VGAS: VARIANCE-REDUCED GUIDANCE AND ADAPTIVE SELECTION FOR TRAINING-FREE REWARD ALIGNMENT IN DISCRETE DIFFUSION

Anonymous authors
Paper under double-blind review

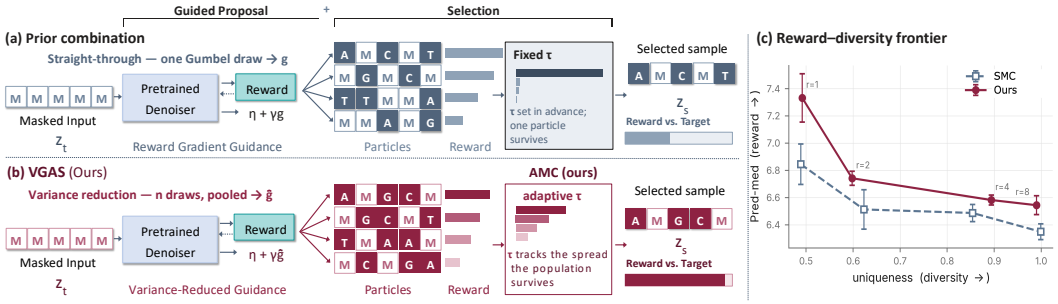


Figure 1: **Overview.** (a) prior combinations pair a high-variance gradient estimate with a fixed resampling temperature. (b) VGAS reduces that variance and sets the temperature per step. (c) it attains the higher reward at every resampling frequency, across the diversity range.

ABSTRACT

Masked discrete diffusion models perform strongly on text, code, and biological sequences, but their training objective rewards only naturalness, and retraining the generator for every new reward is expensive. Inference-time steering of a frozen model either guides the sampler by the reward gradient or searches over several trajectories, and recent samplers combine the two. Such combinations are assembled as pipelines that leave three choices at their defaults: a guidance estimate resting on one Gumbel draw per sample, a reward tilting placed without reference to the distribution the combination then targets, and a selection temperature held fixed although the spread of per-step rewards drifts. We identify that distribution and settle the three choices against it. We therefore propose **V**ariance-reduced **G**uidance and **A**daptive **S**election (VGAS), a simple yet effective inference-time framework that reduces the variance of the guidance estimate for both reward types, applies the reward tilting in the clean-token logits, where the pretrained schedule is preserved, and sets the selection temperature per step. Across regulatory DNA, protein and small-molecule benchmarks, VGAS attains the best training-free reward and matches or surpasses a reward-fine-tuned generator.

1 INTRODUCTION

Masked discrete diffusion generates a sequence by iteratively unmasking it: the forward process replaces tokens with a MASK symbol and the model learns to reverse it (Sahoo et al., 2024; Shi et al., 2024). Simple to train and strong on language and code, these models now also address biological and chemical design problems such as regulatory DNA, proteins, and small molecules (Wang et al., 2025; Lin et al., 2025). Their training objective, however, rewards only naturalness.

In most applications, however, naturalness is not the objective. The goal is samples that also score highly under a downstream reward, such as an enhancer that drives high expression, a protein that folds stably, a molecule that meets a target property, or a text completion that a preference model rates well. One route is post-training: fine-tuning or reinforcement learning of the generator against the reward (Wang et al., 2025; Han et al., 2026; Rector-Brooks et al., 2025), but this is expensive

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

and must be redone for every new reward. The alternative, now standard for continuous diffusion, is inference-time steering: keep the pretrained model frozen and bias only the sampling process toward high-reward regions. In this work, we focus on this inference-time, training-free setting: given a frozen generator p_θ and a reward r, sample sequences that are simultaneously likely under p_θ and high under r, with no retraining.

Specifically, these steering methods fall into three families (detailed in §2). *Gradient guidance* biases sampling by the reward gradient (Chung et al., 2023; Nisonoff et al., 2025; Lin et al., 2025; Tejaswi et al., 2026), and the instantiation we build on is GILC, which applies the reward tilting to the clean-token logits directly (Dou et al., 2026). *Search* keeps several trajectories and resamples them by value, as in SMC (Wu et al., 2023) and SVDD (Li et al., 2025), and a fast-growing line of Feynman–Kac and sequential Monte Carlo resampling extends it (Hasan et al., 2026; Dang et al., 2026; Yadala Chanchu et al., 2026); it improves with more particles but leaves the per-step proposal unchanged. A parallel line steers by gradient-free MCMC instead (Chu et al., 2025; Phunyaphibarn & Sung, 2026). The third family therefore *combines* them, pairing a gradient-guided proposal with search, as in the twisted diffusion sampler TDS (Wu et al., 2023), SMC-DDM (Ou et al., 2026) and TreeG's gradient variant (Guo et al., 2025). Appendix F positions this literature (Table 16) and anatomizes the closest of these samplers (Table 18).

These combinations, however, are assembled as pipelines, and each half is left at its standard setting. The guidance gradient remains a straight-through Monte Carlo estimate whose variance limits how much the guidance helps; the reward tilting is placed without reference to the distribution the combination then targets; and the search resamples at a temperature *set in advance*, even though the spread of per-step rewards varies several-fold along the denoising trajectory, so one temperature is too soft at one end and too sharp at the other.

As shown in Figure 1, all three can be settled at no additional denoiser call. We first establish what the combined sampler targets. For the estimator, we replace the straight-through estimate with a Rao–Blackwellized one for differentiable rewards and a leave-one-out baseline for non-differentiable ones. For the placement, we apply the reward tilting in the clean-token logits, where the undetermined constant in the guidance direction cancels and the pretrained schedule is preserved. For the temperature, we standardize the per-step values, which reduces to a single adaptive resampling temperature. Selection also concentrates the sample set, and VGAS attains the higher reward at every resampling frequency (Figure 1c). Our contributions are as follows.

- We propose VGAS, a training-free sampler that reduces the variance of the guidance estimate for both differentiable and non-differentiable rewards, applies the reward tilting in the clean-token logits, and sets the selection temperature per step.
- We show that a guided proposal paired with selection, run without the proposal correction and at a fixed selection temperature, still targets a reward tilt of the pretrained model, and we establish the effective temperature of that tilt, which is what makes the correction optional.
- Across regulatory DNA, protein design and small-molecule property targeting, VGAS achieves state-of-the-art inference-time reward alignment.

2 Background

Masked Discrete Diffusion. Let $\boldsymbol{x} = (x^1, \ldots, x^L)$ be a length-L sequence with each token x^ℓ in a vocabulary $\mathcal{V} = \{1, \ldots, V\}$, one of which is a special absorbing MASK symbol; write $V_0 = V - 1$ for the number of ordinary tokens. A masked (absorbing) diffusion model (Sahoo et al., 2024; Shi et al., 2024; Wang et al., 2025) defines a forward process that gradually replaces tokens with MASK, and learns to reverse it. Generation therefore proceeds by iterative unmasking: start from the all-MASK sequence and, over T steps, progressively reveal tokens until none are masked.

Concretely, a time variable t runs from $t{=}1$ (all masked) down to $t{=}0$ (fully revealed); write $\boldsymbol{z}_t$ for the partially masked state and $s = t - \Delta t$ for the next, less-noisy time. A denoiser network reads $\boldsymbol{z}_t$ and outputs, for each masked position, a categorical distribution over the clean token,

$$p_\theta(x_0^\ell = k \mid \boldsymbol{z}_t) = \mathrm{softmax}(\boldsymbol{\eta}^\ell)_k, \qquad \boldsymbol{\eta}^\ell = \boldsymbol{\eta}^\ell(\boldsymbol{z}_t) \in \mathbb{R}^V, \tag{1}$$

and these predictions are independent across positions, $p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{z}_t) = \prod_\ell p_\theta(x_0^\ell \mid \boldsymbol{z}_t)$. Let $\bar{\alpha}_t \in [0, 1]$ be the fraction of tokens still present under the schedule ($\bar{\alpha}_1{=}0, \bar{\alpha}_0{=}1$); we write it with a bar

to keep it distinct from the selection temperature α of §3.2. One reverse step either reveals a masked position or leaves it masked:

$$p_\theta(z_s^\ell \mid \boldsymbol{z}_t) = \begin{cases} \delta_{z_s^\ell = z_t^\ell}, & z_t^\ell \neq \text{MASK} \ \text{(already revealed; frozen)}, \\ (1-a_t)\,\delta_{\text{MASK}} + a_t\, p_\theta(x_0^\ell \mid \boldsymbol{z}_t), & z_t^\ell = \text{MASK}. \end{cases} \tag{2}$$

Here $a_t = (\bar\alpha_s - \bar\alpha_t)/(1-\bar\alpha_t) \in [0,1]$ is the per-step reveal probability set by the schedule.

Reward-Tilted Sampling and the Soft Value Function. Given a reward $r : \mathcal{V}^L \to \mathbb{R}$, reward alignment maximizes the expected reward while staying close to the generator. For an inverse temperature β^{-1} (larger $1/\beta$ = stronger guidance), the maximizer of $\mathbb{E}_p[r] - \beta\,\mathrm{KL}(p \,\|\, p_\theta)$ is the reward-tilted distribution (Wang et al., 2025; Dou et al., 2026)

$$p^\star(\boldsymbol{x}_0) \;\propto\; p_\theta(\boldsymbol{x}_0)\, \exp\big(r(\boldsymbol{x}_0)/\beta\big). \tag{3}$$

Post-training methods approximate $p^\star$ by fitting a new generator; the inference-time methods studied here leave p_θ frozen and sample from $p^\star$ during the reverse process instead. The Doob h-transform of Eq. (2) samples Eq. (3) exactly, reweighting each reverse step by the soft value v_s (Wu et al., 2023; Dou et al., 2026):

$$p^\star(\boldsymbol{z}_s \mid \boldsymbol{z}_t) \;\propto\; p_\theta(\boldsymbol{z}_s \mid \boldsymbol{z}_t)\, \exp\big(v_s(\boldsymbol{z}_s)/\beta\big), \qquad v_s(\boldsymbol{z}_s) = \beta \log \mathbb{E}_{p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{z}_s)}\big[e^{r(\boldsymbol{x}_0)/\beta}\big], \tag{4}$$

where $\boldsymbol{x}_0$ denotes a fully revealed sequence, $p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{z}_s)$ the clean posterior of Eq. (1), and v_s the soft value function: the (log-sum-exp) expected future reward if denoising continues from $\boldsymbol{z}_s$ (Dou et al., 2026). In SMC terms, v_s is the optimal twist, and every method below approximates it. The difficulty is that v_s depends on the reward of clean sequences yet has to be evaluated at noisy states. Three families of inference-time methods approximate it, and the first two can be composed.

- **Gradient guidance** edits the reverse kernel itself by the reward gradient. DG (Nisonoff et al., 2025) and TFG-Flow (Lin et al., 2025) tilt the transition rates of the discrete process; for masked diffusion the tilt can instead be applied to the clean-token logits (Tejaswi et al., 2026; Dou et al., 2026), and we build on one such logit-space instantiation, GILC.
- **Search-based selection** keeps the base reverse kernel and samples several trajectories, reweighting or selecting by an inexpensive value estimate $\hat V(\boldsymbol{z})$: SMC resamples a particle population with weights $w \propto \exp(\Delta\hat V/\alpha)$ set by the per-step change in the estimate, while SVDD keeps the best candidate. Both improve with more particles but leave the per-step proposal unchanged.
- **Combining the two** acts on different parts of Eq. (4): in SMC terms, a twisted proposal paired with a selection step. This is the regime of TDS, SMC-DDM and TreeG; each half, however, keeps its default form, a high-variance estimator and a fixed temperature.

Revisiting GILC. Pushing the reveal toward higher reward requires $\nabla_{\boldsymbol{\eta}} r$, but r acts on a discrete sequence and is piecewise constant in $\boldsymbol{\eta}$. The standard fix is the Gumbel–Softmax straight-through (ST-GS) estimator (Jang et al., 2017; Bengio et al., 2013): draw $\hat{\boldsymbol{x}}_{\text{soft}} = \mathrm{softmax}((\boldsymbol{\eta} + \boldsymbol{\zeta})/\tau)$ at a relaxation temperature $\tau > 0$, feed the hard one-hot to the reward, and backpropagate through the soft relaxation,

$$\hat{\boldsymbol{x}} = \mathrm{onehot}\big(\arg\max_k \hat{\boldsymbol{x}}_{\text{soft},k}\big) - \mathrm{sg}(\hat{\boldsymbol{x}}_{\text{soft}}) + \hat{\boldsymbol{x}}_{\text{soft}}, \qquad \partial\hat{\boldsymbol{x}}/\partial\boldsymbol{\eta} = \partial\hat{\boldsymbol{x}}_{\text{soft}}/\partial\boldsymbol{\eta}. \tag{5}$$

GILC (Dou et al., 2026) instantiates gradient guidance through this path: it draws the revealed token from a corrected clean prediction, adding $\boldsymbol{\delta}^\ell = \gamma \boldsymbol{g}^\ell$ to each masked position's logits,

$$\hat p^\ell = \mathrm{softmax}\big(\boldsymbol{\eta}^\ell + \boldsymbol{\delta}^\ell\big), \qquad p^\star_{\text{guided}}(z_s^\ell \mid \boldsymbol{z}_t) = (1-a_t)\,\delta_{\text{MASK}} + a_t\,\hat p^\ell. \tag{6}$$

The logit correction $\boldsymbol{g}^\ell = \big(\partial r(\hat{\boldsymbol{x}}_0)/\partial\hat{\boldsymbol{x}}_0\big)\big(\partial\hat{\boldsymbol{x}}_0/\partial\boldsymbol{\eta}^\ell\big)$ estimates $\nabla_{\boldsymbol{\eta}^\ell}\mathbb{E}[r]$, obtained by approximating the expensive model Jacobian by the identity, $\partial\boldsymbol{\eta}/\partial\boldsymbol{z}_t \approx \boldsymbol{I}$ (hence "Jacobian-free"). This form applies when r is differentiable; when it is not, the same $\boldsymbol{g}^\ell$ is formed by a score-function (policy-gradient) estimator instead, and §3.1 treats the two cases symmetrically. The identity approximation is inherited from GILC, which states it as bypassing the Jacobian, and it is what removes the denoiser from the backward pass (Appendix F.3). The guidance scale $\gamma > 0$ plays the role of the inverse temperature $1/\beta$ in Eq. (3). Both estimators evaluate their Jacobian at one Gumbel draw per sample, so $\boldsymbol{g}^\ell$ carries the variance of that draw, and that variance is what caps the gain from guidance.

Algorithm 1: GUIDANCE-GR

Input: logits $\boldsymbol{\eta}$; state $\boldsymbol{z}_t$; reward r; size n; Rao M

for $i = 1, \dots, n$ **do**
  $\boldsymbol{\zeta} \sim \text{Gumbel}$, $i^\star = \arg\max(\boldsymbol{\eta}+\boldsymbol{\zeta})$;
  $\bar{\boldsymbol{x}} = \frac{1}{M} \sum_{m \le M} \text{softmax}_\tau(\boldsymbol{\eta}+\boldsymbol{\zeta}^{(m)} \,|\, i^\star)$;
  $\hat{\boldsymbol{x}} \leftarrow \text{ST}(i^\star, \bar{\boldsymbol{x}})$ (Eq. 5), $R_i = r(\hat{\boldsymbol{x}})$;
$\boldsymbol{g}_{\text{GR}} = \frac{1}{n} \sum_i \partial R_i / \partial \boldsymbol{\eta}$ (Eq. 8);

Output: correction $\boldsymbol{g}_{\text{GR}}$

Algorithm 2: GUIDANCE-RO

Input: logits $\boldsymbol{\eta}$; state $\boldsymbol{z}_t$; reward r; size n

$\boldsymbol{p} = \text{softmax}(\boldsymbol{\eta})$;
for $i = 1, \dots, n$ **do**
  $\boldsymbol{x} \sim \boldsymbol{p}$, $\hat{\boldsymbol{x}} = \text{onehot}(\boldsymbol{x})$, $R_i = r(\hat{\boldsymbol{x}})$;
$A_i = R_i - \frac{1}{n-1} \sum_{j \ne i} R_j$ (Eq. 9);
$\boldsymbol{g}_{\text{RO}} = \frac{1}{n} \sum_i A_i (\hat{\boldsymbol{x}}_i - \boldsymbol{p})$;

Output: correction $\boldsymbol{g}_{\text{RO}}$

3 MAIN CONTRIBUTION: VGAS

In this section, we present VGAS, which acts on the three choices such combinations leave at their defaults. §3.1 reduces the variance of the guided proposal's logit correction, §3.2 establishes what selection over that proposal targets and why the correction belongs in the clean-token logits, and §3.3 makes the resampling temperature adaptive, none of them at an extra denoiser call.

3.1 VARIANCE-REDUCED GUIDED PROPOSALS

We build on GILC, whose logit correction $\boldsymbol{g}$ of Eq. (6) is a Monte Carlo estimate of $\nabla_{\boldsymbol{\eta}} \mathbb{E}[r]$ formed from a single draw. We therefore introduce one variance reduction per reward type, each reusing the samples the correction already draws (base estimators in Appendix H).

Differentiable reward: the Gumbel–Rao proposal. The pathwise base estimator GUIDANCE-DB (Appendix H) averages straight-through gradients Eq. (5) over n Gumbel–Softmax samples; it is a biased relaxation, evaluated at the single Gumbel draw that produced each sample. The reveal already computes the hard sample $i^\star = \arg\max_k(\eta_k + \zeta_k)$, a sufficient statistic for the discrete outcome, so conditioning on it preserves the mean and cannot increase the variance (Rao, 1945; Blackwell, 1947). We therefore replace the straight-through Jacobian by its conditional expectation given $i^\star$, the Gumbel–Rao construction of Paulus et al. (2021) (GUIDANCE-GR, Algorithm 1). That conditional has a closed-form truncated-Gumbel reparameterization (Maddison et al., 2014; Kool et al., 2019b) that reuses $\boldsymbol{\eta}$ at no forward pass. Writing $v_k = \eta_k + \zeta_k$ for the perturbed logits and $T = v_{i^\star}$ for their maximum, and drawing $U_0, U_1, \dots$ independently from $\text{U}(0,1)$,

$$T = \log \textstyle\sum_j e^{\eta_j} - \log(-\log U_0),$$
$$v_k = -\log\big(e^{-T} - e^{-\eta_k} \log U_k\big) \;\; (k \ne i^\star), \qquad \zeta_k = v_k - \eta_k, \tag{7}$$

where $-\log U_k \sim \text{Exp}(1) > 0$ keeps $v_k \le T$, so the induced argmax is exactly $i^\star$. Repeating this M times gives $\boldsymbol{\zeta}^{(1)}, \dots, \boldsymbol{\zeta}^{(M)}$, and averaging over them gives the correction we use,

$$\boldsymbol{g}_{\text{GR}} = \frac{\partial r(\hat{\boldsymbol{x}}_0)}{\partial \hat{\boldsymbol{x}}_0} \, \mathbb{E}_{\boldsymbol{\zeta} \mid \arg\max(\boldsymbol{\eta}+\boldsymbol{\zeta}) = i^\star} \Big[\tfrac{\partial \, \text{softmax}_\tau(\boldsymbol{\eta}+\boldsymbol{\zeta})}{\partial \boldsymbol{\eta}} \Big] \;\approx\; \frac{\partial r(\hat{\boldsymbol{x}}_0)}{\partial \hat{\boldsymbol{x}}_0} \, \frac{1}{M} \sum_{m=1}^{M} \frac{\partial \, \text{softmax}_\tau(\boldsymbol{\eta} + \boldsymbol{\zeta}^{(m)})}{\partial \boldsymbol{\eta}}, \tag{8}$$

which removes the variance the single Gumbel draw contributes. GUIDANCE-DB is the "base" row of Table 12, where the sweep over M is reported.

