CANDIDATE-SPECIFIC CONTEXT COUPLING FOR DISCRETE DIFFUSION PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimization aligns discrete diffusion models with sequence-level rewards through predictions on partially masked candidates. Existing approaches improve preference objectives and denoising estimates, but uniform masking treats all eligible positions alike. With candidate-specific contexts, the joint sampling of conditional scores also shapes the nonlinear preference objective. We introduce *Candidate-specific Context Coupling* (C^3), which coordinates reveal progress, counts, and orders while preserving each candidate's selected context distribution. Low-confidence-first supplies a one-pass selector. For fixed Plackett–Luce reveal laws, shared Gumbels simultaneously minimize expected Kendall disagreement on common positions. Under score-response conditions, they also minimize a derived preference-discrepancy upper bound at a fixed joint distribution of shared progress and reveal counts. The method supports pairwise and listwise losses and requires one additional confidence evaluation per candidate. Experiments show gains across DNA sequence design, protein inverse folding, and language modeling, including a 17% relative gain over the strongest baseline in protein inverse folding. Our anonymous code is available at `https://anonymous.4open.science/r/C3-C9D7/`.

1 INTRODUCTION

Discrete diffusion models generate sequences through conditional predictions on partially masked inputs (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024; Gat et al., 2024; Nie et al., 2025). Post-training aims to direct these generators toward task-specific goals, such as biological activity or response quality. The feedback used for these tasks evaluates completed sequences as a whole. Preference optimization uses these predictions to learn from rewards on completed candidates (Wallace et al., 2024; Han et al., 2026). Sequence-level rewards specify which outcomes to favor, while the masking rule determines what information the model receives when evaluating them. Each context turns the same sequence-level feedback into a different conditional learning problem. Uniform masking treats all eligible positions alike. This raises two questions: *which tokens should condition each candidate's predictions, and how should these contexts enter a preference comparison?*

Within a candidate, conditioning tokens determine which dependencies the model can use. Finite denoisers learn conditional prediction problems with unequal accuracy (Kim et al., 2025). Masking a position together with those depending on it requires inferring both the local choice and its surrounding relationships. For example, revealing the subject supplies its singular or plural form for verb agreement (Figure 1a). Completed candidates supply these choices during post-training. Revealing a choice fixes its value in the input, so the remaining predictions can be learned conditional on that candidate's decision. We use low completed-input confidence as an empirical rule for selecting these conditioning values. This gives more context exposure to positions with less concentrated completed-input predictions, while keeping the reveal-count distribution unchanged. Each preferred or dispreferred candidate supplies its own tokens, with rewards directing the update.

Across candidates, the selected conditioning positions can differ. Even with fixed candidate score distributions, the expected loss of a nonlinear comparison can depend on how those scores are sampled together. Independent contexts introduce variation in reveal stage, count, and order into their score differences. One candidate may be scored with much of its context visible while another must reconstruct most of its sequence. Their comparison then also reflects the sampled prediction

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

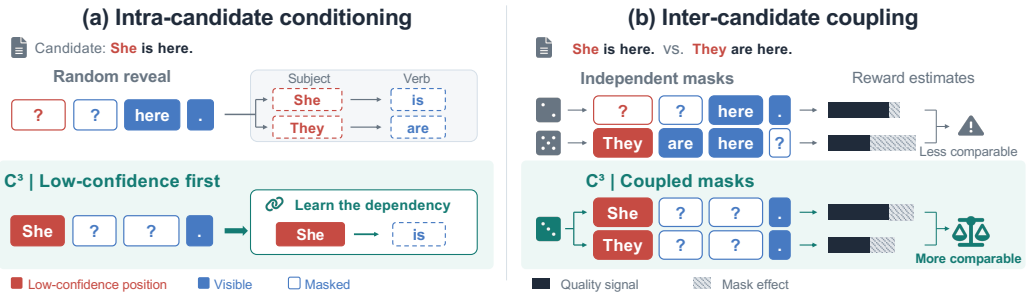


Figure 1: **Candidate-specific conditioning and coupled preference evaluation.** (a) Selected candidate tokens condition the predictions at the remaining positions. (b) Shared progress, count uniforms, and ordering keys coordinate candidate contexts while preserving their individual distributions and reveal budgets. Joint conditional scores receive sequence-level reward supervision. Solid and hatched bar segments schematically illustrate sequence-specific evidence and mask-induced variation.

