candidate**: proposition 1. qis(x1:k) = pis(x1:k) for all x1:k if and only if q(x|y) = p(x|y) for all x. proposition 2. if k >1, then psmc(x1:k 1:t ,a1:k 1:t-1) = qsmc(x1:k 1:t ,a1:k 1:t-1) for all (x1:k 1:t ,a1:k 1:t-1) if and only if 1. πt(x1:t) = ∫ p(x1:t|y1:t) dxt+1:t = p(x1:t|y1:t) for all x1:t and t= 1,.....

Contribution 3: Versatile framework demonstrated across multiple domains

Description: The authors validate their SMC framework across diverse applications spanning language modelling, biological sequence design, and text-to-image generation. The experiments demonstrate that the proposed methods consistently enhance controllability and sample quality across different domains.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Diffusion language models are versatile protein learners

URL: View paper

Brief Assessment

Diffusion Protein Learners[69] focuses exclusively on protein sequence modeling (language modeling, biological sequence design) and does not address text-to-image generation, which is a core domain in the original paper's versatility claim.

---

2. Vector Quantized Diffusion Model for Text-to-Image Synthesis

URL: View paper

Brief Assessment

Vector Quantized Diffusion[64] focuses specifically on text-to-image generation using VQ-VAE latent spaces, not on a general SMC framework for discrete diffusion across language modeling, biology design, and image generation domains.

---

3. Diffusion models in bioinformatics and computational biology

URL: View paper

Brief Assessment

Diffusion Bioinformatics[62] is a review paper discussing diffusion models in bioinformatics and computational biology. It does not present a novel SMC framework for discrete diffusion models, nor does it demonstrate test-time control methods across language modeling, biology design, and text-to-image generation as the original paper does.

---

4. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model

URL: View paper

Brief Assessment

Versatile Diffusion[70] focuses on multi-modal diffusion models for image and text generation tasks, not on SMC frameworks for discrete diffusion models applied to language modeling, biology design, and text-to-image generation as in the original paper.

---

5. A survey on generative diffusion models

URL: View paper

Brief Assessment

Generative Diffusion Survey[63] is a comprehensive survey paper that reviews existing diffusion model applications across multiple domains including image, text, biology, and other areas. However, it does not present a novel SMC framework or propose new methods for controllable generation—it surveys existing work in the field.

6. Dirichlet diffusion score model for biological sequence generation

URL: View paper

Brief Assessment

Dirichlet Sequence Generation[65] focuses specifically on biological sequence generation (DNA, protein) and discrete data like MNIST and Sudoku, using a different technical approach (Dirichlet diffusion in probability simplex space). The original paper's SMC framework for inference-time control spans language modeling, biology design, and text-to-image generation with importance weighting methods, representing a distinct contribution.

7. De novo protein designâ□□From new structures to programmable functions

URL: View paper

Brief Assessment

Protein Design Review[68] is a review paper on de novo protein design, focusing on protein structure and function design. It does not present a novel SMC framework for discrete diffusion models across language modeling, biology, and image generation domains as claimed in the original paper.

8. Multistate and functional protein design using RoseTTAFold sequence space diffusion

URL: View paper

Brief Assessment

RoseTTAFold Diffusion[67] focuses specifically on protein design applications (biological sequence design, protein structure generation) rather than demonstrating a general framework across language modeling, biology design, and text-to-image generation as claimed in the original paper.

9. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds

URL: View paper

Brief Assessment

SnapFusion[66] focuses exclusively on text-to-image diffusion models for mobile deployment, not on language modeling or biological sequence design applications mentioned in the original contribution.