Non-differentiable reward: the policy-gradient proposal. A non-differentiable reward supplies no gradient to relax, so $\boldsymbol{g}$ is obtained from the score-function identity instead, which needs only reward values. Draw n hard samples $\boldsymbol{x}^{(i)} \sim \boldsymbol{p} = \text{softmax}(\boldsymbol{\eta})$, write $\hat{\boldsymbol{x}}^{(i)} = \text{onehot}(\boldsymbol{x}^{(i)})$ and $R_i = r(\hat{\boldsymbol{x}}^{(i)})$, and weight the score $\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{x}^{(i)}) = \hat{\boldsymbol{x}}^{(i)} - \boldsymbol{p}$ by the reward. GILC forms that weight as a group-relative advantage, standardizing R_i by the mean and the standard deviation of the group (GUIDANCE-PG, Appendix H). The price is a variance higher than on the differentiable path. We propose to use a Reinforce leave-one-out (RLOO) baseline (Williams, 1992; Kool et al., 2019a; Ahmadian et al., 2024) in place of that standardization, which weights by the leave-one-out advantage A_i, the deviation of R_i from the mean of the other $n-1$ rewards:

$$A_i = R_i - \tfrac{1}{n-1} \textstyle\sum_{j \ne i} R_j, \qquad \boldsymbol{g}_{\text{RO}} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} A_i \, (\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{p}), \tag{9}$$

which reuses the same n reward evaluations (GUIDANCE-RO, Algorithm 2). The two reductions play the same role, one per estimator, through different arguments, and the asymmetry is structural:

the reveal already computes the sufficient statistic $i^\star$, so conditioning on it is free on the pathwise path, while the score-function estimator admits no such statistic and is reduced by a baseline instead. Gumbel–Rao therefore carries the variance inequality above; the baseline is unbiased for every constant and, for a single sample, reduces variance over an interval that Lemma 3 identifies under a sign condition, for which the leave-one-out mean is the natural plug-in (Appendix A). Table 11 measures its effect on the proposal.

Preserving the unmasking schedule. Placing $\boldsymbol{g}$ inside the softmax, rather than on the kernel after the reveal probability is fixed, leaves the reveal probability at a_t, whereas the post-hoc tilt rescales it to $a_t Z^\ell/(1 - a_t + a_t Z^\ell)$ with $Z^\ell = \mathbb{E}_{p_\theta^\ell}[e^{\gamma g}]$. The additive constant in $\boldsymbol{g}$ is not fixed by its definition: the softmax is unchanged by $\boldsymbol{\eta} \mapsto \boldsymbol{\eta} + c\mathbf{1}$, so a correction expressed in logit coordinates is determined only up to such a constant, and different derivations fix it differently. The logit-space reveal is invariant to it, whereas the post-hoc one tends to 1 or to 0 as the constant grows, so an undetermined choice can drive the post-hoc schedule to a full reveal or to none (Proposition 3, proved in Appendix A). §3.2 shows that this invariance acts on the target itself and not only on the quality of the proposal.

3.2 Selection over a Guided Proposal

Selector. Among SMC, SVDD and Best-of-N we use SMC, the only one that acts throughout the trajectory without extra denoiser calls: SVDD evaluates M candidates at every step and Best-of-N defers all selection to the end. A selector scores a partial state by the one-step value $V(\boldsymbol{z}) = r\big(\arg\max_{\boldsymbol{x}_0} p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{z})\big)$, and the previous step's value is cached, so each step costs one denoiser call. SMC (Del Moral, 2004) carries N particles, scores particle k by the increment $\Delta^k = V(\boldsymbol{z}_s^k) - V(\boldsymbol{z}_t^k)$, and resamples multinomially with $w^k \propto \exp(\Delta^k/\alpha)$, where α is the selection temperature, distinct from the KL weight β of Eq. (3).

Target of the combination. The reward enters through the proposal of §3.1 and again through the weights, so neither fixes the target alone. Three measures are in play: the base model p_θ, the law of the surviving particles that Lemma 1 identifies, and the twisted proposal q_γ, the path measure of the guided reveal of §3.1 run to the end, with terminal marginal $q_\gamma(\boldsymbol{x}_0)$. The log-ratio of that marginal to the base model,

$$\Phi_\gamma(\boldsymbol{x}_0) \;=\; \log \frac{q_\gamma(\boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_0)} \tag{10}$$

is the accumulated log-twist, whose trajectory counterpart is available in closed form (Appendix G.1). The combination then targets

$$\pi_\gamma(\boldsymbol{x}_0) \;\propto\; q_\gamma(\boldsymbol{x}_0)\, e^{r(\boldsymbol{x}_0)/\alpha} \;=\; p_\theta(\boldsymbol{x}_0)\, e^{\Phi_\gamma(\boldsymbol{x}_0) + r(\boldsymbol{x}_0)/\alpha}, \tag{11}$$

a reward tilt of the twisted proposal at every γ. Equivalently $\pi_\gamma = \arg\max_p \mathbb{E}_p[r] - \alpha\,\mathrm{KL}(p \,\|\, q_\gamma)$, which is Eq. (3) with the twisted base measure. Eq. (11) is exact because successive increments cancel along a trajectory, leaving only the terminal reward in the weight.

The alternative is to keep the proposal correction. Adding $\log(p_\theta/q_\gamma)$ to the log-weights returns the target to $p^\star$ for every γ, and the established particle methods for reward alignment retain it (Wu et al., 2023; Ou et al., 2026; Yadala Chanchu et al., 2026). We omit it instead, and doing so does not make the sampler approximate: both weightings are Feynman–Kac changes of measure (Del Moral, 2004; Chopin & Papaspiliopoulos, 2020), stated in general form by Wu et al. (2023), and Lemma 1 identifies the distribution each one targets exactly at finite γ, ours q_γ tilted by the reward and theirs $p^\star$; Proposition 1 then returns ours to the reward-tilted family of Eq. (3).

Lemma 1 (Targets with and without the proposal correction)**.** *Under boundedness of r and $\gamma\boldsymbol{g}$, unbiased multinomial resampling, the deterministic all-*MASK *start and $V(\boldsymbol{z}_0) = r(\boldsymbol{x}_0)$, running the selector over q_γ at fixed α, (i) $\frac{1}{N}\sum_k \varphi(\boldsymbol{x}_0^k) \to \mathbb{E}_{\pi_\gamma}[\varphi]$ almost surely as $N \to \infty$ for every bounded measurable φ; and (ii) adding $\log(p_\theta/q_\gamma)$ to the log-weights replaces π_γ by $p^\star$ of Eq. (3) with $\beta = \alpha$, for every γ.*

Proposition 1 (Effective temperature of the guided proposal)**.** *$\Phi_\gamma(\boldsymbol{x}_0) = \kappa\gamma\, r(\boldsymbol{x}_0) + c + \varepsilon(\boldsymbol{x}_0)$, where c is independent of $\boldsymbol{x}_0$, $\kappa > 0$ is a tangential scale for the estimator that carries the reward gradient into the clean-token logits, and $|\varepsilon| \le E(\gamma) = O(\gamma)$. Omitting the correction therefore leaves the target in the reward-tilted family of Eq. (3) at inverse temperature $\kappa\gamma + 1/\alpha$, up to the factor e^{ε}; at $\gamma = 0$ the bound vanishes and $\beta = \alpha$ exactly.*

Appendix A gives the remaining definitions, including κ and $E(\gamma)$, the remaining hypotheses of both results, and their proofs. There $E(\gamma)$ is controlled by four quantities, one of which is the estimator deviation that the reductions of §3.1 act on, and it grows with the sequence length.

That the effective temperature differs from the nominal guidance scale is known for classifier-free guidance in the continuous setting (Bradley & Nakkiran, 2025), and guidance is known to reshape the sampled distribution of class-conditional masked diffusion in ways the nominal scale does not predict (He et al., 2026); Proposition 1 is the counterpart for a reward-gradient tilt.

To reach the inverse temperature $\kappa\gamma + 1/\alpha$ through the weights alone, α must be smaller, which concentrates them on a few particles, whereas twisting the proposal relocates particles into the reward-relevant region first (§5). The omission also fixes where the tilt is applied: with the correction retained, q_γ is only a proposal and the weights absorb its distortions; without it, q_γ enters the target, so the undetermined constant in $\boldsymbol{g}$ reaches π_γ through the reveal schedule (Proposition 3(c)), which logit-space guidance does not disturb.

3.3 Adaptive-Temperature Selection

The last of the three defaults is the selection temperature, which in SMC, TDS and SMC-DDM alike is prescribed before sampling rather than measured from the particles. Recorded while those samplers run, the per-step spread σ_t of the increments $\{\Delta_t^k\}$ falls by $3.7\times$ to $27.7\times$ along denoising on the three benchmarks, and with it the temperature matched to it (Figure 2a), so a temperature prescribed once exceeds the matched value over most of the trajectory (Appendix F.1 separates this from classical adaptive tempering). We therefore standardize the increment across the particle group, which applies a shift and a scale, and Lemma 2 shows that only one of the two survives the resampling softmax.

Lemma 2 (Baseline invariance of the resampling law)**.** *Fix $c > 0$. For increments $\{\Delta_t^k\}_{k=1}^N \subset \mathbb{R}$ and any constant b, the resampling law $\hat{w}_t^k \propto \exp\big((\Delta_t^k - b)/c\big)$ equals $\mathrm{softmax}_k(\Delta_t^k/c)$ and is independent of b.*

With μ_t, σ_t the per-step mean and standard deviation of $\{\Delta_t^k\}$, the shift therefore cancels and only the scale acts,

$$\hat{w}_t^k \;\propto\; \exp\Big(\frac{\Delta_t^k - \mu_t}{\alpha\sigma_t}\Big) \;\propto\; \exp\Big(\frac{\Delta_t^k}{\alpha_t}\Big), \qquad \alpha_t = \alpha\,\sigma_t, \tag{12}$$

so the fixed temperature becomes a per-step one that tracks the spread, so standardization's only operative effect is $\alpha \mapsto \alpha\sigma_t$, whichever estimate of μ_t is used (σ_t is floored at $\epsilon{=}10^{-6}$). The proposal's contribution $\kappa\gamma$ is unchanged; only the selection term $1/\alpha_t$ now varies along the trajectory. Measured as the per-step effective sample size (ESS), a fixed α either collapses the particle population from the first steps or, at the value that avoids that, leaves the weights nearly uniform by the last ones, so selection is effectively off where the value estimate is most reliable, whereas AMC holds it in between throughout (Figure 2b, Appendix E.5). We call the selector AMC (Adaptive Monte Carlo), and it changes one line of SMC (Appendix G). Proposition 2 identifies the target that a per-step temperature induces.

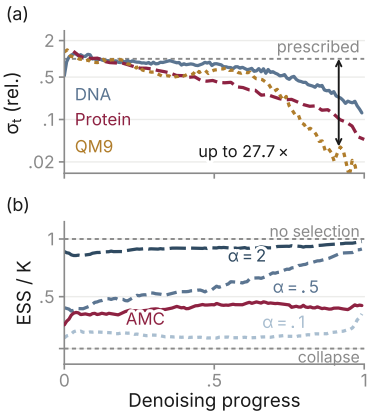


Figure 2: (a) Spread against the prescribed level. (b) ESS on DNA.

Proposition 2 (Target of AMC)**.** *The accumulated weight of a particle is proportional to e^{R_σ} with $R_\sigma = \sum_t \Delta_t/\alpha_t$, a functional of the whole trajectory, and the self-normalized estimator converges in probability under the path measure $\pi^\star_{\mathrm{AMC}} \propto q_\gamma e^{R_{\sigma^\star}}$, whose $\boldsymbol{x}_0$-marginal is the law of the samples. If $\sigma_t^\star \equiv \sigma$, that marginal is π_γ with α replaced by $\alpha\sigma$.*

Adapting the temperature changes what the selector targets. With α_t varying the increments no longer cancel, so R_σ retains the whole trajectory rather than its endpoint, and the reweighting is by the greedy value V of §3.2 rather than the soft value v_s of §2. In return the selector gains a pressure that tracks the spread at every step, which no single α does (§5). When the spread does not drift, $\sigma_t^\star$ constant, the selector reduces to the fixed-temperature one at $\alpha\sigma$, so the adaptation is active exactly where the drift is. Appendix A carries the hypotheses and the proof, and Appendix F.2 places AMC against the selectors of TDS and SMC-DDM.

Standardizing has two further consequences. Dividing by σ_t makes the selector invariant to reward rescaling $r \mapsto ar$, $a > 0$, which scales Δ_t^k and σ_t alike, so α needs no retuning (Table 10). The invariance is a property of α alone; γ multiplies a gradient that scales with the reward. A leave-one-out baseline in place of μ_t is inert to $O(1/(N-1))$ (Remark 2), which Table 11 tests.

Our final method: VGAS. VGAS combines the two: a variance-reduced guided proposal searched under the adaptive selector AMC, as VGAS-GR for a differentiable reward and VGAS-RO (RLOO baseline) for a non-differentiable one.

4 Experiments

We evaluate VGAS on three domains: regulatory DNA sequence design (§4.1), protein sequence design (§4.2), and small-molecule property targeting (§4.3).

Baselines. To rigorously evaluate our proposed method, we compare against the strongest training-free and fine-tuned steering methods for discrete diffusion, reproduced from their official repositories under a common setup. They fall into four groups. Fine-tuning: DRAKES, the standard fine-tuned reference on these benchmarks. Gradient guidance: DG and GILC. Search: Best-of-N, SMC, and SVDD. Combination (guidance + search): TDS, SMC-DDM, and TreeG, of which the last two are closest to ours. Implementation details are in Appendix B.

Synthetic verification. Before the three benchmarks, we check the sampler where the answer is known. On the 64×64 grid of Ou et al. (2026) the reward-tilted target is available in closed form, and the components reduce the earth-mover distance to it one at a time (Appendix E).

4.1 DNA Enhancer Sequence Design

Setup. We evaluate on the genomic enhancer benchmark of Gosai et al. (2024), as adopted by Wang et al. (2025). The backbone is a masked diffusion language model (Sahoo et al., 2024) pretrained on this data with $T=128$ reverse steps. We steer toward high predicted HepG2 enhancer activity, with a held-out chromatin-accessibility (ATAC) oracle reserved for evaluation.

Metrics. Reward is the median predicted HepG2 activity (Pred-med); fidelity is the held-out ATAC pass rate, the 3-mer and JASPAR-motif correlations to natural enhancers (Castro-Mondragon et al., 2022), and the approximate log-likelihood under the frozen generator (App-LL; Appendix C).

Results. As shown in Table 1, VGAS generates sequences with the highest predicted HepG2 activity among all methods, and does so on both paths: the differentiable VGAS-GR and the non-differentiable VGAS-RO each exceed every training-free baseline, the closest combination methods SMC-DDM and TreeG-G, and the reward-fine-tuned DRAKES. Fidelity does not degrade alongside it. The held-out accessibility oracle, which never guides sampling, reaches 99.1% against DRAKES's 92.4%, the 3-mer correlation to natural enhancers is highest under VGAS-RO, and the JASPAR-motif correlation is level with DRAKES. Approximate likelihood is the one metric on which guided samplers trade against the pretrained model, and VGAS concedes more of it than the samplers that leave the reverse kernel intact and no more than the logit-space guidance it builds on.

Table 1: **Quantitative comparison on DNA enhancer design** (Wang et al., 2025) (mean±std, 3 seeds, $N=20$). Bold indicates the best performance.

Method	Pred-med ↑	ATAC-Acc (%) ↑	3-mer Corr ↑	JASPAR ↑	App-LL ↑
Pretrained (MDLM) (Sahoo et al., 2024)	$0.18_{\pm0.02}$	$2.0_{\pm0.4}$	$-0.04_{\pm0.01}$	$0.29_{\pm0.01}$	$-261.9_{\pm0.6}$
DG (Nisonoff et al., 2025)	$1.10_{\pm0.01}$	$0.0_{\pm0.0}$	$-0.09_{\pm0.01}$	$-0.02_{\pm0.01}$	$-267.8_{\pm0.2}$
Best-of-N (Beirami et al., 2025)	$1.81_{\pm0.06}$	$6.8_{\pm1.9}$	$0.35_{\pm0.04}$	$0.76_{\pm0.01}$	$-260.2_{\pm0.1}$
SMC (Del Moral, 2004)	$4.26_{\pm0.13}$	$30.9_{\pm5.9}$	$0.83_{\pm0.03}$	$0.76_{\pm0.02}$	$-257.8_{\pm2.6}$
TDS (Wu et al., 2023)	$4.48_{\pm0.80}$	$36.6_{\pm9.1}$	$0.76_{\pm0.17}$	$0.81_{\pm0.14}$	$\mathbf{-257.4}_{\pm4.5}$
SVDD (Li et al., 2025)	$5.30_{\pm0.02}$	$29.3_{\pm0.2}$	$0.73_{\pm0.01}$	$0.83_{\pm0.05}$	$-259.1_{\pm0.3}$
DRAKES (Wang et al., 2025)	$5.61_{\pm0.08}$	$92.4_{\pm0.7}$	$0.89_{\pm0.01}$	$\mathbf{0.91}_{\pm0.01}$	$-264.2_{\pm0.3}$
GILC-DB (Dou et al., 2026)	$6.21_{\pm0.05}$	$83.8_{\pm1.7}$	$0.79_{\pm0.02}$	$0.90_{\pm0.01}$	$-279.1_{\pm0.5}$
GILC-PG (Dou et al., 2026)	$4.86_{\pm0.07}$	$47.2_{\pm0.7}$	$0.37_{\pm0.02}$	$0.87_{\pm0.01}$	$-276.3_{\pm0.4}$
SMC-DDM (Ou et al., 2026)	$5.71_{\pm0.04}$	$41.8_{\pm0.5}$	$0.53_{\pm0.02}$	$0.77_{\pm0.01}$	$-278.3_{\pm0.2}$
TreeG-G (Guo et al., 2025)	$5.96_{\pm0.02}$	$95.5_{\pm0.8}$	$0.78_{\pm0.02}$	$0.89_{\pm0.01}$	$-286.4_{\pm0.2}$
VGAS-GR	$\mathbf{7.33}_{\pm0.18}$	$\mathbf{99.1}_{\pm1.6}$	$0.88_{\pm0.03}$	$0.88_{\pm0.01}$	$-276.3_{\pm0.7}$
VGAS-RO	$6.96_{\pm0.19}$	$93.9_{\pm2.4}$	$\mathbf{0.94}_{\pm0.01}$	$0.90_{\pm0.01}$	$-275.0_{\pm0.3}$

Table 2: **Quantitative comparison on protein design** (inverse folding) (Wang et al., 2025) (mean±std, 3 seeds, N=20).