conditions. Existing work coordinates noise or generated candidates to improve diffusion preference learning (Wallace et al., 2024; Liang et al., 2025; Hu et al., 2025). Here the design constraint is to coordinate *different candidate-specific context distributions*. Identical subsets are generally incompatible with different prescribed context distributions. Joint sampling should retain each candidate's conditioning rule while reducing avoidable disagreement in the comparison. The relevant reference compares context-averaged candidate scores, rather than requiring identical visible tokens. A fluctuation that shifts all scores equally cancels in the ranking or reward-matching loss, whereas a fluctuation in score differences can change that loss. This distinction motivates coordinating the random inputs to conditional evaluation instead of imposing a common visible subset.

We introduce *Candidate-specific Context Coupling* (C^3), a marginal-preserving construction for nonlinear preference comparisons (Figure 1). Its practical instantiation sets low-confidence reveal priorities from one evaluation of each completed candidate. Candidate-local priorities and token values specify the conditioning content. Shared progress and count uniforms align reveal budgets, while shared Gumbels align orders within those budgets. Each candidate combines the same random keys with its own priorities, producing correlated orders that retain its prescribed distribution. Candidates can consequently reveal different subsets while sharing the randomness used to select them. Priority selection determines candidate-local context marginals. At a fixed model, coupling changes only their joint distribution, preserving average conditional scores while governing how their fluctuations enter a nonlinear preference loss. Comparison is thus part of context design.

The construction operates during preference training, with candidate groups, rewards, scoring support, and the pairwise or listwise loss specified by each task. The score determines which tokens contribute to the loss and how their contributions are weighted. Coupling requires no token-wise reward labels: each completed candidate supplies its own context, and its sequence-level reward supervises the resulting conditional score. External conditions, such as a protein backbone or a language prompt, retain their original role alongside this candidate-local context. After C^3 fine-tuning, each model uses its native decoder. This separates training-time context design from generation-time guidance while allowing the same sampler to serve DNA bases, amino-acid residues, and response tokens.

Contributions. Candidate-specific coupling: C^3 coordinates candidate-specific contexts for nonlinear reward comparisons while preserving each selected context distribution. It couples reveal progress, counts, and orders for pairwise and listwise losses. Low-confidence priorities provide a practical instantiation, requiring one detached model evaluation per candidate. **Theoretical guarantees:** At fixed priorities on a common eligible set, we prove that shared Gumbels simultaneously minimize expected Kendall disagreement for all candidate pairs while preserving their full Plackett–Luce permutation laws. Under score-response conditions, this coupling minimizes a derived upper bound on the gap between sampled and context-averaged preference objectives with the joint distribution of shared progress and reveal counts held fixed. **Empirical evaluation:** We demonstrate gains in DNA design, protein inverse folding, and language modeling. Against SDPO, median DNA activity rises from 6.30 to 6.41 and ATAC-Acc from 94.8% to 97.7%. Median protein Pred-ddG increases by 17.4%, with the highest joint success of 79.7%. C^3 leads four of six language metrics.

2 RELATED WORK

Reward-guided diffusion. At sampling time, classifier-free and discrete guidance use conditional predictions or auxiliary predictors (Ho & Salimans, 2022; Nisonoff et al., 2025; Schiff et al., 2025), while particle methods and logit corrections modify sampling transitions (Wu et al., 2023; Dou et al., 2026). Post-training approaches include pairwise optimization with DPO and Diffusion-DPO (Rafailov et al., 2023; Wallace et al., 2024) and listwise alignment with LAIR (Wang et al., 2026). Discrete-diffusion methods include D2-DPO (Borso et al., 2025), SDPO's stepwise reward matching (Han et al., 2026), and DDPP's reward-tilted posteriors (Rector-Brooks et al., 2025); VRPO and ELBO-KTO address preference estimation (Zhu et al., 2026; Jindal et al., 2025). Other approaches optimize rewards with policy gradients (Black et al., 2024; Zekri & Boullé, 2025; Zhao et al., 2025; Ni et al., 2026), differentiable reward optimization (Wang et al., 2025), or forward-process and local-posterior matching (Zheng et al., 2026; Chen et al., 2026).