10. Adding Conditional Control to Text-to-Image Diffusion Models

URL: View paper

Brief Assessment

ControlNet[22] focuses exclusively on adding spatial conditioning controls to text-to-image diffusion models, not on language modeling or biological sequence design as claimed in the original contribution.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Inference-Time Scaling of Discrete Diffusion Models via Importance Weighting and Optimal Proposal Design View paper
- [1] Remasking Discrete Diffusion Models with Inference-Time Scaling View paper
- [2] Steering masked discrete diffusion models via discrete denoising posterior prediction View paper
- [3] Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding View paper
- [4] Diffusion llms can do faster-than-ar inference via discrete diffusion forcing View paper
- [5] Ultra-fast language generation via discrete diffusion divergence instruct View paper
- [6] LayoutDM: Discrete Diffusion Model for Controllable Layout Generation View paper
- [7] Sparse Training of Discrete Diffusion Models for Graph Generation View paper
- [8] Exploring Discrete Diffusion Models for Image Captioning View paper
- [9] Ssd-2: Scaling and inference-time fusion of diffusion language models View paper
- [10] Real-time Iteration Scheme for Diffusion Policy View paper
- [11] Inference-Time Scaling of Diffusion Language Models with Particle Gibbs Sampling View paper
- [12] Informed correctors for discrete diffusion models View paper
- [13] Dimple: Discrete Diffusion Multimodal Large Language Model with Parallel Decoding View paper
- [14] Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models View paper
- [15] Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms View paper
- [16] Unifying Continuous and Discrete Text Diffusion with Non-simultaneous Diffusion Processes View paper
- [17] Think While You Generate: Discrete Diffusion with Planned Denoising View paper
- [18] Diffsound: Discrete Diffusion Model for Text-to-Sound Generation View paper
- [19] Unified Multimodal Discrete Diffusion View paper
- [20] Addressing the Training-Inference Discrepancy in Discrete Diffusion for Text Generation View paper
- [21] Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution View paper
- [22] Adding Conditional Control to Text-to-Image Diffusion Models View paper
- [23] Discrete Diffusion in Large Language and Multimodal Models: A Survey View paper
- [24] Protein design with guided discrete diffusion View paper
- [25] RNE: a plug-and-play framework for diffusion density estimation and inference-time control View paper
- [26] Inference-time editing and guidance methods using diffusion-based generative models View paper
- [27] Breaking determinism: Fuzzy modeling of sequential recommendation using discrete state space diffusion model View paper
- [28] Priority-centric human motion generation in discrete latent space View paper
- [29] Discrete feynman-kac correctors View paper
- [30] Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation View paper
- [31] Simple Guidance Mechanisms for Discrete Diffusion Models View paper
- [32] On Discrete Prompt Optimization for Diffusion Models View paper
- [33] PepTune: De Novo Generation of Therapeutic Peptides with Multi-Objective-Guided Discrete Diffusion View paper
- [34] Test-Time Anchoring for Discrete Diffusion Posterior Sampling View paper

- [35] CLE Diffusion: Controllable Light Enhancement Diffusion Model View paper
- [36] SIG View paper
- [37] Fast non-autoregressive inverse folding with discrete diffusion View paper
- [38] Fixed Point Diffusion Models View paper
- [39] Controllable Graph Generation with Diffusion Models via Inference-Time Tree Search Guidance View paper
- [40] Toward Diffusion-Based Deep Reinforcement Learning for Discrete Decision-Making: Methods and Evaluations View paper
- [41] Unified Discrete Diffusion for Categorical Data View paper
- [42] DISCO: DISCrete nOise for Conditional Control in Text-to-Image Diffusion Models View paper
- [43] Composer Vector: Style-steering Symbolic Music Generation in a Latent Space View paper
- [44] DiffuseDRAW: Structured Latent Variables Model with Discrete Diffusion Prior View paper
- [45] Reverse Diffusion Sequential Monte Carlo Samplers View paper
- [46] Debiasing guidance for discrete diffusion with sequential monte carlo View paper
- [47] Reinforced sequential Monte Carlo for amortised sampling View paper
- [48] Efficient schemes for stochastic kinetic models View paper
- [49] Importance-Weighted Training of Diffusion Samplers View paper
- [50] Advancing Regularization Methods for Interpretable and Robust Deep Learning View paper
- [51] Computational methods for complex stochastic systems: Alternatives to MCMC View paper
- [52] Auto-Encoding Sequential Monte Carlo View paper
- [53] Sequential Monte Carlo approximations of Wasserstein--Fisher--Rao gradient flows View paper
- [54] Tuning Sequential Monte Carlo Samplers via Greedy Incremental Divergence Minimization View paper
- [55] Online variational sequential monte carlo View paper
- [56] Particle-MALA and Particle-mGRAD: Gradient-based MCMC methods for high-dimensional state-space models View paper
- [57] Parameter Estimation in Hidden Markov Models with Intractable Likelihoods Using Sequential Monte Carlo View paper
- [58] Stochastic gradient Hamiltonian sequential Monte Carlo filter with Earth Mover's Distance sampling for target tracking View paper
- [59] Smcp3: Sequential monte carlo with probabilistic program proposals View paper
- [60] A General-Purpose Fixed-Lag No U-Turn Sampler for Nonlinear Non-Gaussian State Space Models View paper
- [61] Enhanced SMC2: Leveraging Gradient Information from Differentiable Particle Filters Within Langevin Proposals View paper
- [62] Diffusion models in bioinformatics and computational biology View paper
- [63] A survey on generative diffusion models View paper
- [64] Vector Quantized Diffusion Model for Text-to-Image Synthesis View paper
- [65] Dirichlet diffusion score model for biological sequence generation View paper
- [66] Snapfusion: Text-to-image diffusion model on mobile devices within two seconds View paper
- [67] Multistate and functional protein design using RoseTTAFold sequence space diffusion View paper
- [68] De novo protein designâ□□From new structures to programmable functions View paper
- [69] Diffusion language models are versatile protein learners View paper
- [70] Versatile Diffusion: Text, Images and Variations All in One Diffusion Model View paper