Method	Pred-ddG↑	%(ddG>0) (%)↑	scRMSD↓	%(scRMSD<2 Å)↑	Success Rate (%)↑
Pretrained (MDLM) (Sahoo et al., 2024)	$-0.55_{\pm 0.05}$	$36.1_{\pm 0.6}$	$0.85_{\pm 0.01}$	$90.2_{\pm 1.0}$	$33.9_{\pm 0.9}$
DG (Nisonoff et al., 2025)	$-0.54_{\pm 0.05}$	$36.2_{\pm 0.6}$	$0.85_{\pm 0.02}$	$90.3_{\pm 1.0}$	$33.9_{\pm 0.6}$
SMC (Del Moral, 2004)	$0.37_{\pm 0.07}$	$59.9_{\pm 2.6}$	$0.86_{\pm 0.01}$	$\mathbf{93.7}_{\pm 2.5}$	$55.8_{\pm 1.0}$
TDS (Wu et al., 2023)	$0.41_{\pm 0.14}$	$62.8_{\pm 3.7}$	$0.86_{\pm 0.02}$	$92.4_{\pm 2.2}$	$57.1_{\pm 2.3}$
Best-of-N (Beirami et al., 2025)	$0.51_{\pm 0.06}$	$61.7_{\pm 1.6}$	$0.86_{\pm 0.02}$	$92.6_{\pm 0.5}$	$57.2_{\pm 1.3}$
SVDD (Li et al., 2025)	$0.69_{\pm 0.08}$	$69.3_{\pm 2.1}$	$0.85_{\pm 0.03}$	$89.7_{\pm 0.9}$	$65.0_{\pm 3.3}$
DRAKES (Wang et al., 2025)	$1.08_{\pm 0.02}$	$86.1_{\pm 0.4}$	$0.91_{\pm 0.01}$	$92.2_{\pm 0.9}$	$78.6_{\pm 1.3}$
GILC-DB (Dou et al., 2026)	$0.64_{\pm 0.02}$	$77.0_{\pm 1.3}$	$0.89_{\pm 0.01}$	$89.2_{\pm 0.4}$	$67.5_{\pm 1.4}$
GILC-PG (Dou et al., 2026)	$0.16_{\pm 0.02}$	$54.3_{\pm 0.4}$	$0.85_{\pm 0.01}$	$90.8_{\pm 0.5}$	$49.7_{\pm 0.6}$
SMC-DDM (Ou et al., 2026)	$0.62_{\pm 0.05}$	$65.3_{\pm 3.0}$	$0.86_{\pm 0.03}$	$92.5_{\pm 0.5}$	$59.4_{\pm 3.3}$
VGAS-GR	$\mathbf{1.29}_{\pm 0.16}$	$\mathbf{90.2}_{\pm 4.3}$	$0.88_{\pm 0.04}$	$91.6_{\pm 0.9}$	$\mathbf{82.1}_{\pm 3.2}$
VGAS-RO	$1.12_{\pm 0.09}$	$76.0_{\pm 1.0}$	$0.85_{\pm 0.02}$	$92.3_{\pm 1.7}$	$69.7_{\pm 0.3}$

Table 3: **Quantitative comparison on QM9 conditional generation** (Ramakrishnan et al., 2014) (mean±std, 3 seeds); metrics and units in Appendix C.

Method	C_v↓	α↓	μ↓	gap↓	HOMO↓	LUMO↓
Pretrained	$4.17_{\pm 0.04}$	$7.77_{\pm 0.05}$	$1.58_{\pm 0.04}$	$1450_{\pm 16}$	$635_{\pm 20}$	$1422_{\pm 6}$
TFG-Flow (Lin et al., 2025)	$2.20_{\pm 0.03}$	$2.95_{\pm 0.15}$	$0.946_{\pm 0.025}$	$926_{\pm 7}$	$551_{\pm 11}$	$761_{\pm 14}$
SMC (Del Moral, 2004)	$3.48_{\pm 0.03}$	$3.95_{\pm 0.23}$	$1.27_{\pm 0.03}$	$1445_{\pm 13}$	$639_{\pm 10}$	$1416_{\pm 18}$
SVDD (Li et al., 2025)	$3.54_{\pm 0.06}$	$3.79_{\pm 0.08}$	$1.32_{\pm 0.02}$	$1416_{\pm 22}$	$629_{\pm 19}$	$1421_{\pm 40}$
GILC-DB (Dou et al., 2026)	$2.49_{\pm 0.08}$	$2.68_{\pm 0.09}$	$1.19_{\pm 0.06}$	$913_{\pm 15}$	$500_{\pm 2}$	$892_{\pm 8}$
GILC-PG (Dou et al., 2026)	$2.10_{\pm 0.08}$	$3.33_{\pm 0.24}$	$1.09_{\pm 0.05}$	$859_{\pm 41}$	$420_{\pm 9}$	$665_{\pm 10}$
SMC-DDM (Ou et al., 2026)	$2.14_{\pm 0.06}$	$3.61_{\pm 0.10}$	$0.958_{\pm 0.025}$	$879_{\pm 16}$	$520_{\pm 12}$	$641_{\pm 12}$
TreeG-G (Guo et al., 2025)	$3.54_{\pm 0.13}$	$5.63_{\pm 0.19}$	$1.77_{\pm 0.19}$	$1380_{\pm 25}$	$712_{\pm 4}$	$1308_{\pm 20}$
VGAS-GR	$2.06_{\pm 0.07}$	$\mathbf{2.38}_{\pm 0.07}$	$0.921_{\pm 0.014}$	$795_{\pm 32}$	$406_{\pm 4}$	$586_{\pm 11}$
VGAS-RO	$\mathbf{1.22}_{\pm 0.07}$	$2.53_{\pm 0.13}$	$\mathbf{0.718}_{\pm 0.011}$	$\mathbf{790}_{\pm 10}$	$\mathbf{374}_{\pm 5}$	$\mathbf{555}_{\pm 5}$

4.2 Protein Sequence Design

Setup. We evaluate on the protein inverse-folding benchmark of Wang et al. (2025), steering a discrete diffusion model over amino-acid sequences conditioned on a target structure toward predicted stability (ddG) from an oracle trained on Megascale (Tsuboyama et al., 2023); each sequence is re-folded to check that it still folds to that structure.

Metrics. Reward is the median predicted stability (Pred-ddG) and the fraction of stabilizing designs (ddG>0); fidelity is the self-consistency RMSD from re-folding (scRMSD), the well-folded fraction (scRMSD<2 Å), and the joint success rate (Appendix C).

Results. As shown in Table 2, both VGAS paths raise predicted stability above every training-free baseline and above the reward-fine-tuned DRAKES, and VGAS-GR attains the largest fraction of stabilizing designs. Fidelity holds: the joint structure-and-stability success rate is the highest at 82.1%, and the self-consistency RMSD of both paths stays under the fine-tuned model's, with VGAS-RO matching the pretrained model's.

4.3 Small-Molecule Property Targeting

Setup. We evaluate on QM9 (Ramakrishnan et al., 2014) under the conditional-generation protocol of TFG-Flow (Lin et al., 2025), drawing a quantum-chemical property value as the condition and steering the sampler to realize it.

Metrics. We measure reward by the mean absolute error (MAE) between the requested property value and the value realized, over the six properties of Lin et al. (2025); lower is better. Following them and Dou et al. (2026), fidelity is chemical validity, and every method is compared above a 75% validity operating point so that a lower MAE cannot come from invalid molecules.

Results. As shown in Table 3, both VGAS variants attain a lower error than every baseline on all six properties. The ordering across baseline groups matches the sequence domains: search alone improves little, gradient guidance accounts for most of the gain, and the combinations for the remainder. VGAS-RO is the stronger of the two here, leading on five of the six properties at the larger reward budget of Table 6.

Table 4: **Component ablation and effect of AMC.** (a) each proposal to its final VGAS setting; (b) SMC-DDM's selector swapped for AMC.

Stage	DNA Pred-med ↑	Protein Pred-ddG ↑
GILC-DB	6.21	0.64
+ Gumbel–Rao	6.35	0.70
+ SMC	6.84	0.85
+ AMC (VGAS-GR)	**7.33**	**1.29**
GILC-PG	4.86	0.16
+ SMC	6.33	0.41
+ AMC	6.61	1.04
+ RO (VGAS-RO)	**6.96**	**1.12**

(a) Component ablation.

n	selector	Pred-med ↑	ATAC (%) ↑	3-mer ↑	JASPAR ↑
4	SMC-DDM	$5.12_{\pm 0.06}$	$35.1_{\pm 0.6}$	$0.38_{\pm 0.03}$	$0.70_{\pm 0.01}$
4	AMC	$\mathbf{5.37}_{\pm 0.06}$	$\mathbf{42.4}_{\pm 1.8}$	$\mathbf{0.74}_{\pm 0.03}$	$\mathbf{0.90}_{\pm 0.01}$
10	SMC-DDM	$5.71_{\pm 0.04}$	$41.8_{\pm 0.5}$	$0.53_{\pm 0.02}$	$0.77_{\pm 0.01}$
10	AMC	$\mathbf{5.85}_{\pm 0.03}$	$\mathbf{57.5}_{\pm 1.0}$	$\mathbf{0.90}_{\pm 0.01}$	$\mathbf{0.96}_{\pm 0.01}$

(b) AMC vs. SMC-DDM's selector at n reward draws.

5 Ablation Study and Analysis

Ablation study. To separate the contribution of each component, we conduct an ablation study that adds them one at a time to both proposals (Table 4a; full metrics in Tables 7 and 8). The variance reduction lifts the proposal on its own, on both paths, and selection carries the larger share: SMC lifts the weaker policy-gradient proposal the most on DNA, whereas on protein the adaptive temperature contributes more than SMC on both paths. Every component contributes, and the fidelity metrics do not degrade as the reward rises.

Generalizability of AMC. To test whether the selector generalizes beyond its own proposal, we hold SMC-DDM's fixed and replace only the selector (Table 4b). Reward and every fidelity metric rise at both budgets, so an adaptive temperature adds to a first-order approximation of the locally optimal twist (Ou et al., 2026, §3.2.1); and since that sampler carries the proposal correction and ours omits it, the comparison separates the selection rule from that choice.

Effect of the adaptive temperature. To show that the gain is not tied to one scale of α, we sweep it over two decades and run both selectors at each value on the same guided proposal, so SMC is the rung of Table 4a (3 seeds; Figure 3). AMC is higher at every α and at each selector's own best α, on the reward and the held-out oracle alike, and is the flatter of the two; the oracle peaks at our default α=0.5. No prescribed temperature recovers the adaptive one, which confirms Eq. (12).

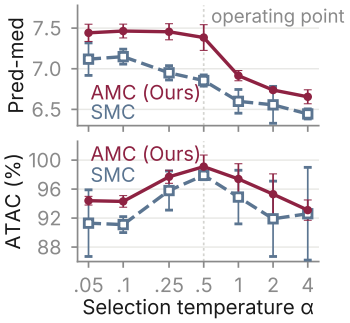


Figure 3: Adaptive selection temperature beats every fixed α on the DNA benchmark.

Selection and diversity. To separate reward from diversity, we measure diversity three ways (Tables 13 and 14). The benchmark's mean pairwise Hamming distance does not register the collapse, while the nearest-neighbour distance and the uniqueness rate do. Sweeping the resampling frequency, AMC attains the higher reward at every setting (Figure 1c), so the concentration is an operating point rather than a property of the selector.

Limitations. VGAS costs a particle population rather than a single trajectory, and selection concentrates the sample set, which the resampling frequency controls (Appendix E). We evaluate on biological sequences and small molecules, and leave text untested.

6 Conclusion

We proposed VGAS, a training-free sampler for inference-time reward tilting of masked discrete diffusion models. Guided proposals paired with selection leave three choices at their defaults, so we first established what such a combination targets once the proposal correction is omitted, and settled each against that target: a Rao–Blackwellized guidance estimate for differentiable rewards and a leave-one-out baseline for non-differentiable ones, the tilting applied in the clean-token logits where the pretrained schedule is preserved, and a selection temperature standardized by the spread of the per-step rewards. Across regulatory DNA, protein and small-molecule benchmarks, VGAS attains the best training-free reward on all three and exceeds the reward-fine-tuned reference where one exists, without any training and without degrading fidelity. We expect it to extend to further domains and rewards through a selector that needs no retuning.

ETHICS STATEMENT

Our benchmarks are regulatory DNA enhancer design, inverse protein folding and small-molecule property targeting. Reward-guided sampling is indifferent to the sign of the objective: the same procedure that raises a benign reward would raise a harmful one if such an oracle were supplied. The rewards and the held-out evaluation oracles used here are published academic predictors and all results are in silico. We release the method rather than any designed sequence, and we regard oracle choice and biosecurity screening as prerequisites for downstream use.

AI USE STATEMENT

We used generative AI tools to edit prose, including grammatical correction, and to write plotting code. The method, the theory, the proofs and every experimental result are the authors' own. We take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

Every result is reported as a mean over three seeds with the standard deviation alongside, except for a few secondary tables in Appendix E that report point estimates. Our method is stated as an algorithm (Algorithms 1–4), with its theory and its experiments given side by side, and every proof is in Appendix A. Backbones, oracles, data splits, hyperparameters and per-step inference costs are in Appendix B and metric definitions in Appendix C. We will release the code.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE-Style Optimization for Learning from Human Feedback in LLMs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical Guarantees on the Best-of-N Alignment Policy. In *International Conference on Machine Learning (ICML)*, 2025.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *arXiv preprint arXiv:1308.3432*, 2013.

Alexandros Beskos, Ajay Jasra, Nikolas Kantas, and Alexandre Thiery. On the convergence of adaptive sequential Monte Carlo methods. *The Annals of Applied Probability*, 26(2):1111–1146, 2016.

David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18(1):105–110, 1947.

Arwen Bradley and Preetum Nakkiran. Classifier-Free Guidance is a Predictor-Corrector. *Transactions on Machine Learning Research (TMLR)*, 2025.

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.

Jaime A. Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, Oriol Fornes, Tiffany Y. Leung, Alejandro Aguirre, Fayrouz Hammal, Daniel Schmelter, Damir Baranasic, Benoit Ballester, Albin Sandelin, Boris Lenhard, Klaas Vandepoele, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50 (D1):D165–D173, 2022.

Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer, 2020.

Wenda Chu, Zihui Wu, Yifan Chen, Yang Song, and Yisong Yue. Split Gibbs Discrete Diffusion Posterior Sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Meihua Dang, Jiaqi Han, Minkai Xu, Kai Xu, Akash Srivastava, and Stefano Ermon. Inference-Time Scaling of Diffusion Language Models via Trajectory Refinement. In *Conference on Language Modeling (COLM)*, 2026. arXiv:2507.08390.

Pierre Del Moral. *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Probability and Its Applications. Springer, 2004.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

Hongkun Dou, Zike Chen, Fengji Li, Hongjue Li, and Yue Deng. Plug-and-Play Guidance for Discrete Diffusion Models via Gradient-Informed Logit Correction. In *International Conference on Machine Learning (ICML)*, 2026.

Sager J. Gosai, Rodrigo I. Castro, Natalia Fuentes, John C. Butts, Kousuke Mouri, Michael Alasoadura, Susan Kales, Thanh Thanh L. Nguyen, Ramil R. Noche, Arya S. Rao, Mary T. Joy, Pardis C. Sabeti, Steven K. Reilly, and Ryan Tewhey. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, 634:1211–1220, 2024.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

Nate Gruver, Samuel Stanton, Nathan C. Frey, Tim G. J. Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein Design with Guided Discrete Diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yingqing Guo, Yukang Yang, Hui Yuan, and Mengdi Wang. Training-Free Guidance Beyond Differentiability: Scalable Path Steering with Tree Search in Diffusion and Flow Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jiaqi Han, Austin Wang, Minkai Xu, Wenda Chu, Meihua Dang, Haotian Ye, Huayu Chen, Yisong Yue, and Stefano Ermon. Discrete Diffusion Trajectory Alignment via Stepwise Decomposition. In *International Conference on Learning Representations (ICLR)*, 2026.

Mohsin Hasan, Viktor Ohanesian, Artem Gazizov, Yoshua Bengio, Alán Aspuru-Guzik, Roberto Bondesan, Marta Skreta, and Kirill Neklyudov. Discrete Feynman-Kac Correctors. *arXiv preprint arXiv:2601.10403*, 2026.

Ye He, Kevin Rojas, and Molei Tao. What Exactly Does Guidance Do in Masked Discrete Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2026.

Peter Holderrieth, Douglas Chen, Luca Eyring, Ishin Shah, Giri Anantharaman, Yutong He, Zeynep Akata, Tommi Jaakkola, Nicholas M. Boffi, and Max Simchowitz. Diamond Maps: Efficient Reward Alignment via Stochastic Flow Maps. In *International Conference on Machine Learning (ICML)*, 2026.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. In *International Conference on Machine Learning (ICML)*, 2022.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Jaihoon Kim, Taehoon Yoon, Prin Phunyaphibarn, Seungjun Kim, Morteza Mardani, and Minhyuk Sung. Contrastive Distribution Matching for Amortized Sequential Monte Carlo in Discrete Diffusion. *arXiv preprint arXiv:2605.23346*, 2026.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE Samples, Get a Baseline for Free! In *ICLR Deep Reinforcement Learning Meets Structured Prediction Workshop*, 2019a.

Wouter Kool, Herke van Hoof, and Max Welling. Stochastic Beams and Where to Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement. In *International Conference on Machine Learning (ICML)*, 2019b.

Sanghyun Lee, Sunwoo Kim, Seungryong Kim, Jongho Park, and Dongmin Park. Effective Test-Time Scaling of Discrete Diffusion through Iterative Refinement. *arXiv preprint arXiv:2511.05562*, 2025.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-Free Guidance in Continuous and Discrete Diffusion Models with Soft Value-Based Decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2408.08252.

Haowei Lin, Shanda Li, Haotian Ye, Yiming Yang, Stefano Ermon, Yitao Liang, and Jianzhu Ma. TFG-Flow: Training-free Guidance in Multimodal Generative Flow. In *International Conference on Learning Representations (ICLR)*, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* Sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking Guidance for Discrete State-Space Diffusion and Flow Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Zijing Ou, Chinmay Pani, and Yingzhen Li. Inference-Time Scaling of Discrete Diffusion Models via Importance Weighting and Optimal Proposal Design. In *International Conference on Learning Representations (ICLR)*, 2026.

Max B Paulus, Chris J Maddison, and Andreas Krause. Rao-Blackwellizing the Straight-Through Gumbel-Softmax Gradient Estimator. In *International Conference on Learning Representations (ICLR)*, 2021.

Prin Phunyaphibarn and Minhyuk Sung. Reward-Guided Discrete Diffusion via Clean-Sample Markov Chain for Molecule and Biological Sequence Design. *arXiv preprint arXiv:2602.09424*, 2026.

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

C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Jarrid Rector-Brooks, Mohsin Hasan, Zhangzhi Peng, Cheng-Hao Liu, Sarthak Mittal, Nouha Dziri, Michael Bronstein, Pranam Chatterjee, Alexander Tong, and Avishek Joey Bose. Steering Masked Discrete Diffusion Models via Discrete Denoising Posterior Prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander M Rush, and Volodymyr Kuleshov. Simple and Effective Masked Diffusion Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and Generalized Masked Diffusion for Discrete Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A General Framework for Inference-time Scaling and Steering of Diffusion Models. In *International Conference on Machine Learning (ICML)*, 2025.

Atula Tejaswi, Litu Rout, Constantine Caramanis, Sanjay Shakkottai, and Sujay Sanghavi. Entropy Aware Reward Guidance for Diffusion Language Model Alignment. *arXiv preprint arXiv:2602.05000*, 2026.

Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J Weinstein, Niall M Mangan, Sergey Ovchinnikov, and Gabriel J Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620:434–444, 2023.

Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Tommaso Biancalani, Avantika Lal, Tommi Jaakkola, Sergey Levine, Hanchen Wang, and Aviv Regev. Fine-Tuning Discrete Diffusion Models via Reward Optimization with Applications to DNA and Protein Design. In *International Conference on Learning Representations (ICLR)*, 2025.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696.

Luhuan Wu, Brian L Trippe, Christian A Naesseth, David M Blei, and John P Cunningham. Practical and Asymptotically Exact Conditional Sampling in Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lohithsai Yadala Chanchu, Hany Abdulsamad, and Christian A. Naesseth. Discrete Diffusion Inference-Time Control with Nested Sequential Monte Carlo. *arXiv preprint arXiv:2608.20123*, 2026.

APPENDIX CONTENTS

A Deferred Proofs . . . 14
B Implementation Details . . . 21
C Metric Definitions . . . 23
D Full-Result Ablation Study . . . 24
E Additional Analysis . . . 25
E.1 Reward-Scale Invariance . . . 25
E.2 Isolating the Matched Variance Reducers . . . 25
E.3 Protein Structure versus Stability . . . 26
E.4 Diversity Metric Blindness . . . 26
E.5 What a Fixed Temperature Costs . . . 26
F Related Work and Positioning . . . 28
F.1 Relation to Adaptive-Tempering SMC . . . 29
F.2 Anatomy of Guidance and Selection in Prior Samplers . . . 29
F.3 Explicit Comparison of the Guidance Signals . . . 30
G Sampler and Selector Algorithms . . . 31
H Base Proposal Estimators . . . 32

The four results declared in the main text are restated and then proved, in the order they appear there; two further results, declared here, follow.