Training contexts and comparison coupling. Within a training example, LIFT selects targets by difficulty and time, restoring unselected masked tokens as context (Parashar et al., 2026). PAPL incorporates planner-based reverse dynamics into training (Peng et al., 2026); context-sensitive training and adaptive unmasking further modify conditioning (Oh et al., 2026; Koh et al., 2025; Jazbec et al., 2026; Hong et al., 2026). Across preference candidates, Diffusion-DPO shares noise and timesteps (Wallace et al., 2024), while SPO and D-Fusion coordinate noisy states or sample consistency (Liang et al., 2025; Hu et al., 2025). DiffuCoder couples complementary masks within one completion for probability estimation (Gong et al., 2026). C^3 designs the joint contexts of nonlinear preference comparisons under fixed candidate-specific marginals. For Plackett–Luce reveal orders, it attains simultaneous Kendall minima and, under score-response conditions, minimizes a preference-discrepancy upper bound. Low-confidence conditioning instantiates these marginals.

3 PRELIMINARIES

Masked diffusion. Let $y \in \mathcal{V}^L$ be a completed sequence, c its condition, I its eligible positions, and $\mathtt{m} \notin \mathcal{V}$ the mask token. The partial input x keeps conditioning tokens visible and padding unmasked. Reveal progress $t \in [0,1]$ runs from fully masked to revealed. Under absorbing-mask corruption, each eligible position is visible independently with probability t: $q_t^{\mathrm{mask}}(x \mid y) = \prod_{\ell \in I}[t\mathbf{1}\{x_\ell = y_\ell\} + (1-t)\mathbf{1}\{x_\ell = \mathtt{m}\}]$. The denoiser predicts $p_\theta(Y_\ell = a \mid x, t, c)$. Masked diffusion and discrete flow matching learn these conditionals (Sahoo et al., 2024; Shi et al., 2024; Gat et al., 2024).

Denoising evidence. With a fully masked prior, carry-over parameterization, and $t \sim \mathrm{Unif}(0,1)$, the continuous-time masked-diffusion evidence lower bound is (Sahoo et al., 2024; Shi et al., 2024)

$$\mathcal{E}_\theta(y \mid c) = \mathbb{E}_{t,\, x \sim q_t^{\mathrm{mask}}(\cdot \mid y)} \left[\frac{1}{1-t} \sum_{\ell : x_\ell = \mathtt{m}} \log p_\theta(Y_\ell = y_\ell \mid x, t, c) \right] \le \log p_\theta(y \mid c). \tag{1}$$

The likelihood belongs to the corresponding reverse process. The bound connects conditional token predictions to sequence likelihood through the specified masking law and time weights.

Candidates and conditional scores. A group $\mathcal{G}_g = (c_g, \{(y_{gk}, R_{gk})\}_{k=1}^{K_g})$ contains completed candidates and fixed rewards. A preference pair has $K_g = 2$. Candidates may be pre-collected or obtained during training; their sequences and rewards remain fixed within an update. A minibatch $\mathcal{B}$ contains group indices. We omit g when discussing one group. Policy p_θ and frozen reference $p_{\bar{\theta}}$ evaluate the same partial input (Wallace et al., 2024; Han et al., 2026), giving

$$A_{gk}(\theta; x_{gk}, t_{gk}) = \sum_{\ell \in I_{gk}} v_{gk\ell} \log \frac{p_\theta(Y_\ell = y_{gk\ell} \mid x_{gk}, t_{gk}, c_g)}{p_{\bar{\theta}}(Y_\ell = y_{gk\ell} \mid x_{gk}, t_{gk}, c_g)}. \tag{2}$$

Nonnegative weights $v_{gk\ell}$ specify scoring support, time weighting, and normalization. Masked-only scores exclude visible positions. In Equation 2, probabilities at scored positions are raw network outputs before deterministic copying. All-valid scores include visible-token reconstruction terms. With the masking law and score weights of Equation 1, the expected relative score equals the policy–reference ELBO difference.

Post-training objectives. A loss Φ_{R_g} maps the group scores to a training objective. We use pairwise logistic ranking and listwise reward-distribution matching. Both are convex in the candidate scores and invariant to a common score shift. Appendix A.3 specifies these objectives, and Appendix C records their application conventions. The score, context distribution, and loss together define the conditional surrogate. Score-weight rules are fixed for the update. Realized weights, contexts, and reference parameters are held constant during differentiation. Section 4 constructs their joint distribution without token-wise reward labels or additive trajectory rewards.