A DEFERRED PROOFS

Lemma 1 (Targets with and without the proposal correction, restated). *Assume* $\|r\|_\infty < \infty$, $\|\gamma \boldsymbol{g}\|_\infty < \infty$, *unbiased multinomial resampling, the deterministic all-*MASK *initial state, and the terminal identity* $V(\boldsymbol{z}_0) = r(\boldsymbol{x}_0)$. *Run the filter with proposal* q_γ *and incremental weight* $w_t^k = \exp(\Delta_t^k/\alpha)$, $\Delta_t^k = V(\boldsymbol{z}_s^k) - V(\boldsymbol{z}_t^k)$, *at fixed* α. *Then for every bounded* φ, $\frac{1}{N}\sum_k \varphi(\boldsymbol{x}_0^k) \to \mathbb{E}_{\pi_\gamma}[\varphi]$ *almost surely as* $N \to \infty$, *where* $\pi_\gamma \propto q_\gamma\, e^{r/\alpha}$; *and* $\pi_0 = p^\star$ *of Eq. (3) with* $\beta = \alpha$. *Moreover, adding* $\log(p_\theta/q_\gamma)$ *to the log-weights replaces* π_γ *by* $p^\star$ *with* $\beta = \alpha$, *for every* γ.

Proof. **Step 1 (accumulated weight).** Fix a particle k with trajectory $\boldsymbol{z}_T \to \cdots \to \boldsymbol{z}_0$. Its accumulated unnormalized weight is a product of successive ratios:

$$\prod_t w_t^k = \exp\Big(\tfrac{1}{\alpha}\textstyle\sum_t \big(V(\boldsymbol{z}_s^k) - V(\boldsymbol{z}_t^k)\big)\Big) \tag{13}$$

$$= \exp\Big(\tfrac{1}{\alpha}\big(V(\boldsymbol{z}_0^k) - V(\boldsymbol{z}_T^k)\big)\Big) \tag{14}$$

$$= \exp\Big(\tfrac{1}{\alpha}\big(r(\boldsymbol{x}_0^k) - c_0\big)\Big) \tag{15}$$

$$= e^{-c_0/\alpha}\; e^{r(\boldsymbol{x}_0^k)/\alpha}. \tag{16}$$

The second equality follows from the cancellation of successive increments; the third uses the terminal identity $V(\boldsymbol{z}_0^k) = r(\boldsymbol{x}_0^k)$ and writes $c_0 := V(\boldsymbol{z}_T^k)$, a constant shared by all particles because the chain starts from the fully masked state.

Therefore, the accumulated weight is a function of $\boldsymbol{x}_0^k$ alone, up to a factor common to every particle.

Step 2 (the potential is positive and bounded). Since $V(\boldsymbol{z})$ is a value of r at the greedy sequence, $\|V\|_\infty \le \|r\|_\infty$. Therefore,

$$w_t^k \le e^{2\|r\|_\infty/\alpha}, \qquad G(\boldsymbol{x}_0) := e^{r(\boldsymbol{x}_0)/\alpha} \in (0, \infty), \tag{17}$$

so the per-step weights are positive and uniformly bounded, as the cited convergence result requires, and the common factor $e^{-c_0/\alpha}$ cancels under self-normalization.

Step 3 (Feynman–Kac limit). The scheme is therefore a Feynman–Kac particle model with mutation kernel q_γ and per-step potentials w_t, whose successive factors cancel to leave the terminal potential G. The intermediate resampling changes the genealogy but not the terminal measure, which is

$$\eta(\mathrm{d}\boldsymbol{x}_0) = \frac{q_\gamma(\mathrm{d}\boldsymbol{x}_0)\, G(\boldsymbol{x}_0)}{\int q_\gamma(\mathrm{d}\boldsymbol{x}'_0)\, G(\boldsymbol{x}'_0)}, \tag{18}$$

which is exactly

$$\eta(\boldsymbol{x}_0) \;\propto\; q_\gamma(\boldsymbol{x}_0)\, e^{r(\boldsymbol{x}_0)/\alpha} \;=\; \pi_\gamma \tag{19}$$

of Eq. (11). The accumulated potential enters as the density of this Feynman–Kac measure relative to the proposal, not as a weight carried past the selection: resampling at every step leaves the surviving particles unweighted, which is what Algorithm 4 returns. Since G is bounded and strictly positive and resampling is unbiased, the particle estimator is consistent (Del Moral, 2004, Ch. 7). Therefore, for every bounded measurable φ,

$$\tfrac{1}{N}\sum_k \varphi(\boldsymbol{x}_0^k) \;\xrightarrow[N\to\infty]{\text{a.s.}}\; \eta(\varphi) = \mathbb{E}_{\pi_\gamma}[\varphi]. \tag{20}$$

Finally, at $\gamma = 0$ the guided reveal reduces to the base kernel, $q_0 = p_\theta$, so

$$\pi_0 \;\propto\; p_\theta\, e^{r/\alpha} \;=\; p^\star \tag{21}$$

of Eq. (3) with $\beta = \alpha$, which completes Part (i).

Part (ii). The guided reveal renormalizes a strictly positive tilt of p_θ inside the vocabulary, so $q_\gamma(\cdot \mid \boldsymbol{z}) > 0$ wherever $p_\theta(\cdot \mid \boldsymbol{z}) > 0$, and the density ratio is bounded at a step revealing m_t positions by $e^{2m_t\|\gamma \boldsymbol{g}\|_\infty}$, hence along the whole path by

$$\frac{\mathrm{d}p_\theta}{\mathrm{d}q_\gamma}(\boldsymbol{z}_{T:0}) \;\le\; e^{2L\|\gamma \boldsymbol{g}\|_\infty} \;<\; \infty. \tag{22}$$

Augmenting the incremental weight by the proposal correction gives

$$\tilde{w}_t^k = w_t^k\, \frac{p_\theta(\boldsymbol{z}_s^k \mid \boldsymbol{z}_t^k)}{q_\gamma(\boldsymbol{z}_s^k \mid \boldsymbol{z}_t^k)}, \qquad \prod_t \tilde{w}_t^k = e^{-c_0/\alpha}\, e^{r(\boldsymbol{x}_0^k)/\alpha}\, \frac{\mathrm{d}p_\theta}{\mathrm{d}q_\gamma}(\boldsymbol{z}_{T:0}^k), \tag{23}$$

so the Feynman–Kac model of Steps 1–3 is run with mutation kernel q_γ and path potential $G \cdot \mathrm{d}p_\theta/\mathrm{d}q_\gamma$. Its normalized terminal measure is obtained by integrating the path measure over the intermediate states, and the ratio is a path quantity rather than a ratio of terminal marginals, so the integral is taken before it is cancelled:

$$\begin{aligned}\eta(\boldsymbol{x}_0) \;&\propto\; e^{r(\boldsymbol{x}_0)/\alpha} \int q_\gamma(\boldsymbol{z}_{T:0})\, \frac{\mathrm{d}p_\theta}{\mathrm{d}q_\gamma}(\boldsymbol{z}_{T:0})\, \mathrm{d}\boldsymbol{z}_{T:1} \;=\; e^{r(\boldsymbol{x}_0)/\alpha} \int p_\theta(\boldsymbol{z}_{T:0})\, \mathrm{d}\boldsymbol{z}_{T:1} \\ &=\; p_\theta(\boldsymbol{x}_0)\, e^{r(\boldsymbol{x}_0)/\alpha} \;=\; p^\star. \end{aligned} \tag{24}$$

The correction is given in closed form in Eq. (65).

Therefore, the target is $p^\star$ of Eq. (3) with $\beta = \alpha$ for every γ, which is Part (ii). □

Proposition 1 (Effective temperature of the guided proposal, restated). *Under* (A1)–(A3) *below, let $E(\gamma)$ be the bound of Eq. (33), which is $O(\gamma)$. Then the terminal log-twist satisfies $\Phi_\gamma(\boldsymbol{x}_0) = \kappa\gamma\, r(\boldsymbol{x}_0) + c + \varepsilon(\boldsymbol{x}_0)$ with c independent of $\boldsymbol{x}_0$, $|\varepsilon| \le E(\gamma)$, and $\kappa > 0$ any tangential scale for the estimator that carries the reward gradient into the clean-token logits, against which δ_P is measured. Consequently $\pi_\gamma(\boldsymbol{x}_0) \propto p_\theta(\boldsymbol{x}_0)\, e^{(\kappa\gamma + 1/\alpha) r(\boldsymbol{x}_0) + \varepsilon(\boldsymbol{x}_0)}$, so omitting the proposal correction leaves the target in the reward-tilted family of Eq. (3) at inverse temperature $\kappa\gamma + 1/\alpha$, exactly when ε is constant in $\boldsymbol{x}_0$ and up to the factor e^{ε} otherwise. At $\gamma = 0$, $E(0) = 0$ and $\beta = \alpha$ exactly.*

Notation. Let $\tilde{r}$ be the relaxation of r to the product simplex that the guidance path differentiates, so that $\tilde{r}(\mathrm{onehot}(\boldsymbol{x})) = r(\boldsymbol{x})$. Write the soft representation of a state $\boldsymbol{z}$ as $S(\boldsymbol{z})$, with row ℓ equal to $\boldsymbol{e}_{z^\ell}$ at a revealed position and to $\boldsymbol{p}^\ell = \mathrm{softmax}(\boldsymbol{\eta}^\ell)$ at a masked one, and set $\tilde{V}(\boldsymbol{z}) = \tilde{r}(S(\boldsymbol{z}))$, so that $\tilde{V}(\boldsymbol{z}_0) = r(\boldsymbol{x}_0)$. It is distinct from the selector's greedy value V of §3.2; the two agree at $\boldsymbol{z}_0$. Write $\boldsymbol{u}^\ell$ for the simplex coordinate at position ℓ, that is row ℓ of $S(\boldsymbol{z})$, and let $\boldsymbol{h}^\ell = \nabla_{\boldsymbol{u}^\ell} \tilde{r}(S(\boldsymbol{z}))$

be the simplex gradient at position ℓ, with $\boldsymbol{h}$ the concatenation of the $\boldsymbol{h}^\ell$ over positions. The chain starts from the deterministic all-MASK state, so $\tilde{V}(\boldsymbol{z}_T)$ is common to every trajectory.

Two log-twists. The *trajectory* twist $\bar{\Phi}_\gamma(\boldsymbol{z}_{T:0}) = \log \frac{q_\gamma(\boldsymbol{z}_{T:0})}{p_\theta(\boldsymbol{z}_{T:0})}$ is what the guided reveal accumulates step by step; the *terminal* twist $\Phi_\gamma(\boldsymbol{x}_0)$ of Eq. (10) is what Eq. (11) needs. Steps 1–4 bound the first and Step 5 passes to the second.

Assumptions. We assume:

(A1) *Estimator form.* $\mathbb{E}[\tilde{\boldsymbol{g}}^\ell] = \boldsymbol{g}^\ell = \boldsymbol{P}^\ell \boldsymbol{h}^\ell$, the expectation being over the estimator's own randomness. Let Π^ℓ be the orthogonal projector onto the tangent space of the simplex at $\boldsymbol{p}^\ell$. Take $\kappa > 0$ to be the scale that minimizes the anisotropy, $\kappa = \arg\min_{\kappa'>0} \sup_\ell \|\boldsymbol{P}^\ell - \kappa'\Pi^\ell\|_{\mathrm{op}}$, so that κ and hence $\kappa\gamma + 1/\alpha$ are determined rather than free, and set the anisotropy $\delta_P = \sup_\ell \|\boldsymbol{P}^\ell - \kappa\Pi^\ell\|_{\mathrm{op}}$, one pair for all positions and steps.

(A2) *Smoothness.* $\nabla\tilde{r}$ is L_2-Lipschitz on the product simplex.

(A3) *Boundedness.* $\|\tilde{\boldsymbol{g}}^\ell\|_\infty \le G$, and write $\Delta_g = \max_\ell \|\tilde{\boldsymbol{g}}^\ell - \boldsymbol{g}^\ell\|_\infty$, so $\Delta_g \le 2G$.

Proof. Steps 1 and 2 make the trajectory twist explicit and expand it in γ; Step 3 replaces the realized estimator by κ times the simplex gradient, which is where the reward enters linearly; Step 4 sums over positions to produce $r(\boldsymbol{x}_0) - \tilde{V}(\boldsymbol{z}_T)$; Step 5 passes to the terminal twist.

Step 1 (exact, along a trajectory). Eq. (25) below is an identity between path densities, not between terminal marginals; the two are related in Step 5. Every position is revealed exactly once. Summing Eq. (65) over the trajectory, negating, and writing $k_\ell = x_0^\ell$ for the token committed at position ℓ, we obtain

$$\bar{\Phi}_\gamma(\boldsymbol{z}_{T:0}) = \sum_{\ell=1}^{L} \Big[\gamma \tilde{g}^\ell_{k_\ell} - \big(\mathrm{logsumexp}(\boldsymbol{\eta}^\ell + \gamma\tilde{\boldsymbol{g}}^\ell) - \mathrm{logsumexp}(\boldsymbol{\eta}^\ell)\big)\Big], \tag{25}$$

with $\boldsymbol{\eta}^\ell$ evaluated at the step at which ℓ was revealed. Eq. (65) is written in the gauge-shifted $\hat{\boldsymbol{g}} = \boldsymbol{g} - c\mathbf{1}$; since both terms above shift together under $\boldsymbol{g} \mapsto \boldsymbol{g} + c\mathbf{1}$, the realized estimator $\tilde{\boldsymbol{g}}$ may be substituted for it without changing the value.

Therefore, $\bar{\Phi}_\gamma$ is exactly the displayed sum over positions.

Step 2 (expansion). Since $\mathrm{logsumexp}(\boldsymbol{\eta}+\gamma\tilde{\boldsymbol{g}}) - \mathrm{logsumexp}(\boldsymbol{\eta}) = \gamma\langle \boldsymbol{p}, \tilde{\boldsymbol{g}}\rangle + \frac{\gamma^2}{2}\mathrm{Var}_{\boldsymbol{p}}(\tilde{\boldsymbol{g}}) + O(\gamma^3)$, Eq. (25) gives

$$\bar{\Phi}_\gamma = \gamma \sum_\ell \big\langle \tilde{\boldsymbol{g}}^\ell,\ \boldsymbol{e}_{k_\ell} - \boldsymbol{p}^\ell \big\rangle - \frac{\gamma^2}{2} \sum_\ell \mathrm{Var}_{\boldsymbol{p}^\ell}(\tilde{\boldsymbol{g}}^\ell) + O(\gamma^3), \tag{26}$$

still in the realized $\tilde{\boldsymbol{g}}$. Both terms are invariant under $\tilde{\boldsymbol{g}} \mapsto \tilde{\boldsymbol{g}} + c\mathbf{1}$, the gauge freedom of the softmax parameterization; the constant in $\tilde{\boldsymbol{g}}$ therefore does not reach π_γ.

Therefore, to second order, $\bar{\Phi}_\gamma$ is the displayed linear form in the realized estimator.

Step 3 (identification). Split the realized estimator into its mean and its deviation. Since $\|\boldsymbol{e}_k - \boldsymbol{p}^\ell\|_1 \le 2$, Hölder's inequality gives

$$\big|\langle \tilde{\boldsymbol{g}}^\ell - \boldsymbol{g}^\ell,\ \boldsymbol{e}_k - \boldsymbol{p}^\ell\rangle\big| \ \le\ 2\|\tilde{\boldsymbol{g}}^\ell - \boldsymbol{g}^\ell\|_\infty \ \le\ 2\Delta_g. \tag{27}$$

Accordingly, the deviation contributes at most $2\gamma L\Delta_g$ in total. For the mean, Π^ℓ is an orthogonal projector, hence self-adjoint, and $\boldsymbol{e}_k - \boldsymbol{p}^\ell$ lies in its range, so $\langle \kappa\Pi^\ell\boldsymbol{h}^\ell, \boldsymbol{e}_k - \boldsymbol{p}^\ell\rangle = \kappa\langle \boldsymbol{h}^\ell, \boldsymbol{e}_k - \boldsymbol{p}^\ell\rangle$. Therefore, by (A1) and $\|\boldsymbol{e}_k - \boldsymbol{p}^\ell\| \le \sqrt{2}$,

$$\big|\langle \boldsymbol{g}^\ell, \boldsymbol{e}_k - \boldsymbol{p}^\ell\rangle - \kappa\langle \boldsymbol{h}^\ell, \boldsymbol{e}_k - \boldsymbol{p}^\ell\rangle\big| \ \le\ \sqrt{2}\,\delta_P\|\boldsymbol{h}^\ell\|. \tag{28}$$

Therefore, the realized estimator may be replaced by $\kappa\boldsymbol{h}^\ell$, at a cost of $2\Delta_g$ per position from the deviation and $\sqrt{2}\,\delta_P\|\boldsymbol{h}^\ell\|$ from the anisotropy. This is where the reward gradient enters, and with it the linear dependence on r.

Step 4a (per-step bounds). Let m_t be the number of positions revealed at step t, so $\sum_t m_t = L$. Let $\boldsymbol{d}_t$ collect the displacements $\boldsymbol{e}_{k_\ell} - \boldsymbol{p}^\ell$ of those positions, zero-padded to all L rows, and let

$\boldsymbol{D}_t = S(\boldsymbol{z}_s) - S(\boldsymbol{z}_t)$ be the full displacement of the soft representation. The two differ because revealing a token re-runs the denoiser, so the rows that stay masked move as well; write

$$\delta_S = \sum_t \big\|\boldsymbol{D}_t - \boldsymbol{d}_t\big\|, \tag{29}$$

a quantity determined by the trajectory.

The linear form is additive across positions, so $\sum_{\ell\in t}\langle \boldsymbol{h}^\ell, \boldsymbol{e}_{k_\ell} - \boldsymbol{p}^\ell\rangle = \langle \boldsymbol{h}, \boldsymbol{d}_t\rangle$ exactly. By (A2),

$$\big|\langle \boldsymbol{h}, \boldsymbol{D}_t\rangle - \big(\tilde{V}(\boldsymbol{z}_s) - \tilde{V}(\boldsymbol{z}_t)\big)\big| \;\le\; \tfrac{L_2}{2}\|\boldsymbol{D}_t\|^2, \tag{30}$$

and by Cauchy–Schwarz,

$$\big|\langle \boldsymbol{h}, \boldsymbol{D}_t\rangle - \langle \boldsymbol{h}, \boldsymbol{d}_t\rangle\big| \;\le\; \|\boldsymbol{h}\|\,\|\boldsymbol{D}_t - \boldsymbol{d}_t\|, \qquad \|\boldsymbol{h}\| = \Big(\textstyle\sum_\ell \|\boldsymbol{h}^\ell\|^2\Big)^{1/2}. \tag{31}$$

Therefore, at each step, the linear form differs from the increment of $\tilde{V}$ by a curvature term and a clean-prediction drift term.

Step 4b (accumulation). The blocks of $\boldsymbol{d}_t$ are orthogonal and $\|\boldsymbol{e}_k - \boldsymbol{p}\|^2 \le 2$, so $\|\boldsymbol{d}_t\|^2 \le 2m_t$. Therefore, $\|\boldsymbol{D}_t\|^2 \le 2\|\boldsymbol{d}_t\|^2 + 2\|\boldsymbol{D}_t - \boldsymbol{d}_t\|^2$, and summing over steps,

$$\sum_t \|\boldsymbol{D}_t\|^2 \;\le\; 4\sum_t m_t + 2\sum_t \|\boldsymbol{D}_t - \boldsymbol{d}_t\|^2 \;\le\; 4L + 2\delta_S^2, \tag{32}$$

where the second inequality uses $\sum_t a_t^2 \le (\sum_t a_t)^2$ for $a_t \ge 0$.