4 CANDIDATE-SPECIFIC CONTEXT COUPLING

We present Candidate-specific Context Coupling (C^3) for preference learning with candidate-specific conditioning. Section 4.1 separates the context laws from their coupling. Sections 4.2 and 4.3 construct these laws and their shared sampler, connecting Kendall-optimal order alignment to a preference-discrepancy bound. Section 4.4 gives the training update.

4.1 DESIGNING EVIDENCE FOR PREFERENCE LEARNING

Sequence rewards evaluate completed candidates, while the denoiser learns from partial inputs. A candidate-specific rule determines which choices condition its predictions. Once these rules are selected, the design problem is to coordinate their joint realization without altering any candidate's context distribution. We fix the candidate group, rewards, scoring rules, and context-construction parameters throughout each update.

Let $Z_k = (X_k, T_k)$ have candidate-specific law q_k, and let Π be a joint law with these marginals. Write $A_k = A_k(\theta; Z_k)$ as in Equation 2 and $\bar{A}_k = \mathbb{E}_{q_k} A_k$. Averaging each candidate's scores before comparison gives the reference $\mathcal{L}_{\mathrm{avg}} = \Phi_R(\bar{A}_1, \ldots, \bar{A}_K)$. Practical training compares a few sampled contexts (Zhu et al., 2026; Han et al., 2026), giving

$$\mathcal{L}_\Pi(\theta) = \mathbb{E}_\Pi \Phi_R(A_1, \ldots, A_K), \qquad \Delta_\Pi(\theta) = \mathcal{L}_\Pi(\theta) - \mathcal{L}_{\mathrm{avg}}(\theta), \qquad \Pi_k = q_k. \tag{3}$$

Convex ranking and reward-distribution losses give $\Delta_\Pi \geq 0$. Independent sampling preserves each marginal but leaves their joint realization uncoordinated. Our design variable is this dependence: coupling can change the discrepancy while preserving every candidate's score distribution and hence its mean. After specifying the marginals and sharing progress and count uniforms, Section 4.3 derives a score-sensitive upper bound on Δ_Π and a shared-order construction that minimizes it.

4.2 BUDGETED CANDIDATE-SPECIFIC CONDITIONING

Which candidate choices should provide context? Completed preference candidates provide token values that can condition the remaining predictions. We use low-confidence-first as a plug-in selector, ranking positions by completed-input output concentration. Its motivation is to expose candidate choices for conditional learning; its task effects are tested by the priority ablations. Preferred and dispreferred candidates each supply their own tokens, with rewards directing the update.

At checkpoint θ_s, let I_k contain the n_k eligible positions, excluding external conditioning and padding. Evaluate the completed candidate $(y_k, 1, c)$ once, with dropout, augmentation, and gradients disabled. The resulting concentration statistic and reveal priorities are

$$\begin{aligned} \gamma_{k\ell} &= \max_{v \in \mathcal{V}} p^{\mathrm{raw}}_{\theta_s}(Y_\ell = v \mid y_k, 1, c), \\ \widetilde{\gamma}_{k\ell} &= \mathrm{clip}\left(\frac{\gamma_{k\ell} - \bar{\gamma}_k}{s_k + \epsilon}, -a, a \right), \qquad \eta_{k\ell} = \zeta \widetilde{\gamma}_{k\ell}, \quad \zeta < 0. \end{aligned} \tag{4}$$

The raw non-mask probabilities are read before visible-token copying. Here $\bar{\gamma}_k$ and s_k are the mean and population standard deviation over I_k; $\epsilon > 0$ stabilizes normalization and $a > 0$ bounds the standardized range. This statistic measures output concentration with the candidate token visible, not a masked-token reconstruction probability. Its maximizing vocabulary token need not equal $y_{k\ell}$, but that alternative is not substituted into the candidate. Confidence selects positions only: the revealed value remains $y_{k\ell}$, while completed sequences and their rewards stay fixed within the update. Negative ζ gives lower-concentration positions higher priority, and $|\zeta|$ controls the strength of that preference in the randomized order.

A fixed budget with candidate-specific content. Let P_k be the Plackett–Luce (PL) law with weights $e^{\eta_{k\ell}}$ (Kool et al., 2019). It selects successive positions with probabilities proportional to their weights among the remaining indices. Independently of the order, retain $t \sim \nu$ and $N_k \mid t \sim \mathrm{Binomial}(n_k, t)$. Revealing the first N_k positions defines q_k; its explicit subset probabilities are given in Appendix A.5. Each position keeps the candidate's own token value, while the scoring support remains specified by Equation 2. At $\zeta = 0$, uniform prefixes recover the absorbing-mask marginal. Changing priority direction or strength changes the selected information while retaining full subset support and the prescribed reveal-count distribution.

Priority choice and coupling. Priorities specify each candidate's context law before cross-candidate sampling. Low-confidence-first is a practical choice within the PL family. The shared sampler accepts any finite priorities fixed before its draws. Section 5.4 tests priority choice and coupling separately.

4.3 Preference-Aware Context Coupling

Shared sampling. The candidate laws specify which contexts each score should receive. We coordinate their realizations by sharing progress, count uniforms, and ordering keys. For each group independently, draw $t \sim \nu$ and two arrays $u^{\mathrm{cnt}}, u^{\mathrm{ord}}$ on $I = \bigcup_k I_k$. All entries are independent uniforms on $(0, 1)$ and independent of t. With $G^{(k)} = (G_\ell)_{\ell \in I_k}$, set

$$N_k(t) = \sum_{\ell \in I_k} \mathbf{1}\{u_\ell^{\mathrm{cnt}} \le t\}, \qquad G_\ell = -\log(-\log u_\ell^{\mathrm{ord}}),$$
$$\pi_k = \mathrm{argsort}_{\downarrow}(\eta_k + G^{(k)}), \quad S_k(t) = \{\pi_k(1), \ldots, \pi_k(N_k(t))\}. \tag{5}$$

Copy y_k on $S_k(t)$ and mask the other eligible positions; conditioning and padding remain unchanged. Shared progress aligns expected reveal fractions, and shared count uniforms match counts for equal eligible sets (*budget alignment*). Shared Gumbels (Glasserman & Yao, 1992; Daliri et al., 2024) coordinate the candidate-specific orders (*order alignment*). Each binomial count remains independent of its PL order, preserving q_k and each candidate's conditioning content.

Proposition 1 (Kendall-optimal coupling at fixed priorities)**.** *Fix finite priorities on a common eligible set I. Let $p_{k,\ell m} = \sigma(\eta_{k\ell} - \eta_{km})$, with σ the logistic function, and let d_{K} count pairwise order reversals. The shared Gumbels attain, simultaneously for every candidate pair,*

$$\min_{\Lambda:\, \pi_k \sim P_k \ \forall k} \mathbb{E}_\Lambda d_{\mathrm{K}}(\pi_i, \pi_j) = \sum_{\substack{\ell < m \\ \ell, m \in I}} |p_{i,\ell m} - p_{j,\ell m}|. \tag{6}$$

Shared Gumbel differences make the pairwise ordering events nested, attaining every disagreement lower bound through compatible complete PL permutations. The minimum measures ordering disagreement required by the different candidate marginals. Independent orders incur additional reversals; shared Gumbels remove that excess simultaneously for all pairs. This optimum covers nonnegative candidate-pair weights, including reward-ordered comparisons (Appendix A.2).

A preference-sensitive order objective. Fix the candidate PL marginals and joint distribution of shared progress and reveal counts; vary only their permutation coupling $\Lambda \in \mathcal{C}(P_1, \ldots, P_K)$ independently of the budget draws. Write $a_k(S, t) = A_k(\theta; x_k^S, t)$ for deterministic scores on a common eligible set. Assume finite second moments and finite $L_j \ge 0$ satisfying $|a_j(S, t) - a_j(S', t)| \le L_j |S \triangle S'|$ for equal-sized sets at sampled progress values. Define $V_{ij} = \mathrm{Var}[a_i(S_i, t) - a_j(S_i, t)]$ under candidate i's marginal, using each candidate's tokens on the same visible indices. Both V_{ij} and L_j are fixed as Λ varies.