The increments cancel along the trajectory, leaving $\sum_t \big(\tilde{V}(\boldsymbol{z}_s) - \tilde{V}(\boldsymbol{z}_t)\big) = \tilde{V}(\boldsymbol{z}_0) - \tilde{V}(\boldsymbol{z}_T) = r(\boldsymbol{x}_0) - \tilde{V}(\boldsymbol{z}_T)$, with $\tilde{V}(\boldsymbol{z}_T)$ the value of the all-MASK state, common to every trajectory. Collecting the remainders of Steps 2–4 and taking the supremum over trajectories, we obtain

$$\sup\Big|\bar{\Phi}_\gamma - \kappa\gamma\big(r(\boldsymbol{x}_0) - \tilde{V}(\boldsymbol{z}_T)\big)\Big| \;\le\; E(\gamma),$$

$$E(\gamma) = \gamma\kappa\Big(L_2\big(2L + \delta_S^2\big) + \|\boldsymbol{h}\|\,\delta_S\Big) + \gamma\Big(\sqrt{2}\,\delta_P\,\textstyle\sum_\ell \|\boldsymbol{h}^\ell\| + 2L\Delta_g\Big) + \tfrac{\gamma^2}{2}LG^2 + O\big(\gamma^3 LG^3\big), \tag{33}$$

where $\|\boldsymbol{h}\|$, $\sum_\ell \|\boldsymbol{h}^\ell\|$, δ_S and Δ_g are read at their trajectory suprema, each finite because the product simplex is compact and T is finite.

Therefore, $\bar{\Phi}_\gamma$ equals $\kappa\gamma(r(\boldsymbol{x}_0) - \tilde{V}(\boldsymbol{z}_T))$ up to $E(\gamma)$, uniformly over trajectories.

Step 5 (trajectory to terminal). Marginalizing the path measure over the trajectories that end at $\boldsymbol{x}_0$ gives $e^{\Phi_\gamma(\boldsymbol{x}_0)} = \mathbb{E}_{p_\theta(\boldsymbol{z}_{T:0}|\boldsymbol{x}_0)}\big[e^{\bar{\Phi}_\gamma}\big]$. The leading term of Eq. (33) is a function of $\boldsymbol{x}_0$ alone, hence factors out of the expectation, and the remainder is bounded uniformly over trajectories. Therefore,

$$\big|\Phi_\gamma(\boldsymbol{x}_0) - \kappa\gamma\big(r(\boldsymbol{x}_0) - \tilde{V}(\boldsymbol{z}_T)\big)\big| \;\le\; \log\mathbb{E}\big[e^{|\bar{\Phi}_\gamma - \kappa\gamma(r - \tilde{V}(\boldsymbol{z}_T))|}\big] \;\le\; E(\gamma). \tag{34}$$

Therefore, $\Phi_\gamma(\boldsymbol{x}_0) = \kappa\gamma\, r(\boldsymbol{x}_0) + c + \varepsilon(\boldsymbol{x}_0)$ with $c = -\kappa\gamma\tilde{V}(\boldsymbol{z}_T)$ and $|\varepsilon| \le E(\gamma)$, which is Proposition 1. □

Remark 1 (What the bound does and does not give). *$E(\gamma)$ is $O(\gamma)$ as $\gamma \to 0$ and grows with the sequence length L, while the leading term does not. The bound therefore constrains the* form *of the relation, namely that Φ_γ is linear in r with slope $\kappa\gamma$, rather than certifying a small relative error at large L, and it does not predict κ. Of the four controlling quantities, the reductions of §3.1 act on Δ_g, so the variance reduction and the omitted correction bear on the same quantity.*

Lemma 2 (Baseline invariance of the resampling law, restated with the leave-one-out case). *Fix a scale $c > 0$, and for increments $\{\Delta^k\}_{k=1}^N \subset \mathbb{R}$ and baselines $\{b^k\}_{k=1}^N$ let $\hat{w}^k \propto \exp\big((\Delta^k - b^k)/c\big)$, with $\mu = \frac{1}{N}\sum_j \Delta^j$. Then* (i) *for a constant baseline $b^k \equiv b$, $\hat{w}^k = \mathrm{softmax}_k(\Delta^k/c)$, independent of b and a function of the centered increments $\Delta^k - \mu$ only; and* (ii) *for the leave-one-out baseline $b^k = \frac{1}{N-1}\sum_{j\neq k}\Delta^j$ with bounded increments, $\log\hat{w}^k$ equals its value in* (i) *up to an additive $O\big(1/(N-1)\big)$.*

Proof. **Part (i): constant baseline.** With $b^k \equiv b$, the factor $e^{-b/c}$ appears in both numerator and denominator and cancels:

$$\hat{w}^k = \frac{e^{(\Delta^k - b)/c}}{\sum_j e^{(\Delta^j - b)/c}} = \frac{e^{\Delta^k/c}}{\sum_j e^{\Delta^j/c}} = \mathrm{softmax}_k\big(\Delta^k/c\big). \tag{35}$$

Thus $\hat{w}^k$ is independent of b. Since softmax is invariant under a common additive shift of its arguments,

$$\mathrm{softmax}_k\big(\Delta^k/c\big) = \mathrm{softmax}_k\big((\Delta^k - \mu)/c\big), \tag{36}$$

so the law depends on $\{\Delta^k\}$ only through the centered increments $\{\Delta^k - \mu\}$.

Part (ii): leave-one-out baseline. Rewrite the baseline in terms of the group mean μ:

$$b^k = \frac{1}{N-1}\sum_{j \neq k} \Delta^j \tag{37}$$

$$= \frac{1}{N-1}\big(N\mu - \Delta^k\big) \tag{38}$$

$$= \mu + \frac{\mu - \Delta^k}{N-1}. \tag{39}$$

Define $\rho^k := \dfrac{\mu - \Delta^k}{(N-1)c}$. Then the exponent differs from part (i) by exactly ρ^k,

$$\frac{\Delta^k - b^k}{c} = \frac{\Delta^k - \mu}{c} - \rho^k, \tag{40}$$

and if $|\Delta^k| \le B$ for all k,

$$|\rho^k| \le \frac{|\mu| + |\Delta^k|}{(N-1)c} \le \frac{2B}{(N-1)c} = O\big(1/(N{-}1)\big). \tag{41}$$

Let $\ell^k := \log \hat{w}^k$ under the leave-one-out baseline and ℓ_0^k its value from part (i). Subtracting the two normalized log-weights gives

$$\ell^k - \ell_0^k = -\rho^k + \log \frac{\sum_j e^{(\Delta^j - \mu)/c}}{\sum_j e^{(\Delta^j - \mu)/c}\, e^{-\rho^j}}. \tag{42}$$

Because each summand in the denominator is the corresponding numerator summand scaled by $e^{-\rho^j} \in \big[e^{-\max_j |\rho^j|},\, e^{\max_j |\rho^j|}\big]$, the log-ratio lies in $\big[-\max_j |\rho^j|,\, \max_j |\rho^j|\big]$, so

$$\big|\ell^k - \ell_0^k\big| \le 2 \max_j |\rho^j| = O\big(1/(N{-}1)\big). \tag{43}$$

Hence, the resampling law is exactly invariant to any constant baseline, and invariant to the leave-one-out baseline up to $O(1/(N{-}1))$. □

Remark 2 (Leave-one-out baseline in the selector)**.** *Replacing the group mean μ by the leave-one-out baseline changes $\log \hat{w}^k$ by $O\big(1/(N{-}1)\big)$, so the selector is insensitive to which of the two is used. This is the prediction tested in Table 11. Since the group-mean shift cancels exactly, the only operative effect of standardization on the resampling law is the replacement $\alpha \mapsto \alpha\sigma$.*

Proposition 2 (Target of AMC, restated). *Assume $\|r\|_\infty < \infty$, $\alpha > 0$, $\sigma_t \ge \epsilon > 0$, and multinomial resampling at every step, and assume as in Lemma 1 that the chain starts from the deterministic all-*MASK *state and that $V(\boldsymbol{z}_0) = r(\boldsymbol{x}_0)$. The accumulated weight of a particle is proportional to e^{R_σ} with $R_\sigma = \sum_t \Delta_t/\alpha_t$ and $\alpha_t = \alpha\sigma_t$.* (i) *Let φ be bounded and measurable, and assume the empirical μ_t, σ_t admit deterministic limits $\mu_t^\star, \sigma_t^\star$ and that the map from the empirical measure to (μ_t, σ_t) satisfies the regularity conditions of Beskos et al. (2016). Let $\pi^\star_{\mathrm{AMC}} \propto q_\gamma e^{R_{\sigma^\star}}$ be the path measure built from those limits. Then the self-normalized* AMC *estimator converges* in probability *to $\mathbb{E}_{\pi^\star_{\mathrm{AMC}}}[\varphi(\boldsymbol{x}_0)]$.* (ii) *If $\sigma_t^\star \equiv \sigma$, then the $\boldsymbol{x}_0$-marginal of $\pi^\star_{\mathrm{AMC}}$ is π_γ at temperature $\alpha\sigma$.*

Setup. At each step AMC alternates a mutation (one guided reveal $\boldsymbol{z}_s^k \sim q_\gamma(\cdot \mid \boldsymbol{z}_t^k)$ of Lemma 1) with a selection that resamples the N particles with probabilities proportional to the potential

$$G_t(\boldsymbol{z}_s^k) = \exp\Big(\tfrac{\Delta_t^k - \mu_t}{\alpha\sigma_t}\Big), \qquad \Delta_t^k = V(\boldsymbol{z}_s^k) - V(\boldsymbol{z}_t^k), \tag{44}$$

with μ_t, σ_t the mean and standard deviation of $\{\Delta_t^k\}_{k=1}^N$, so the scheme is a Feynman–Kac model with per-step potentials G_t. The accumulated potential $G_{0:T}^k = \prod_t w_t^k$ is the density of the target path measure relative to the proposal; with resampling at every step the surviving particles are unweighted.

Proof. **Step 1 (accumulated weight).** The population constants $e^{-\mu_t/\alpha_t}$ are common to the N particles at step t, so they cancel in $G_{0:T}^k$ by Lemma 2(i). Therefore,

$$G_{0:T}^k \;\propto\; \exp\big(R_\sigma^k\big), \qquad R_\sigma^k = \sum_t \frac{\Delta_t^k}{\alpha_t}, \qquad \alpha_t = \alpha\sigma_t, \tag{45}$$

which is the first claim of the proposition. Because α_t varies with t, the increments do not cancel and R_σ is a functional of the whole trajectory, so the reweighted measure lives on paths; the estimator averages $\varphi(\boldsymbol{x}_0)$, a bounded function of the terminal state, so its limit is an expectation under the $\boldsymbol{x}_0$-marginal of that path measure.

Step 2: the potentials are bounded above and below. Since V is bounded by $\|r\|_\infty$, both Δ_t^k and μ_t lie in $[-2\|r\|_\infty, 2\|r\|_\infty]$, so

$$\big|\Delta_t^k - \mu_t\big| \;\le\; 4\|r\|_\infty. \tag{46}$$

The floor $\sigma_t \ge \epsilon$ gives $\alpha\sigma_t \ge \alpha\epsilon > 0$. Therefore,

$$0 \;<\; e^{-4\|r\|_\infty/(\alpha\epsilon)} \;\le\; G_t \;\le\; e^{4\|r\|_\infty/(\alpha\epsilon)} \;<\; \infty. \tag{47}$$

Step 3: placing the scheme in the framework of Beskos et al. (2016). Their framework covers a Feynman–Kac flow whose potential at step t is parameterized by a bounded summary statistic of the particle population that is about to be reweighted, resampled multinomially at every step. It asks for three things.

(a) *A bounded summary statistic.* Ours is (μ_t, σ_t), the first two empirical moments of $\{\Delta_t^k\}$, bounded by Step 2 through $\|r\|_\infty < \infty$ and the floor ϵ. *Verified.*

(b) *Multinomial resampling at every step.* A hypothesis of the proposition. *Assumed as a hypothesis.*

(c) *Deterministic limits and regularity.* That the empirical moments admit deterministic limits, and that the map from the empirical measure to (μ_t, σ_t), which is non-linear in σ_t, satisfies their regularity conditions. *Assumed, and carried in the hypotheses of part (i).*

Therefore, by their Theorems 3.1 and 3.2, for every bounded φ,

$$\tfrac{1}{N}\sum_k \varphi(\boldsymbol{x}_0^k) \;\xrightarrow[N\to\infty]{\mathbb{P}}\; \mathbb{E}_{\pi^\star_{\mathrm{AMC}}}[\varphi], \tag{48}$$

Therefore, the self-normalized estimator converges in probability to $\mathbb{E}_{\pi^\star_{\mathrm{AMC}}}[\varphi]$ under the $\boldsymbol{x}_0$-marginal of $\pi^\star_{\mathrm{AMC}}$, which is part (i).

Step 4: where our statistic sits. Their adaptation uses the population that is about to be reweighted. The same holds here:

$$(\mu_t, \sigma_t) = \text{moments of } \{\Delta_t^k\}_{k=1}^N, \qquad \Delta_t^k = V(\boldsymbol{z}_s^k) - V(\boldsymbol{z}_t^k), \qquad \hat{w}^k \propto G_t(\boldsymbol{z}_s^k), \tag{49}$$

so the statistic is read off the mutated states $\{\boldsymbol{z}_s^k\}$ and the potential is applied to those same N particles. Accordingly, the statistic and the reweighting act on one population, as their structure requires it.

Step 5: the constant-σ case. Let $\sigma_t^\star \equiv \sigma$, and evaluate the limiting potentials at that value. Then

$$G_t(\boldsymbol{z}) = \exp\Big(\tfrac{\Delta_t(\boldsymbol{z}) - \mu_t}{\alpha\sigma}\Big) = e^{-\mu_t/(\alpha\sigma)}\,\exp\Big(\tfrac{\Delta_t(\boldsymbol{z})}{\alpha\sigma}\Big) \;\propto\; \exp\Big(\tfrac{\Delta_t(\boldsymbol{z})}{\alpha\sigma}\Big), \tag{50}$$

the last step because the population constant $e^{-\mu_t/(\alpha\sigma)}$ cancels in the resampling softmax by Lemma 2(i). This is the fixed-temperature potential of Lemma 1 with α replaced by $\alpha\sigma$, whose increments cancel along the trajectory. Therefore,

$$R_\sigma = \sum_t \frac{\Delta_t}{\alpha\sigma} = \frac{r(\boldsymbol{x}_0) - V(\boldsymbol{z}_T)}{\alpha\sigma}, \tag{51}$$

which is $r/(\alpha\sigma)$ up to an additive constant.

Therefore, the $\boldsymbol{x}_0$-marginal of $\pi^\star_{\mathrm{AMC}}$ is π_γ at temperature $\alpha\sigma$, which is part (ii). □

Lemma 3 (Optimal constant baseline for the score-function correction)**.** *Let $x \sim \boldsymbol{p}$ on a finite vocabulary with $\boldsymbol{p}$ non-degenerate, $\hat{\boldsymbol{x}} = \mathrm{onehot}(x)$ and $R = r(x)$. For $b \in \mathbb{R}$ define the single-sample correction $\boldsymbol{u}(b) = (R - b)(\hat{\boldsymbol{x}} - \boldsymbol{p})$ and write $Q = \|\hat{\boldsymbol{x}} - \boldsymbol{p}\|^2$, so $\mathbb{E}[Q] > 0$. Then* (i) *$\mathbb{E}[\boldsymbol{u}(b)] = \nabla_{\boldsymbol{\eta}}\mathbb{E}[r]$ for every b; and* (ii) *the variance of $\boldsymbol{u}(b)$ is minimized at $b^\star = \mathbb{E}[RQ]/\mathbb{E}[Q]$, and if $b^\star > 0$ it is strictly smaller than that of $\boldsymbol{u}(0)$ for every $b \in (0, 2b^\star)$.*

Part (i) is the Reinforce unbiasedness of Williams (1992, Thm. 1) and part (ii) is Greensmith et al. (2004, Thm. 11) specialized to a categorical distribution, where $\nabla_{\boldsymbol{\eta}} \log p(x) = \hat{\boldsymbol{x}} - \boldsymbol{p}$; the interval $(0, 2b^\star)$ is a one-line consequence of the excess-variance identity in that theorem rather than a separate result of theirs. We give the short proof in our notation.

Proof. **Part (i).** Since $\mathbb{E}[\hat{\boldsymbol{x}} - \boldsymbol{p}] = \mathbf{0}$, the baseline term vanishes. Therefore,

$$\nabla_{\eta_j}\mathbb{E}[r] = \sum_k r_k\, p_k(\delta_{kj} - p_j) = \mathbb{E}\big[R(\hat{x}_j - p_j)\big] = \mathbb{E}\big[u_j(b)\big], \tag{52}$$

the second equality being the softmax Jacobian $\partial p_k/\partial\eta_j = p_k(\delta_{kj} - p_j)$.

Part (ii). By part (i) the mean of $\boldsymbol{u}(b)$ does not depend on b, so comparing variances is comparing second moments. Accordingly,

$$\mathbb{E}\|\boldsymbol{u}(b)\|^2 = \mathbb{E}\big[(R - b)^2 Q\big] = \mathbb{E}[R^2 Q] - 2b\,\mathbb{E}[RQ] + b^2\,\mathbb{E}[Q], \tag{53}$$

a quadratic in b, strictly convex because $\mathbb{E}[Q] > 0$, whose minimizer is $b^\star = \mathbb{E}[RQ]/\mathbb{E}[Q]$. Subtracting the value at $b = 0$ gives

$$\mathbb{E}\|\boldsymbol{u}(b)\|^2 - \mathbb{E}\|\boldsymbol{u}(0)\|^2 = b\,\mathbb{E}[Q]\,\big(b - 2b^\star\big), \tag{54}$$

which is negative exactly when $0 < b < 2b^\star$. □

Remark 3 (Leave-one-out as a plug-in)**.** *The leave-one-out baseline estimates $\mathbb{E}[R]$ from the remaining samples of the group. When R and Q are uncorrelated, $b^\star = \mathbb{E}[R]$, so it is then an unbiased plug-in for the variance-optimal constant baseline; in general it removes the reward's mean offset without matching $b^\star$ exactly, and part (ii) requires $b^\star > 0$. Two further gaps separate the lemma from $\boldsymbol{g}_{\mathrm{RO}}$ of Eq. (9): its baseline is random rather than constant, and the n leave-one-out baselines are coupled, so the variance of the averaged estimator carries cross terms the single-sample comparison does not see. Unbiasedness survives both, since b^i is independent of sample i; the variance reduction is the motivation for the estimator, and Table 11 is what tests it.*

The last result concerns the placement of the tilt and is used in Appendix F.3, where the two proposal forms it compares, Eq. (63) and Eq. (64), are derived.

Proposition 3 (Schedule leakage)**.** *Fix a masked position ℓ and a step with $a_t \in (0, 1)$, let $\gamma > 0$, and write $Z^\ell = \mathbb{E}_{p^\ell_\theta}[e^{\gamma g}]$ for the tilt normalizer of Eq. (63). Let $\tilde{a}^\ell_t = 1 - q^\ell(\mathrm{MASK})$ be the reveal probability actually induced by the guided proposal. Then*

$$\tilde{a}^{\ell,\,\mathrm{post}}_t = \frac{a_t Z^\ell}{1 - a_t + a_t Z^\ell}, \qquad \tilde{a}^{\ell,\,\mathrm{ours}}_t = a_t. \tag{55}$$

Consequently:

(a) $\tilde{a}^{\ell,\,\mathrm{post}}_t$ is strictly increasing in Z^ℓ, with $\tilde{a}^{\ell,\,\mathrm{post}}_t \gtrless a_t$ according as $Z^\ell \gtrless 1$.

(b) Under the gauge $\mathbb{E}_{p_\theta^\ell}[g] = 0$, Jensen's inequality gives $Z^\ell \geq 1$, with equality iff g is p_θ^ℓ-almost surely constant, so unless g is degenerate guidance strictly accelerates unmasking.

(b′) Under that same gauge, Z^ℓ is nondecreasing in γ, so by part (a) the acceleration grows with γ by a state- and step-dependent amount.

(c) q^ℓ of Eq. (64) is exactly invariant under $\boldsymbol{g}^\ell \mapsto \boldsymbol{g}^\ell + c\mathbf{1}$, whereas $\tilde{a}_t^{\ell,\,\mathrm{post}} \to 1$ as $c \to +\infty$ and $\to 0$ as $c \to -\infty$.