Corollary 1 (A preference-discrepancy objective)**.** *For pairwise logistic loss with fixed weights $\omega_{ij} \ge 0$ on reward-ordered pairs $\mathcal{P}_R$, the induced discrepancy satisfies, under the conditions above,*

$$0 \le \Delta_\Lambda^{\mathrm{pair}} \le \mathcal{U}(\Lambda) := \frac{\beta^2}{8} \sum_{(i,j) \in \mathcal{P}_R} \omega_{ij} \left(\sqrt{V_{ij}} + 2L_j \sqrt{\mathbb{E}_\Lambda d_{\mathrm{K}}(\pi_i, \pi_j)} \right)^2. \tag{7}$$

Shared Gumbels minimize $\mathcal{U}$ when the candidate priority laws and the joint distribution of shared progress and reveal counts are fixed.

Common-index score fluctuations contribute V_{ij}, while mismatched positions contribute through L_j and Kendall distance. Combining this decomposition with logistic curvature gives the bound. Every term of $\mathcal{U}$ is nondecreasing in its distance, so Proposition 1 selects its minimizing coupling without estimating V_{ij} or L_j. Appendix A.6 gives the proof and extensions to listwise matching and stochastic scoring. Proposition 2 extends the ordering and loss bounds to unequal response lengths with an explicit length-mismatch term.

4.4 Preference Learning from Coupled Evidence

At update s, evaluate the policy and frozen reference on the coupled inputs. Substituting their conditional scores into the selected task loss gives

$$\widehat{\mathcal{L}}_s(\theta) = \frac{1}{|\mathcal{B}|} \sum_{g \in \mathcal{B}} \Phi_{R_g} \left(\left[\sum_{\ell \in I_{gk}} v_{gk\ell} \log \frac{p_\theta(Y_\ell = y_{gk\ell} \mid X_{gk}, t_g, c_g)}{p_{\bar{\theta}}(Y_\ell = y_{gk\ell} \mid X_{gk}, t_g, c_g)} \right]_{k=1}^{K_g} \right). \tag{8}$$

During differentiation, contexts, priorities, realized score weights, and reference parameters are fixed. Under Appendix A.6's expectation-interchange conditions, the expected raw gradient differentiates $\mathcal{L}_{\Pi_s}$ with the update-s law Π_s frozen. The updated policy supplies the next step's priorities.

Each candidate contributes one score reused across its comparisons. The joint score vector determines the preference coefficients multiplying the conditional score gradients, including interactions among pairs that share a candidate. Section 5.4 evaluates this full-group direction against $\nabla_\theta \mathcal{L}_{\mathrm{avg}}$, which averages candidate scores before applying the loss (Appendix A.4). These measurements evaluate the aggregate response separately from the ordering guarantee.

Algorithm 1 gives the low-confidence instantiation: Conf and $\mathrm{StdClip}$ implement Equation 4, Top_n selects the n largest-entry indices, and $\mathrm{Mask}(y, E)$ masks E. Set $I_g = \bigcup_k I_{gk}$ and apply candidate-indexed operations to all k.

Algorithm 1 Candidate-specific Context Coupling update

Require: $\mathcal{B}, \theta_s, \bar{\theta}, \zeta < 0, \nu, \Phi_R$

1: **for** $g \in \mathcal{B}$ **do**
2: $\quad \eta_{gk} \leftarrow \zeta\, \mathrm{StdClip}(\mathrm{Conf}_{\theta_s}(y_{gk}, 1, c_g); I_{gk})$
3: $\quad t_g \sim \nu; \quad u_g^{\mathrm{cnt}}, u_g^{\mathrm{ord}} \overset{\mathrm{ind}}{\sim} \mathrm{Unif}(0,1)^{I_g}$
4: $\quad G_g \leftarrow -\log(-\log u_g^{\mathrm{ord}}); \quad G_g^{(k)} \leftarrow (G_{g\ell})_{\ell \in I_{gk}}; \quad N_{gk} \leftarrow \sum_{\ell \in I_{gk}} \mathbf{1}\{u_{g\ell}^{\mathrm{cnt}} \le t_g\}$
5: $\quad S_{gk} \leftarrow \mathrm{Top}_{N_{gk}}(\eta_{gk} + G_g^{(k)}); \quad X_{gk} \leftarrow \mathrm{Mask}(y_{gk}, I_{gk} \setminus S_{gk})$
6: $\quad a_{gk}(\theta) \leftarrow A_{gk}(\theta; X_{gk}, t_g)$
7: $\quad \ell_g(\theta) \leftarrow \Phi_{R_g}((a_{gk}(\theta))_{k=1}^{K_g})$
8: **end for**
9: $\theta_{s+1} \leftarrow \mathrm{Update}(\theta_s, \nabla \widehat{\mathcal{L}}_s(\theta_s))$

Lines 2–5 use no gradients. Context construction adds one detached policy evaluation and one sort per candidate. Sharing requires no extra denoiser evaluations. The candidate scores are reused in $O(K_g^2)$ scalar operations per group for all-pairs ranking and $O(K_g)$ for listwise matching.

5 Experiments

We evaluate C^3 on DNA sequence design (Section 5.1), protein inverse folding (Section 5.2), and language modeling (Section 5.3). Protein uses rank-gap-weighted all-pairs ranking, while language and DNA use listwise reward-distribution matching. For C^3, generation uses each base model's native decoder. Appendix C gives the task-specific scoring and training protocols. Evaluation separates the reward used for alignment from complementary checks on the generated sequences. DNA includes sequence-statistic agreement and diversity, protein includes structural consistency and joint success, and language combines response judging with task-specific tests.

Table 1: DNA sequence design. C^3 combines high predicted activity with strong accessibility and motif agreement. Baseline values are from Han et al. (2026), except for our GILC-DB reproduction. Means and standard deviations follow the reported seed protocols. Bold/underline denote the best/second-best results. † Local entropy sums position-wise nucleotide entropies.

Method	Pred-Activity (med) ↑	ATAC-Acc ↑ (%)	3-mer Corr ↑	JASPAR Corr ↑	App-Log-Lik (med) ↑	Entropy† ↑
Pretrained (Sahoo et al., 2024)	0.17 (0.040)	1.5 (0.20)	−0.061 (0.034)	0.249 (0.015)	−261 (0.6)	<u>390</u> (6.2)
Test-time guidance						
CG (Nisonoff et al., 2025)	3.30 (0.000)	0.0 (0.00)	−0.065 (0.001)	0.212 (0.035)	−266 (0.6)	12 (4.1)
SMC (Wu et al., 2023)	4.15 (0.330)	39.9 (8.70)	0.840 (0.045)	0.756 (0.068)	−259 (2.5)	351 (6.5)
TDS (Wu et al., 2023)	4.64 (0.210)	45.3 (16.40)	0.848 (0.008)	0.846 (0.044)	−257 (1.5)	340 (5.4)
CFG (Ho & Salimans, 2022)	5.04 (0.060)	92.1 (0.90)	0.746 (0.001)	0.864 (0.011)	−265 (0.6)	363 (6.1)
GILC-DB (Dou et al., 2026)	6.19 (0.040)	85.8 (1.00)	0.794 (0.006)	0.893 (0.003)	−278 (0.3)	**397** (0.1)
Post-training alignment						
D2-DPO (Borso et al., 2025)	2.97 (0.030)	35.6 (0.90)	**0.944** (0.002)	0.883 (0.005)	−252 (0.4)	362 (4.9)
VRPO (Zhu et al., 2026)	4.60 (0.010)	15.8 (0.20)	0.838 (0.002)	0.865 (0.005)	−255 (0.8)	289 (13.5)
DDPP-IS (Rector-Brooks et al., 2025)	4.07 (0.020)	50.0 (0.30)	0.711 (0.001)	0.723 (0.004)	−253 (0.9)	378 (5.8)
DRAKES (Wang et al., 2025)	5.61 (0.070)	92.5 (0.60)	0.887 (0.002)	0.911 (0.002)	−264 (0.6)	375 (5.2)
diffu-GRPO (Zhao et al., 2025)	5.86 (0.040)	33.0 (0.80)	0.783 (0.001)	0.903 (0.004)	**−245** (0.4)	310 (8.6)
SDPO (Han et al., 2026)	<u>6.30</u> (0.003)	<u>94.8</u> (0.01)	0.900 (0.003)	<u>0.936</u> (0.003)	<u>−246</u> (0.5)	365 (4.4)
C^3	**6.41** (0.020)	**97.7** (0.36)	<u>0.911</u> (0.005)	**0.956** (0.002)	−249 (0.3)	356 (0.4)

5.1 DNA SEQUENCE DESIGN

Setup and metrics. We design 200-base-pair DNA sequences for predicted HepG2 enhancer activity using MPRA data (Gosai et al., 2024) and the pretrained masked-diffusion model released with DRAKES (Sahoo et al., 2024; Wang et al., 2025). We use the same HepG2 activity reward as SDPO (Han et al., 2026). Following its evaluation protocol, we report median predicted activity (Pred-Activity) and the fraction with predicted chromatin accessibility above 0.5 (ATAC-Acc). Pearson 3-mer and Spearman JASPAR motif correlations (Castro-Mondragon et al., 2022) compare generated sequence statistics with the top 0.1% activity set. App-Log-Lik is the median pretrained-model ELBO estimate; positional entropy measures diversity.