Therefore, the post-hoc reveal probability is not determined by the guidance signal itself, whereas the logit-space one is.

Proof. Both lines of Eq. (55) are read off Eq. (63) and Eq. (64).

Part (a). Write $f(z) = a_t z/(1 - a_t + a_t z)$. Then

$$f'(z) = \frac{a_t(1-a_t)}{(1-a_t+a_t z)^2} \;>\; 0 \qquad \text{for } a_t \in (0,1), \tag{56}$$

and $f(1) = a_t$. Therefore, f is strictly increasing and crosses a_t at $z = 1$, which is the claim.

Part (b). By Jensen's inequality and the gauge $\mathbb{E}_{p_\theta^\ell}[g] = 0$,

$$Z^\ell = \mathbb{E}\big[e^{\gamma g}\big] \;\geq\; e^{\gamma \mathbb{E}[g]} = 1, \tag{57}$$

with equality if and only if γg is degenerate.

Part (b′). For the dependence on γ, let $\Lambda(\gamma) = \log \mathbb{E}[e^{\gamma g}]$ be the cumulant generating function. Accordingly,

$$\Lambda(0) = 0, \qquad \Lambda'(0) = \mathbb{E}[g] = 0, \qquad \Lambda''(\gamma) = \mathrm{Var}_{\text{tilted}}(g) \geq 0. \tag{58}$$

The third identity makes Λ' nondecreasing, so $\Lambda'(\gamma) \geq \Lambda'(0) = 0$ for $\gamma > 0$, and Λ is nondecreasing there, strictly so unless g is degenerate.

Part (c). The shift $\boldsymbol{g}^\ell \mapsto \boldsymbol{g}^\ell + c\mathbf{1}$ multiplies every $e^{\gamma g_k}$, and hence Z^ℓ, by $e^{\gamma c}$. In Eq. (64) the two factors cancel,

$$q^\ell(k)\big|_{\boldsymbol{g}^\ell + c\mathbf{1}} = \frac{a_t\, p_\theta^\ell(k)\, e^{\gamma(g_k + c)}}{e^{\gamma c} Z^\ell} = \frac{a_t\, p_\theta^\ell(k)\, e^{\gamma g_k}}{Z^\ell} = q^\ell(k), \tag{59}$$

so $\tilde{a}_t^{\ell,\,\mathrm{ours}} = a_t$ is unmoved. In Eq. (55) the factor survives,

$$\begin{aligned} \tilde{a}_t^{\ell,\,\mathrm{post}}\big|_{\boldsymbol{g}^\ell + c\mathbf{1}} &= \frac{a_t\, e^{\gamma c} Z^\ell}{1 - a_t + a_t\, e^{\gamma c} Z^\ell}, \\ &\longrightarrow\; 1 \;\; (c \to +\infty), \qquad \longrightarrow\; 0 \;\; (c \to -\infty). \end{aligned} \tag{60}$$

Therefore, a constant that the guidance signal does not fix moves the post-hoc reveal schedule across its whole range, while leaving the logit-space one at a_t. □

B Implementation Details

All methods are reproduced from their official repositories under a common setup, following the protocol of Wang et al. (2025). We carry N=20 particles or candidates for every search method, which is the particle count N of §3.2 for ours. Table 5 lists every hyperparameter of VGAS and its value on each domain. Gumbel–Rao lifts the protein *proposal* at no extra denoiser cost ($0.64 \to 0.70$), and we read M off the proposal, where the effect is resolved, adopting M=8 on both sequence benchmarks (Table 12).

Tuning protocol. All VGAS hyperparameters (γ, α, M, and the resampling frequency) were selected on a held-out validation split of generations that is disjoint from the evaluation set whose numbers we report; the reported metrics are never used for selection. Baselines are run at their official default configurations. This keeps the comparison free of test-set tuning and avoids re-tuning any baseline downward.

Table 5: **Hyperparameters of VGAS.** The guidance configuration follows the one GILC (Dou et al., 2026) releases for these benchmarks, with γ selected on the held-out validation split of §B; GILC is run at the same values wherever it is compared against. The selector's α and r-freq are shared across all three domains.

Symbol	Controls	DNA	Protein	QM9
T	denoising steps	128	50	100
N	particles carried by the selector	20	20	20
γ	guidance scale on the logit correction	11000	1000	10
n	Monte Carlo samples, VGAS-GR	10	5	5
n	Monte Carlo samples, VGAS-RO	20	20	20
M	Gumbel–Rao samples (VGAS-GR)	8	8	4
τ	Gumbel–Softmax straight-through temperature	1.0	1.0	1.0
α	base selection temperature, $\alpha_t = \alpha\sigma_t$	0.5	0.5	0.5
r-freq	steps between resampling events	1	1	1

Table 6: **Inference cost per denoising step**, in the convention of Dou et al. (2026, Tab. 2).

Method	Denoiser calls	Reward calls	Denoiser backward
Pretrained (MDLM) (Sahoo et al., 2024)	1	0	✗
DRAKES (Wang et al., 2025)	1	0	✗
Best-of-N (Beirami et al., 2025)	N	0	✗
SMC (Del Moral, 2004)	$3N$	$2N$	✗
SVDD (Li et al., 2025)	$N{+}1$	N	✗
GILC-DB (Dou et al., 2026)	1	n	✗
GILC-PG (Dou et al., 2026)	1	n	✗
TDS (Wu et al., 2023)	$4N$	$3N$	✓
SMC-DDM (Ou et al., 2026)	N	Nn	✓
TreeG-G (Guo et al., 2025)	N	Nn	✓
VGAS-GR (ours)	N	$N(n{+}1)$	✗
VGAS-RO (ours)	N	$N(n{+}1)$	✗

Results are means over three seeds (640 generated sequences per seed on DNA). The reward oracles for guidance and the held-out oracles for evaluation follow Wang et al. (2025) on DNA and protein, and Lin et al. (2025) on QM9.

TreeG releases no official protein-design pipeline, so on that benchmark the combination comparison is against SMC-DDM only.

The protein benchmark is the Megascale corpus of Tsuboyama et al. (2023) as curated by Wang et al. (2025), whose wild-type structures are clustered with Foldseek and split by cluster, leaving a test set of 12 proteins. On QM9 the generator is the multimodal flow model of Lin et al. (2025), which adapts Multiflow (Campbell et al., 2024) to small molecules with the equivariant architecture of Hoogeboom et al. (2022). A molecule there is a three-dimensional point cloud of heavy atoms carrying continuous coordinates and discrete atom types. The coordinates are guided as in Lin et al. (2025) and our logit correction is added on the atom-type logits.

DG. DG (Nisonoff et al., 2025) is run at its official default configuration like every other baseline. On DNA it behaves as a guidance method: the reward rises from 0.18 to 1.10 while every fidelity metric falls (Table 1). On protein neither axis moves beyond one standard deviation of the pretrained model (Table 2).

SMC-DDM and TreeG-G. We run SMC-DDM (Ou et al., 2026) and TreeG's gradient variant TreeG-G (Guo et al., 2025) from their official repositories on the shared backbone, matching the denoiser-call budget of the other search methods and otherwise following the setup above. TreeG-G releases no official protein-design pipeline, so we omit it from the protein comparison in Table 2. No schedule shape is imposed on top of $\alpha_t = \alpha\sigma_t$: the selector's only free parameter is the scalar α, and the same value is used on all three domains. Table 6 reports the per-step inference cost of every method underlying Tables 1–2: calls to the denoising network and to the reward oracle, per step and per generated sequence, with a third column for whether the method backpropagates through

the *denoiser*, which the call counts do not capture. There n is each method's own Monte Carlo sample count per guidance estimate. Best-of-N scores only completed sequences, so its reward calls fall outside the per-step count, and the extra reward call in our rows is the selector's value $V(\boldsymbol{z}_s)$, the previous step's being cached. We report this budget rather than claim per-call optimality. The budgets are not equal across rows. On DNA, VGAS-GR and SMC-DDM both run at n=10 while VGAS-RO runs at n=20, so VGAS-RO issues about twice the reward calls of either; on QM9 the two paths run at n=5 and n=20, a factor of three and a half. Where the non-differentiable path leads, it leads at the larger budget.

C METRIC DEFINITIONS

This appendix gives the precise definition of every reward and fidelity metric used in §4 and in the full ablation of §D. Throughout, "reward" metrics measure how strongly generation is pushed toward the guidance objective, while "fidelity" metrics measure whether the resulting designs stay realistic rather than becoming adversarial artifacts that merely fool the reward oracle.

DNA enhancer design.

- *Pred-med*, the median, over generated sequences, of the predicted HepG2 enhancer activity from the reward oracle. Higher is better.
- *ATAC%*, the fraction of designs judged accessible by a held-out chromatin-accessibility (ATAC) oracle that is never used for guidance. Because it is independent of the guidance reward, a rise in ATAC% together with reward indicates genuine generalization rather than reward hacking.
- *3-mer, JASPAR*, Pearson correlations between the designs and natural enhancers, in 3-mer nucleotide frequency and in JASPAR (Castro-Mondragon et al., 2022) transcription-factor motif occurrence, respectively; both measure how closely the sequence statistics match natural DNA.
- *App-LL*, the approximate log-likelihood of the designs under the frozen pretrained generator, a measure of how natural the samples remain to the base model (less negative is better). It is the evidence lower bound of the masked diffusion model of Sahoo et al. (2024), computed with the released implementation of Wang et al. (2025): the bound is discretized over 128 timesteps with a fresh masking pattern drawn at each, and a single Monte Carlo replicate. Following Wang et al. (2025) and Dou et al. (2026), it is reported on DNA only.

Small-molecule property targeting (QM9).

- *MAE*, the mean absolute error between the property value requested as the condition and the value realized by the generated molecule, evaluated with the property predictor of Lin et al. (2025); reported per property, lower is better. Heat capacity C_v is in $\mathrm{cal\,mol^{-1}K^{-1}}$, the isotropic polarizability α in Bohr^3, the dipole moment μ in Debye, and the HOMO–LUMO gap and the two orbital energies in meV.
- *Validity*, the fraction of generated graphs that are chemically valid. It is a fidelity constraint rather than a reported column: every method is compared at an operating point with validity above 75%, so that MAE gains cannot come from emitting invalid molecules. Table 9 gives the realized values.

Protein sequence design.

- *Pred*-ddG, the median, over generated sequences, of the predicted folding stability from the Megascale-trained oracle (Tsuboyama et al., 2023); higher is more stable.
- *%(*ddG$>$0*)*, the fraction of stabilizing designs (positive predicted ddG).
- *scRMSD*, self-consistency RMSD: each generated sequence is re-folded with ESMFold (Lin et al., 2023) and its backbone compared to the wild-type backbone, as in Wang et al. (2025); lower is better.
- *%(scRMSD*$<$2*)*, the fraction of designs that re-fold well (scRMSD < 2 Å).
- *Success Rate*, the joint structure-and-stability success rate: the fraction of designs that are simultaneously stable (ddG$>$0) and well-folded (scRMSD < 2 Å).

Diversity. Three further metrics measure how concentrated a generated set is, and are used in Appendix E.4.

- *mean-Ham*, the mean pairwise Hamming distance over all pairs of generated sequences, normalized by sequence length. It is the diversity metric reported by the benchmark of Wang et al. (2025).
- *NN*, the mean nearest-neighbour distance: for each generated sequence, the normalized Hamming distance to its closest other generated sequence, averaged over the set. Unlike mean-Ham it is sensitive to duplicates and near-duplicates.
- *uniq*, the uniqueness rate: the fraction of generated sequences that are distinct.

D Full-Result Ablation Study

The main-text ablation (Table 4a) reports one metric per stage. Here we give the complete ablation of our method: every stage of the build-up for both guided proposals, over all metrics. Shaded rows are our two final settings.

Two details are referenced from §5. *(i) The selector's share is largest where the proposal is weakest:* on protein the AMC increment exceeds the SMC increment on both paths, and by the wider margin on the non-differentiable one. *(ii) Recovery.* With recovery defined as $1-(\text{final gap})/(\text{proposal gap})$, where the proposal gap is the difference between the variance-reduced differentiable proposal and the plain policy-gradient one, and the final gap is the same difference at the +AMC rung of each ladder, selection closes 51.4% of the gap on DNA and 54.9% on protein, matching despite the non-differentiable proposal being $4.4\times$ weaker on protein.

Table 7: **Full DNA ablation of VGAS** (mean±std, 3 seeds, $N=20$). Every stage of the build-up for both guided proposals, all metrics; shaded are our two finals.

Method	Pred-med ↑	ATAC (%) ↑	3-mer ↑	JASPAR ↑	App-LL ↑
GILC-DB proposal	$6.214_{\pm 0.046}$	$83.80_{\pm 1.66}$	$0.788_{\pm 0.024}$	$0.898_{\pm 0.002}$	$-279.1_{\pm 0.5}$
+ Gumbel–Rao	$6.345_{\pm 0.008}$	$87.33_{\pm 1.37}$	$0.797_{\pm 0.011}$	$0.887_{\pm 0.002}$	$-278.8_{\pm 0.3}$
+ SMC	$6.845_{\pm 0.149}$	$97.87_{\pm 0.59}$	$0.871_{\pm 0.023}$	$0.892_{\pm 0.010}$	$-279.3_{\pm 0.9}$
+ AMC (VGAS-GR)	$7.332_{\pm 0.177}$	$99.10_{\pm 1.56}$	$0.875_{\pm 0.025}$	$0.882_{\pm 0.010}$	$-276.3_{\pm 0.7}$
GILC-PG proposal	$4.859_{\pm 0.074}$	$47.20_{\pm 0.70}$	$0.373_{\pm 0.020}$	$0.865_{\pm 0.004}$	$-276.3_{\pm 0.4}$
+ SMC	$6.326_{\pm 0.096}$	$86.00_{\pm 1.04}$	$0.908_{\pm 0.055}$	$0.936_{\pm 0.008}$	$-275.8_{\pm 1.0}$
+ AMC	$6.610_{\pm 0.037}$	$89.50_{\pm 3.84}$	$0.926_{\pm 0.032}$	$0.900_{\pm 0.009}$	$-274.7_{\pm 0.2}$
+ RO (VGAS-RO)	$6.956_{\pm 0.188}$	$93.93_{\pm 2.44}$	$0.943_{\pm 0.006}$	$0.900_{\pm 0.009}$	$-275.0_{\pm 0.3}$

Table 8: **Full protein ablation of VGAS** (mean±std, 3 seeds, $N=20$). Every stage of the build-up for both guided proposals, all metrics; shaded are our two finals. −: metric not run for that row.

Method	Pred-ddG ↑	%(ddG>0) ↑	scRMSD ↓	%(scRMSD<2 Å) ↑	Success Rate (%) ↑
GILC-DB proposal	$0.644_{\pm 0.018}$	$77.00_{\pm 1.29}$	$0.893_{\pm 0.002}$	$89.17_{\pm 0.36}$	$67.51_{\pm 1.35}$
+ Gumbel–Rao	$0.703_{\pm 0.022}$	$79.77_{\pm 1.18}$	$0.902_{\pm 0.012}$	$89.30_{\pm 0.26}$	$70.05_{\pm 1.41}$
+ SMC	$0.846_{\pm 0.108}$	$83.36_{\pm 2.23}$	$0.901_{\pm 0.024}$	$90.36_{\pm 1.18}$	$74.50_{\pm 1.33}$
+ AMC (VGAS-GR)	$1.285_{\pm 0.163}$	$90.23_{\pm 4.30}$	$0.881_{\pm 0.035}$	$91.64_{\pm 0.87}$	$82.07_{\pm 3.19}$
GILC-PG proposal	$0.160_{\pm 0.022}$	$54.34_{\pm 0.38}$	$0.853_{\pm 0.012}$	$90.80_{\pm 0.54}$	$49.65_{\pm 0.63}$
+ SMC	$0.405_{\pm 0.175}$	$58.77_{\pm 3.68}$	$0.869_{\pm 0.010}$	$90.13_{\pm 0.75}$	$53.10_{\pm 3.08}$
+ AMC	$1.040_{\pm 0.056}$	$72.14_{\pm 2.13}$	$0.857_{\pm 0.012}$	$92.17_{\pm 2.45}$	$67.43_{\pm 2.13}$
+ RO (VGAS-RO)	$1.116_{\pm 0.088}$	$76.04_{\pm 1.04}$	$0.852_{\pm 0.016}$	$92.30_{\pm 1.74}$	$69.73_{\pm 0.34}$

Every method clears the floor, so the errors of Table 3 are read at comparable validity. VGAS is not the most valid of them: on the gap and the two orbital energies SMC-DDM and TreeG-G stay above 95% where VGAS sits between 83% and 88%, and their error on those same properties is the larger by a wide margin. The guidance that lowers the error does cost validity relative to leaving the kernel alone, and the floor is what keeps that cost bounded.

Table 9: **Realized chemical validity on QM9** (%, mean±std, 3 seeds), one column per targeted property. The pretrained model generates unconditionally, so its value does not vary with the target. Every method clears the 75% operating point of Lin et al. (2025).

Method	C_v	α	μ	gap	HOMO	LUMO
Pretrained (MDLM) (Sahoo et al., 2024)	$88.61_{\pm 1.37}$	$88.61_{\pm 1.37}$	$88.61_{\pm 1.37}$	$88.61_{\pm 1.37}$	$88.61_{\pm 1.37}$	$88.61_{\pm 1.37}$
SMC (Del Moral, 2004)	$88.90_{\pm 1.97}$	$90.53_{\pm 1.95}$	$84.24_{\pm 0.75}$	$86.98_{\pm 0.85}$	$87.21_{\pm 0.68}$	$87.34_{\pm 0.65}$
SVDD (Li et al., 2025)	$88.93_{\pm 1.05}$	$88.67_{\pm 1.06}$	$85.32_{\pm 1.13}$	$87.27_{\pm 1.03}$	$87.24_{\pm 0.94}$	$87.27_{\pm 0.76}$
GILC-DB (Dou et al., 2026)	$75.12_{\pm 0.41}$	$79.59_{\pm 1.70}$	$75.46_{\pm 1.77}$	$87.76_{\pm 1.09}$	$88.51_{\pm 0.84}$	$89.06_{\pm 0.80}$
GILC-PG (Dou et al., 2026)	$75.00_{\pm 0.98}$	$78.87_{\pm 1.96}$	$76.81_{\pm 1.96}$	$86.59_{\pm 0.65}$	$87.92_{\pm 0.73}$	$85.12_{\pm 0.66}$
SMC-DDM (Ou et al., 2026)	$88.41_{\pm 1.10}$	$92.06_{\pm 0.30}$	$76.66_{\pm 0.71}$	$96.22_{\pm 0.41}$	$95.83_{\pm 0.65}$	$95.96_{\pm 0.54}$
TreeG-G (Guo et al., 2025)	$88.83_{\pm 1.63}$	$92.12_{\pm 0.25}$	$75.84_{\pm 1.32}$	$95.73_{\pm 0.62}$	$95.35_{\pm 0.65}$	$95.71_{\pm 0.26}$
VGAS-GR	$78.17_{\pm 1.09}$	$79.33_{\pm 2.50}$	$78.23_{\pm 0.48}$	$87.21_{\pm 0.20}$	$88.05_{\pm 1.39}$	$85.29_{\pm 1.33}$
VGAS-RO	$77.18_{\pm 1.19}$	$78.22_{\pm 0.26}$	$78.33_{\pm 0.83}$	$82.98_{\pm 1.07}$	$87.53_{\pm 2.69}$	$83.37_{\pm 0.83}$

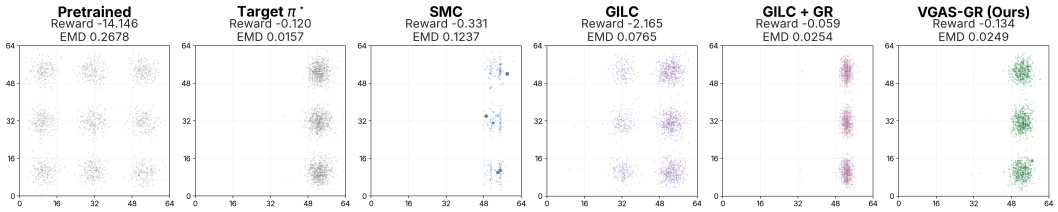


Figure 4: **Steering on a synthetic grid.** Samples against the tilted target $p^\star$, with reward and earth-mover distance to $p^\star$ above each panel.

E Additional Analysis

Synthetic verification. We first reproduce the 64×64 grid of Ou et al. (2026), on which the reward-tilted target is available in closed form, and add our components one at a time (Figure 4): guidance alone recovers the target region but leaves mass on off-target modes, the Gumbel–Rao proposal sharpens it, and the adaptive-temperature selector brings the samples closest to the target, whereas SMC matches the reward only by collapsing onto a few points.

This appendix reports the ablation-study details, the shape of the effective-temperature schedule and a check of reward-scale invariance, the isolation of the two matched variance reducers, and the reward–diversity frontier.

E.1 Reward-Scale Invariance

The per-step spread σ_t that AMC divides by drifts by $3.7\times$ to $27.7\times$ along the trajectory (Figure 2a, Appendix E.5). One further check corroborates the mechanism. Because standardization divides by σ_t, AMC is invariant to a reward rescaling $r \mapsto ar$ (the shift b cancels in the softmax by Lemma 2). Table 10 confirms the scale part empirically: a ninefold change in reward scale moves Pred-med by 0.078 under AMC versus 0.471 under SMC, so α needs no retuning when the reward is rescaled.

Table 10: **Reward-scale invariance.** DNA Pred-med under a reward rescaling $r \mapsto ar$. AMC barely moves ($|\Delta|$=0.078) whereas SMC shifts by 0.471, so the base temperature α needs no retuning when the reward is rescaled.

	$r\times\frac{1}{3}$	$r\times3$
AMC	7.268	7.346
SMC	6.678	7.149
$\lvert\Delta\rvert$	0.078 vs 0.471	

E.2 Isolating the Matched Variance Reducers

RLOO placement (non-differentiable path). Table 11 varies the placement of the leave-one-out baseline over the two-by-two of proposal and selector. On the proposal, at the group-mean selector, it gives a clear gain ($6.610 \to 6.956$, $+0.346$, t=3.13). On the selector it moves the result by $+0.211$ in one row and -0.296 in the other, of comparable size and opposite sign, so the table identifies no

selector effect. This is the shape the leave-one-out case of Lemma 2 (restated in Appendix A) leads one to expect, the selector baseline being inert up to $O(1/(N-1))$ in the log-weights, although the lemma's constant is in log-weight rather than reward units and we do not convert it. We therefore place RLOO on the proposal alone.

Gumbel–Rao sample count (differentiable path). On the proposal the estimator improves over the base row on both domains, by 0.161 on DNA and 0.064 on protein against seed spreads of 0.008–0.046 and 0.010–0.022, and is saturated by $M{=}8$. Under AMC the five rows span less than the spread of any one of them, so the sweep resolves no effect of M on the final method (Table 12); we therefore set $M{=}8$ from the proposal and use it on both sequence benchmarks.

Table 11: **Placement of the RLOO baseline** (DNA, PG path). DNA Pred-med with the baseline on the proposal, the selector, both, or neither. The proposal placement carries a consistent gain; the selector placement moves the result in opposite directions in the two rows.

(DNA, PG)	sel = group mean	sel = leave-one-out
prop = group mean	6.610	6.821
prop = leave-one-out	**6.956**	6.660

Table 12: **Gumbel–Rao sample count** M (backward-only; NFE fixed; mean±std, 3 seeds). Pred-med (DNA) and ddG (protein), proposal alone and under AMC. On the proposal the gain over the base estimator is resolved on both domains and saturates; under AMC the columns lie within the seed spread, so M is set from the proposal and carried over.

M	DNA prop	DNA +AMC	Protein prop	Protein +AMC
base	$6.214_{\pm 0.046}$	$7.227_{\pm 0.108}$	$0.644_{\pm 0.018}$	$1.243_{\pm 0.070}$
2	$6.328_{\pm 0.016}$	$7.265_{\pm 0.062}$	$0.664_{\pm 0.017}$	$1.235_{\pm 0.050}$
4	$6.335_{\pm 0.016}$	$7.132_{\pm 0.165}$	$0.684_{\pm 0.020}$	$1.264_{\pm 0.128}$
8	$6.345_{\pm 0.008}$	$7.332_{\pm 0.177}$	$0.703_{\pm 0.022}$	$1.285_{\pm 0.163}$
16	$\mathbf{6.375}_{\pm 0.027}$	$7.224_{\pm 0.028}$	$\mathbf{0.708}_{\pm 0.010}$	$1.278_{\pm 0.048}$

E.3 Protein Structure versus Stability

On protein the adaptive selector lifts stability while holding structure. VGAS-GR attains the best training-free ddG and surpasses the trained DRAKES on it (Table 2), and also exceeds DRAKES on the joint structure-and-stability success rate (82.1% vs. 78.6%). In the protein ladder (Table 8), the selector raises ddG from 0.85 to 1.29 while the joint success rate rises ($74.5\% \to 82.1\%$) and the well-folded rate holds ($90.4\% \to 91.6\%$); on DNA the same pattern holds (ATAC and 3-mer both improve). VGAS therefore attains the best training-free stability at no cost in structural success.

E.4 Diversity Metric Blindness

Table 13 is the diversity evidence for §5: the standard mean-Hamming metric stays flat once selection is switched on, while the nearest-neighbour distance over the same runs falls by more than two orders of magnitude, under either selector. Mean-Hamming therefore does not detect the concentration that the frontier of Table 14 then quantifies.

Table 13: **Metric blindness.** DNA mean-Hamming distance is unchanged by selection, although the nearest-neighbour (NN) distance over the same runs collapses.

	mean-Ham ↑	NN ↑
Guided proposal (no selection)	0.73	0.62
SMC	0.72	0.0044
AMC	0.72	0.0043

E.5 What a Fixed Temperature Costs

Protocol. At every denoising step we record the whole array $\{\Delta_t^k\}_{k=1}^N$ as it enters the resampling softmax, which needs no additional denoiser call. We run four selectors on each benchmark, AMC and the fixed temperatures $\alpha \in \{0.1, 0.5, 2.0\}$, at three seeds. The three benchmarks parameterize

Table 14: **Reward–diversity frontier at matched resampling frequency** (DNA, 3 seeds; the standard deviation is given for the reward column only). AMC against SMC on the same guided proposal; lower r-freq resamples more often. AMC attains the higher reward at every r-freq.

r-freq	selector	Pred-med ↑	uniq ↑	NN ↑
1	AMC	$\mathbf{7.332}_{\pm 0.177}$	0.492	0.0043
	SMC	$6.845_{\pm 0.149}$	0.489	0.0044
2	AMC	$\mathbf{6.742}_{\pm 0.052}$	0.598	0.0113
	SMC	$6.513_{\pm 0.145}$	0.623	0.0101
4	AMC	$\mathbf{6.582}_{\pm 0.036}$	0.894	0.0274
	SMC	$6.487_{\pm 0.063}$	0.854	0.0276
8	AMC	$\mathbf{6.544}_{\pm 0.069}$	0.990	**0.0575**
	SMC	$6.349_{\pm 0.058}$	**0.999**	0.0558

time differently, so we report everything against denoising progress, from 0 at the all-MASK state to 1 at $\boldsymbol{x}_0$; without that normalization two of the domains drift in opposite directions. Drift is the ratio of the mean of σ_t over the first fifth of the trajectory to its mean over the last fifth, which is insensitive to the noise in any single step. Curves are drawn with a five-step moving average. Table 15 reports the drift and Figure 5 the per-step effective sample size.

Table 15: **Drift of the per-step spread.** Mean σ_t over the first and the last fifth of the trajectory, and their ratio. The main text quotes the fixed-α row, the selector being replaced.

Domain	Selector	first fifth	last fifth	drift
DNA	fixed α=0.5	0.792	0.213	3.7×
	AMC	0.809	0.187	4.3×
Protein	fixed α=0.5	0.156	0.0173	9.0×
	AMC	0.121	0.0091	13.3×
QM9	fixed α=0.5	12.58	0.454	27.7×
	AMC	12.58	0.439	28.6×

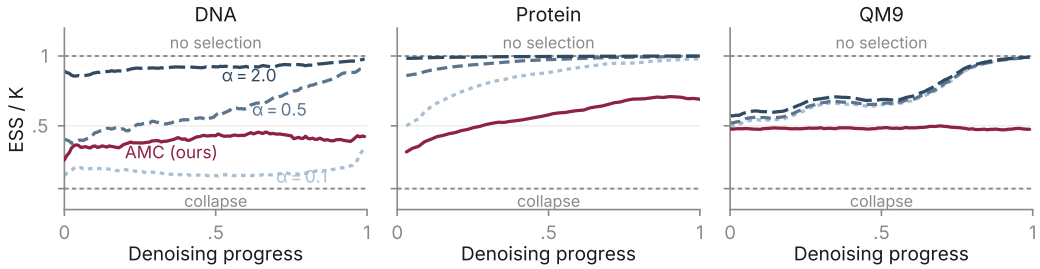


Figure 5: **Per-step effective sample size.** Four selectors on each benchmark. On QM9 the fixed-α curves are recomputed from the AMC increments rather than run separately.

Reading the figure. The effective sample size $\mathrm{ESS} = (\sum_k \hat{w}^k)^2 / \sum_k (\hat{w}^k)^2$ counts the particles that survive the resampling at that step. At N the weights are uniform and selection does nothing; at 1 a single particle carries the population. DNA shows the two failure modes at once. With α=0.1 the effective sample size sits near $0.16N$ for most of the trajectory, so the population is collapsed from the first steps, while with α=2.0 it stays above $0.92N$, so selection barely acts anywhere. Every fixed value rises toward the end. On DNA the two larger ones finish above $0.93N$ and α=0.1 finishes at $0.43N$, on protein all three finish above $0.97N$ and on QM9 above $0.99N$, whereas AMC finishes at $0.37N$, $0.66N$ and $0.47N$. The temperature that is correct at one end of the trajectory is therefore wrong at the other, and no single value is correct throughout.

The shape of the increments moves too. Standardization is not a fixed shape rescaled. The skewness and excess kurtosis of $(\Delta_t^k - \mu_t)/\sigma_t$ drift along the trajectory as well, and on protein and QM9 the excess kurtosis sits near $+4$, far from Gaussian. The selector holds its selection pressure while dividing by σ_t alone, and neither Lemma 2 nor Proposition 2 assumes anything about the shape: the first uses only that the shift cancels, the second only that σ_t admits a deterministic limit.

What the late trajectory means. A fixed α could be read as behaving correctly at the end: once σ_t is small there is little left to choose between, uniform weights would be the right response, and dividing by σ_t would amplify noise. The samples say otherwise. Were the late selection acting on noise, AMC would not beat every fixed α in Figure 3 nor lead Table 1.

On QM9. The QM9 diagnostic runs at γ=100 rather than the γ=10 of Table 5, and its fixed-α curves are recomputed from the AMC increments under the fixed weighting rule rather than run separately, so those curves share one trajectory. The drift is a property of $\{\Delta_t^k\}$ itself and the four selectors agree on it to within 27.6–28.6×, so Table 15 is unaffected.

F Related Work and Positioning

Reading Table 16. Prior work guides, selects, or combines the two. *Training-free* means the generative model is not retrained. The *Gradient* column marks use of the reward's gradient: through the denoiser for the guidance rows, through the sampling trajectory for DRAKES; the gradient-free MCMC methods (SGDD, CSMC, IterRef) and the pure-search methods use none. *Adaptive sel.* marks a selection rule whose temperature is measured from the particle population at run time; an ESS-triggered resampling schedule, which adapts when to resample rather than how sharply to select, is not counted (Appendix F.1). Unlike the combined methods (TDS, TreeG, SMC-DDM), our selection uses an adaptive rather than a pre-set rule, TreeG's hard top-A beam or the SMC family's pre-set temperature. Four entries carry a qualification. †DG does not retrain the generator but fits a noise-conditional predictor, and we tabulate its gradient-based DG-TAG variant. ‡TFG-Flow uses the gradient on the continuous part only; its discrete part is guided by importance sampling. §SMC-DDM is our shorthand for Ou et al. (2026), after their earlier preprint title. ¶TDS is formulated for continuous diffusion, and the discrete rows report the adaptation obtained by post-hoc kernel tilting (Eq. (63)).

Table 16: **Positioning of reward-steering methods.** Columns and markers are defined above.

Method	Training-free	Gradient	Search	Discrete	Combination	Adaptive sel.
Fine-tuning / learning-based						
DRAKES (Wang et al., 2025)	✗	✓	✗	✓	✗	✗
SDPO (Han et al., 2026)	✗	✗	✗	✓	✗	✗
DDPP (Rector-Brooks et al., 2025)	✗	✗	✗	✓	✗	✗
Gradient guidance						
DPS (Chung et al., 2023)	✓	✓	✗	✗	✗	✗
DG-TAG (Nisonoff et al., 2025)	✓†	✓	✗	✓	✗	✗
TFG-Flow (Lin et al., 2025)	✓	✓‡	✗	✓	✗	✗
EntRGi (Tejaswi et al., 2026)	✓	✓	✗	✓	✗	✗
GILC (Dou et al., 2026)	✓	✓	✗	✓	✗	✗
Sampling-based guidance (gradient-free MCMC)						
SGDD (Chu et al., 2025)	✓	✗	✗	✓	✗	✗
CSMC (Phunyaphibarn & Sung, 2026)	✓	✗	✗	✓	✗	✗
IterRef (Lee et al., 2025)	✓	✗	✗	✓	✗	✗
Search / selection						
Best-of-N (Beirami et al., 2025)	✓	✗	✓	✓	✗	✗
SMC (Del Moral, 2004; Chopin & Papaspiliopoulos, 2020)	✓	✗	✓	✓	✗	✗
SVDD (Li et al., 2025)	✓	✗	✓	✓	✗	✗
Guidance + search						
TDS (Wu et al., 2023)¶	✓	✓	✓	✗	✓	✗
TreeG (Guo et al., 2025)	✓	✓	✓	✓	✓	✗
SMC-DDM§ (Ou et al., 2026)	✓	✓	✓	✓	✓	✗
Ours	✓	✓	✓	✓	✓	✓

Gradient guidance. The families differ in where the reward gradient is applied: at the denoiser input (Chung et al., 2023), in the hidden states of the denoiser, with a discriminative head fitted for the purpose (Gruver et al., 2023), on the rate matrix (Nisonoff et al., 2025; Lin et al., 2025), or on the clean-token logits (Tejaswi et al., 2026; Dou et al., 2026). We adopt GILC's placement, which bypasses the model Jacobian, stated here as approximating it by the identity (§F.3), and equip it with Gumbel–Rao. A separate, gradient-free line steers by MCMC over clean or intermediate states (Chu et al., 2025; Phunyaphibarn & Sung, 2026; Lee et al., 2025), admitting the reward through accept/reject rather than a gradient. Against GILC specifically, the difference is where the standardization acts: GILC-PG already standardizes by a group standard deviation, but on the *proposal*, across the Monte Carlo samples drawn at one state, whereas ours is on the *selector*, across particles and per step, and it is the drift of that spread that §3.3 acts on.

Inference-time search. The families differ in the selection rule. SMC and the Twisted Diffusion Sampler (Wu et al., 2023) resample a particle population by a value twist, value-based decoding (Li et al., 2025) selects per step, TreeG (Guo et al., 2025) unifies candidate proposal, value evaluation and selection under a tree search, and a general Feynman–Kac framework covers the family across modalities, discrete text diffusion among them (Singhal et al., 2025). TreeG is the closest prior work: its gradient-based variant TreeG-G establishes the gradient-plus-search combination for discrete diffusion, and we run it from the official repository. What separates us from all of them is the selection rule itself. The prior combinations select by a rule *set in advance*, hard top-A for TreeG's beam search and a prescribed resampling temperature for the SMC family, whereas we measure the temperature from the particles; our contribution is that rule, not a new proposal.

Learning-based steering. These differ from us on the training-free axis rather than on the gradient or the search axis, amortizing the reward-tilted posterior into a trained sampler (Rector-Brooks et al., 2025) or retraining the generator (Wang et al., 2025; Han et al., 2026). We compare against DRAKES, the standard fine-tuned reference on these benchmarks. The value-function view of Diamond Maps (Holderrieth et al., 2026) connects these directions through the soft value.

Concurrent training-free SMC steerers. These share our stance but not our selection rule, steering through pre-set potentials (Hasan et al., 2026), trajectory-level resampling (Dang et al., 2026), or a corrected weighting of prior bootstrap formulations (Yadala Chanchu et al., 2026). None of them measures the temperature from the particles, which is the σ_t-standardized selection of §3.3 (Appendix F.1). Orthogonally, CDM (Kim et al., 2026) *learns* an amortized twist, the same axis on which we position against SMC-DDM's amortized proposal.

F.1 Relation to Adaptive-Tempering SMC

Temperature adaptation within SMC is well established: when a sampler anneals to a static target through a ladder $\pi_t \propto \pi_0^{1-\lambda_t}\pi_{\text{target}}^{\lambda_t}$, the exponents are commonly chosen *adaptively* to hold a target effective sample size (Beskos et al., 2016). AMC differs on three axes. *(i) Object.* We adapt the per-step *resampling* temperature inside a reverse process we do not design. *(ii) Statistic.* Adaptive tempering solves iteratively for the exponent that hits a user-chosen ESS level; AMC uses a closed-form, parameter-free statistic, the per-step spread σ_t of the group advantage. *(iii) Motivation.* Our diagnosis is specific to reward-guided diffusion: σ_t drifts by $3.7\times$ to $27.7\times$ along the trajectory (Figure 2a), a phenomenon absent from static-target annealing. Table 17 makes the distinction concrete. The ESS-adaptive tolerance ladder of Del Moral et al. (2012) is in the same spirit; we cite Beskos et al. (2016) for the convergence of AMC (Proposition 2).

Table 17: AMC **versus classical adaptive-tempering SMC.**

	Adaptive-tempering SMC	AMC (ours)
Adapted quantity	annealing exponent λ_t, on a static target	resampling temperature, inside the diffusion reverse process
Selection rule	solve λ_t for a target ESS, iteratively	$\alpha_{\text{eff}} = \alpha\sigma_t$ from the advantage spread, closed-form
Tuning	a target ESS level, retuned per setting	one scalar α, the same value on all three domains
Motivation	sample-health control	σ_t drifts 3.7–27.7$\times$ along the trajectory

F.2 Anatomy of Guidance and Selection in Prior Samplers

Table 18 resolves what Table 16 leaves open: which gradient and which search. The closest baselines are often described as sharing one guidance and differing only in selection, and neither half of that description holds. The guidance signals are distinct objects, and all three baselines differentiate through the denoiser at one backward pass per step, whereas $\partial\boldsymbol{\eta}/\partial\boldsymbol{z}_t \approx \boldsymbol{I}$ removes the denoiser from ours. The selection rules differ too, and TreeG departs furthest, keeping the top-A candidates by value with no importance weight, a beam search rather than a particle filter. The two substantive differences, which derivative is taken and where the tilt is inserted, are made precise in §F.3.

Table 18: **Guidance versus selection in the closest samplers.** Every gradient signal is written in the common form $\nabla_{(\cdot)}\hat{r}$ ($\hat{r} = \frac{1}{n}\sum_i r(\hat{\boldsymbol{x}}_0^{(i)})$); methods differ only in the variable of differentiation ($\boldsymbol{x}_t$, the noisy state $\boldsymbol{z}_t$, or the clean logits $\boldsymbol{\eta}$) and in where the tilt is inserted. Ours is the only Jacobian-free entry and the only one with an adaptive resampling temperature. The backward row counts passes through the denoiser only; every differentiable-reward method, ours included, still backpropagates through the reward oracle.

Axis	**TDS** (Wu et al., 2023)	**SMC-DDM** (Ou et al., 2026)	**TreeG-G** (Guo et al., 2025)	**Ours** (§3.1–3.3)
State space	continuous (Gaussian)	discrete (masked)	discrete (flow / masked)	discrete (masked)
Guidance signal	$\nabla_{\boldsymbol{x}_t}\hat{r}$ (continuous score at $\hat{\boldsymbol{x}}_0$)	$\nabla_{\boldsymbol{z}_t}\hat{r}$ (ST-Gumbel relaxation)	$\nabla_{\boldsymbol{z}_t}\hat{r}$ (ST-Gumbel relaxation)	$\nabla_{\boldsymbol{\eta}}\hat{r}$ (clean logits; $\partial\boldsymbol{\eta}/\partial\boldsymbol{z}_t \approx \boldsymbol{I}$)
Model Jacobian	✓	✓	✓	✗
Denoiser backward passes / step	1	1	1	0
Reveal factorized over positions	✓	✓	✓	✓
Selection	SMC (systematic), pre-set twist	SMC (multinomial, every step), pre-set λ_t schedule	top-A by value (their default; a value-proportional option exists), no temp.	SMC (multinomial), **adaptive temp.** (AMC)
Proposal correction $\log(p_\theta/q)$	✓ (post-hoc e^g; drops Z)	✓	✗	✗ by design (Lemma 1)

F.3 Explicit Comparison of the Guidance Signals: TDS, SMC-DDM, TreeG-G, and Ours

Here we state the guidance signals as formulas, so that the two substantive differences, which derivative is taken and where the tilt is inserted, can be verified line by line. Throughout, $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{z}_t)$ are the clean logits emitted by the denoiser at $\boldsymbol{z}_t$, $p_\theta^\ell = \mathrm{softmax}(\boldsymbol{\eta}^\ell)$ is the per-position clean prediction, a_t is the schedule's reveal probability from Eq. (2), and we work at a single masked position ℓ.

(i) Which derivative. All three gradient methods target the sensitivity of the multi-sample reward $\hat{r}(\boldsymbol{z}_t) = \frac{1}{n}\sum_i r(\boldsymbol{x}_0^{(i)})$, $\boldsymbol{x}_0^{(i)} \sim p_\theta(\cdot \mid \boldsymbol{z}_t)$, routing the categorical draw through the same Gumbel–Softmax relaxation, and differ only in where the chain rule is stopped. The decomposition is GILC's (Dou et al., 2026, §3.3), which names the last factor the model Jacobian; SMC-DDM keeps it, tilting the kernel from the state (Ou et al., 2026, Eq. 8 and App. C.2),

$$\nabla_{\boldsymbol{z}_t}\hat{r} \;\approx\; \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\partial r(\boldsymbol{x}_0^{(i)})}{\partial \boldsymbol{x}_0^{(i)}}}_{\text{reward backward}} \underbrace{\frac{\partial \boldsymbol{x}_0^{(i)}}{\partial \boldsymbol{\eta}}}_{\text{ST-Gumbel}} \underbrace{\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{z}_t}}_{\textbf{denoiser backward}}, \tag{61}$$

and TreeG-G (Guo et al., 2025, Modules 3–4) forms exactly the same three-factor product, then contracts it against $(\boldsymbol{z}_t^{\setminus\ell}(k) - \boldsymbol{z}_t)$ to obtain a per-token score $g_k^{(\ell)} = (\boldsymbol{z}_t^{\setminus\ell}(k) - \boldsymbol{z}_t)^\top \nabla_{\boldsymbol{z}_t}\hat{r}$. Ours keeps the first two factors and replaces the third by the identity,

$$\boldsymbol{g}^\ell = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial r(\boldsymbol{x}_0^{(i)})}{\partial \boldsymbol{x}_0^{(i)}} \frac{\partial \boldsymbol{x}_0^{(i)}}{\partial \boldsymbol{\eta}^\ell} \;=\; \Big[\nabla_{\boldsymbol{z}_t}\hat{r}\Big]_{\partial\boldsymbol{\eta}/\partial\boldsymbol{z}_t \to \boldsymbol{I}}, \tag{62}$$

which is Eq. (6). Ours is therefore Eq. (61) truncated after two factors, the "Jacobian-free" row of Table 18, and two things follow from the truncation.

First, the discarded factor is the only one requiring a backward pass through the denoiser. Second, the surviving product lives in logit space, $\boldsymbol{g}^\ell \in \mathbb{R}^{|\mathcal{V}|}$, whereas Eq. (61) lives in state space. Since the reveal is parameterized by $\boldsymbol{\eta}$, a logit-space correction can be added to $\boldsymbol{\eta}$ and reused verbatim, while a state-space one must be applied externally to the transition kernel.

The truncation is an approximation, exact only for an identity denoiser, and it changes the quantity estimated: it gives the direction in which the clean logits should move rather than the direction in which the noisy state should move, and the former is the one the reveal can use as it stands. Its cost in reward and fidelity is measured in §4.

(ii) Where the tilt is inserted, and its effect on the schedule. At a masked position the base kernel Eq. (2) puts mass $a_t\, p_\theta^\ell(k)$ on each token $k \in \mathcal{V}$ and mass $1 - a_t$ on staying masked. Every method tilts this by $e^{\gamma g_k}$ and they differ only in whether the MASK branch is tilted with it. Write $Z^\ell = \mathbb{E}_{p_\theta^\ell}[e^{\gamma g}]$. The post-hoc form of TDS (Wu et al., 2023) and SMC-DDM (Ou et al., 2026, Eq. 8) multiplies the tilt onto the kernel after the reveal probability is fixed, so it renormalizes over $\mathcal{V} \cup \{\text{MASK}\}$:

$$q^\ell_{\text{post}}(k) = \frac{a_t\, p_\theta^\ell(k)\, e^{\gamma g_k}}{1 - a_t + a_t Z^\ell}, \qquad q^\ell_{\text{post}}(\text{MASK}) = \frac{1 - a_t}{1 - a_t + a_t Z^\ell}. \tag{63}$$

Ours Eq. (6) inserts the tilt inside the softmax that defines the clean prediction and leaves the MASK branch untouched:

$$q^\ell(k) = a_t\, \text{softmax}(\boldsymbol{\eta}^\ell + \gamma \boldsymbol{g}^\ell)_k = \frac{a_t\, p_\theta^\ell(k)\, e^{\gamma g_k}}{Z^\ell}, \qquad q^\ell(\text{MASK}) = 1 - a_t, \tag{64}$$

where the second equality is the elementary identity $\text{softmax}(\boldsymbol{\eta}^\ell + \gamma \boldsymbol{g}^\ell)_k = p_\theta^\ell(k) e^{\gamma g_k} / Z^\ell$. Comparing Eq. (63) and (64): the numerators are identical. Both are the same exponential tilt of the same clean prediction, so both belong to a single twist family. The entire difference is the denominator, whether the normalizer Z^ℓ is retained on the revealed branch alone, or distributed across the revealed and masked branches together. Proposition 3, stated and proved in Appendix A, turns that difference into a statement about the reveal probability $\tilde{a}_t$ each form actually induces.

Part (a) of that proposition makes $\tilde{a}_t^{\ell,\,\text{post}}$ strictly increasing in Z^ℓ, crossing a_t at $Z^\ell = 1$, so the direction of the leakage is decided by whether Z^ℓ exceeds one. Part (b), that $Z^\ell \geq 1$ under the gauge $\mathbb{E}_{p_\theta^\ell}[g] = 0$, settles that question: the leakage is systematic and one-signed, and the post-hoc form reveals earlier than the schedule intends. This matters because the masked-diffusion ELBO and the $T \to \infty$ consistency of the reverse process are stated for a_t, so replacing it by $\tilde{a}_t > a_t$ commits tokens earlier than intended, exactly when guidance is strongest and the clean prediction least reliable.

Part (c), the invariance of Eq. (64) to $\boldsymbol{g}^\ell \mapsto \boldsymbol{g}^\ell + c\mathbf{1}$, is the sharpest, because that constant is arbitrary: the softmax is unchanged by $\boldsymbol{\eta} \mapsto \boldsymbol{\eta} + c\mathbf{1}$, so a correction written in logit coordinates carries no canonical choice of it, and the gauge shift $\hat{\boldsymbol{g}} = \boldsymbol{g} - c\mathbf{1}$ fixes it only by fiat. Under the post-hoc form that convention moves the unmasking schedule, all the way to "reveal everything" or "reveal nothing"; under Eq. (64) it is invisible.

Both parts describe a discrepancy that can be repaired downstream rather than avoided. TDS and SMC-DDM carry the $\log(p_\theta/q)$ correction that repairs the leakage, but the repair is only as good as the particle count and is unavailable under rules that do not reweight, including TreeG's beam search and vanilla SVDD. Eq. (64) removes the discrepancy at the proposal level instead.

(iii) Is putting the correction in the logits ours alone? Not as an idea, and we do not claim it: the logit-space placement is GILC's (Dou et al., 2026), restated in Eq. (6). The narrower, checkable claim is that among the training-free gradient-plus-search samplers of Table 18, none places the correction there. TDS tilts a Gaussian transition; SMC-DDM tilts $p_\theta(x_{t-1} \mid x_t)$ from outside (Ou et al., 2026, Eq. 8); TreeG-G adds its per-token score to candidate generation and selects by beam search (Guo et al., 2025, Modules 3–4). Proposition 3 is the consequence of that placement, not of the tilt, which is shared. Our own contribution sits elsewhere: $\boldsymbol{g}^\ell$ is a per-position object, so Eq. (64) remains a product over positions, and we act through selection instead (§3.3).

G Sampler and Selector Algorithms

This appendix gives the single-trajectory GUIDED SAMPLER loop referenced in §3.1 (Algorithm 3) and our selector AMC (Algorithm 4); both call the guided reveal of §3.1 in place of the base model's unmasking, with the correction $\boldsymbol{g}$ supplied by either reduced-variance estimator (Algorithms 1–2). For AMC the single line that departs from vanilla SMC (the adaptive-temperature weight) is flagged; resampling is multinomial.

Algorithm 3: GUIDED SAMPLER (single trajectory)

Input: denoiser p_θ; reward r; guidance scale γ; MC size n; steps T

$\boldsymbol{z}_T = \text{all-MASK}$;
for $t = T, \dots, 1$ **do**
  $\boldsymbol{\eta} = \text{denoiser}(\boldsymbol{z}_t)$, $\boldsymbol{g} = \text{GUIDANCE}(\boldsymbol{\eta}, r, n)$ (Alg. 1/2);
  reveal masked ℓ from $(1-a_t)\delta_{\text{MASK}} + a_t \,\text{softmax}(\boldsymbol{\eta}^\ell + \gamma \boldsymbol{g}^\ell)$ by Eq. (6);
  keep revealed frozen; $\boldsymbol{z}_t = \boldsymbol{z}_s$;

Output: $\boldsymbol{z}_0$

Algorithm 4: AMC

Input: particles N; reward r; guidance scale γ; MC size n; selection temperature α; variance floor ϵ; steps T

$\boldsymbol{z}_T^k \leftarrow \text{all-MASK}$; $\boldsymbol{\eta}^k \leftarrow \text{denoiser}(\boldsymbol{z}_T^k)$; $V^k \leftarrow r(\arg\max \boldsymbol{\eta}^k)$, $k = 1, \dots, N$;
for $t = T, \dots, 1$ **do**
  for $k = 1, \dots, N$ **do**
    $\boldsymbol{g} \leftarrow \text{GUIDANCE-GR}(\boldsymbol{\eta}^k, \boldsymbol{z}_t^k, r, n)$; `// GR proposal; η^k cached`
    reveal $\boldsymbol{z}_s^k$ from $(1-a_t)\delta_{\text{MASK}} + a_t \,\text{softmax}(\boldsymbol{\eta}^k + \gamma \boldsymbol{g})$ by Eq. (6);
    $\boldsymbol{\eta}_s^k \leftarrow \text{denoiser}(\boldsymbol{z}_s^k)$; $V_s^k \leftarrow r(\arg\max \boldsymbol{\eta}_s^k)$; $\Delta^k \leftarrow V_s^k - V^k$; `// Feynman--Kac increment`
  $\mu \leftarrow \text{mean}(\{\Delta^k\})$, $\sigma \leftarrow \max\big(\text{std}(\{\Delta^k\}), \epsilon\big)$; $\hat{w}^k \propto \exp\big((\Delta^k - \mu)/(\alpha\sigma)\big)$; resample $\{(\boldsymbol{z}_s^k, \boldsymbol{\eta}_s^k, V_s^k)\} \sim \hat{w}$; `// adaptive temperature; vanilla SMC uses a prescribed α`
  $\boldsymbol{z}_t^k \leftarrow \boldsymbol{z}_s^k$; $\boldsymbol{\eta}^k \leftarrow \boldsymbol{\eta}_s^k$; $V^k \leftarrow V_s^k$; `// one denoiser call per particle per step`

Output: $\boldsymbol{z}_0^1, \dots, \boldsymbol{z}_0^N$

G.1 CLOSED FORM OF THE IMPORTANCE CORRECTION

Lemma 1 states that adding $\log(p_\theta/q_\gamma)$ to the log-weights restores $p^\star$ for every γ. Both kernels leave the masked/unmasked split untouched, the guided reveal renormalizes inside the vocabulary and multiplies by the same schedule factor, so the ratio is supported on the positions revealed at that step. For the product-form proposal, writing $\hat{\boldsymbol{g}} = \boldsymbol{g} - c\mathbf{1}$ for an arbitrary gauge constant c and $z_s^\ell = k$,

$$\log \frac{p_\theta(\boldsymbol{z}_s \mid \boldsymbol{z}_t)}{q_\gamma(\boldsymbol{z}_s \mid \boldsymbol{z}_t)} = \sum_{\ell \,\in\, \text{revealed}} \Big(-\gamma \hat{g}_k^\ell + \text{lse}\big(\boldsymbol{\eta}^\ell + \gamma \hat{\boldsymbol{g}}^\ell\big) - \text{lse}\big(\boldsymbol{\eta}^\ell\big) \Big). \tag{65}$$

Summing Eq. (65) over the trajectory and negating gives the trajectory log-twist $\bar{\Phi}_\gamma$ of Appendix A, whose terminal counterpart is Φ_γ of Eq. (10). Every quantity in Eq. (65) is already materialized by the guided reveal, so the correction costs no additional forward or backward pass. Under AMC it must be added to the log-weights after the standardization of $\{\Delta^k\}$, since it carries an absolute scale that the standardization would otherwise destroy.

H BASE PROPOSAL ESTIMATORS

§3.1 presents the two reduced-variance proposals we run, GUIDANCE-GR and GUIDANCE-RO. This appendix states the two base estimators they reduce: the differentiable GUIDANCE-DB and the non-differentiable GUIDANCE-PG. Both share the mean $\nabla_{\boldsymbol{\eta}} \mathbb{E}[r]$ and are interchangeable in Algorithm 3.

Differentiable base (GUIDANCE-DB). GUIDANCE-DB (Alg. 5), the estimator GILC uses in practice, differentiates the reward through the ST-GS path Eq. (5): draw n Gumbel–Softmax samples, evaluate the reward on the hard one-hot, and average the straight-through gradients. Evaluating its ST-GS Jacobian at a single Gumbel draw per sample makes it high-variance; the Gumbel–Rao reduction of §3.1 replaces that Jacobian by an M-sample conditional average without an extra pass.

Algorithm 5: GUIDANCE-DB (differentiable base estimator)

Input: clean logits $\boldsymbol{\eta}$; state $\boldsymbol{z}_t$; reward r; sample size n; Gumbel temp τ

for $i = 1, \dots, n$ **do**
- draw $\boldsymbol{\zeta}^{(i)} \sim \mathrm{Gumbel}(0, 1)$ i.i.d.;
- $\hat{\boldsymbol{x}}^{(i)}_{\mathrm{soft}} \leftarrow \mathrm{softmax}\big((\boldsymbol{\eta} + \boldsymbol{\zeta}^{(i)})/\tau\big)$;
- $\hat{\boldsymbol{x}}^{(i)} \leftarrow$ ST estimator of $\hat{\boldsymbol{x}}^{(i)}_{\mathrm{soft}}$ by Eq. (5);
- $R_i \leftarrow r(\hat{\boldsymbol{x}}^{(i)})$;

$\boldsymbol{g} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \partial R_i / \partial \boldsymbol{\eta}$; // $= \nabla_{\boldsymbol{\eta}} \mathbb{E}[r]$

Output: logit correction $\boldsymbol{g}$

Algorithm 6: GUIDANCE-PG

Input: clean logits $\boldsymbol{\eta}$; state $\boldsymbol{z}_t$; reward r; sample size n

$\boldsymbol{p} \leftarrow \mathrm{softmax}(\boldsymbol{\eta})$; // proxy clean-token dist. $\hat{\boldsymbol{x}}_\theta$

for $i = 1, \dots, n$ **do**
- draw $\boldsymbol{x}^{(i)} \sim \boldsymbol{p}$; $\hat{\boldsymbol{x}}^{(i)} \leftarrow \mathrm{onehot}(\boldsymbol{x}^{(i)})$; // hard; no relaxation
- $R_i \leftarrow r(\hat{\boldsymbol{x}}^{(i)})$;

$A_i \leftarrow \big(R_i - \mathrm{mean}(\{R_j\})\big) \big/ \mathrm{std}(\{R_j\})$; // group-relative advantage

$\boldsymbol{g} \leftarrow \frac{1}{n} \sum_{i=1}^{n} A_i \dfrac{\partial \langle \log \boldsymbol{p}, \hat{\boldsymbol{x}}^{(i)} \rangle}{\partial \boldsymbol{\eta}}$; // $= \frac{1}{n} \sum_i A_i (\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{p})$; same as GILC

Output: logit correction $\boldsymbol{g}$

Substituting the softmax Jacobian for the derivative of the argmax leaves the ST-GS path biased, and conditioning on $i^\star$ preserves that mean: Gumbel–Rao removes the variance of the draw, not the bias of the relaxation. Two implementation points carry into every estimator here. The conditional perturbations $\boldsymbol{\zeta}^{(m)}$ of Eq. (7) depend on $\boldsymbol{\eta}$ through T and v_k, so they are detached from the autograd graph before the reward is differentiated; leaving them attached backpropagates through the sampler itself and corrupts the Jacobian. And the candidate $\hat{\boldsymbol{x}}$ is assembled from the current state: positions already revealed in $\boldsymbol{z}_t$ are held at their committed tokens and only the masked positions receive a sample, so the absorbing structure of Eq. (2) is preserved. The base estimator is the "base" row of Table 12. The conditional $p(\boldsymbol{\zeta} \mid \arg\max = i^\star)$ used there is drawn by the truncated-Gumbel reparameterization of Eq. (7), where T is the maximum $v_{i^\star}$.

Non-differentiable base (GUIDANCE-PG). When the reward is non-differentiable (a black box, or expensive to backprop through), the correction can instead be formed by the score-function / policy-gradient (Reinforce) estimator. Sampling hard tokens $\boldsymbol{x}^{(i)}$ from the proxy clean distribution $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{\eta})$ and using $\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{x}^{(i)}) = \mathrm{onehot}(\boldsymbol{x}^{(i)}) - \boldsymbol{p}$, the reward-weighted score is an unbiased estimate of $\nabla_{\boldsymbol{\eta}} \mathbb{E}[r]$. Following the group-relative formulation of GILC (Dou et al., 2026), itself after group-relative policy optimization (Shao et al., 2024), we standardize the rewards into a *group-relative advantage* before weighting (GUIDANCE-PG, Alg. 6). This needs no relaxation, no Gumbel temperature and no model Jacobian, at the price of higher variance than the differentiable path. The leave-one-out baseline $A_i = R_i - \frac{1}{n-1} \sum_{j \neq i} R_j$ is the variance-reduction counterpart, for this proposal, of Gumbel–Rao for the differentiable one: Table 11 shows it outperforms the full-group baseline on the proposal, while on the selector it moves the result in opposite directions in the two rows, as the inertness of Remark 2 would lead one to expect, which is why our non-differentiable final method places it on the proposal and standardization only on the selector.