Results. C^3 leads Pred-Activity, ATAC-Acc, and JASPAR Corr, with the second-highest 3-mer Corr (Table 1). Relative to SDPO, activity rises from 6.30 to 6.41, accessibility from 94.8% to 97.7%, and motif correlation from 0.936 to 0.956. The 3-mer correlation also increases from 0.900 to 0.911, showing agreement with high-activity sequences at both short-subsequence and motif scales.

For DNA, C^3 uses its selected candidate bases as context for the surrounding sequence. Listwise reward matching compares predictions conditional on these candidate-specific choices using each completed sequence's activity reward. Higher 3-mer and motif correlations support target-pattern compatibility alongside the activity and accessibility gains.

5.2 PROTEIN INVERSE FOLDING

Setup and metrics. We optimize amino-acid sequences conditioned on backbone structures, starting from DRAKES' pretrained inverse-folding checkpoint (Wang et al., 2025). Training uses 438,540 reward-labeled records from 311 Megascale structures (Han et al., 2026; Tsuboyama et al., 2023). Pred-ddG measures predicted stability improvement. Self-consistency RMSD (scRMSD) compares an ESMFold-predicted structure (Lin et al., 2023) with its conditioning backbone in Å. We report the fractions with Pred-ddG > 0 and scRMSD < 2 Å, together with joint success requiring both conditions. Median positional entropy across backbones measures sequence diversity.

Results. C^3 leads predicted stability, positive-stability rate, joint success, and entropy (Table 2). Joint success reaches 79.7%, exceeding SDPO by 4.2 percentage points and DRAKES by 1.1 points. Against SDPO, median Pred-ddG rises from 1.400 to 1.644, positive stability from 87.1% to 92.0%, and entropy from 42.3 to 45.4 bits. Structural success is 86.4% versus SDPO's 88.9%.

The backbone condition is shared, but each candidate supplies its own amino-acid choices at low-confidence positions. Other residue predictions are evaluated conditional on these choices and the same backbone structure. Coupled contexts retain this candidate-specific information while coordinating the stability-based comparison. Higher joint success with greater residue diversity matches this aim of learning multiple successful sequences for a backbone.

Table 2: Protein inverse folding. C^3 achieves the highest joint success for predicted stability and structural consistency. Baseline values are from Han et al. (2026), except for our GILC-DB reproduction. Means and standard deviations follow the reported seed protocols. Bold/underline denote the best/second-best results. † Entropy is summarized by the median across backbones.

Method	Pred-ddG (med) ↑	%(ddG > 0) (%) ↑	scRMSD (med) ↓	%(scRMSD < 2) (%) ↑	Success Rate (%) ↑	Entropy (med)† ↑
Pretrained (Campbell et al., 2024)	−0.544 (0.037)	36.6 (1.0)	0.849 (0.013)	90.9 (0.6)	34.4 (0.5)	35.2 (8.1)
Test-time guidance						
CG (Nisonoff et al., 2025)	−0.561 (0.045)	36.9 (1.1)	0.839 (0.012)	90.9 (0.6)	34.7 (0.9)	34.6 (7.1)
SMC (Wu et al., 2023)	0.659 (0.044)	68.5 (3.1)	0.841 (0.006)	<u>93.8</u> (0.4)	63.6 (4.0)	24.9 (6.9)
TDS (Wu et al., 2023)	0.674 (0.086)	68.2 (2.4)	<u>0.834</u> (0.001)	**94.4** (1.2)	62.9 (2.8)	24.9 (7.2)
CFG (Ho & Salimans, 2022)	−1.186 (0.035)	11.0 (0.4)	3.146 (0.062)	29.4 (1.0)	1.3 (0.4)	8.4 (5.9)
GILC-DB (Dou et al., 2026)	0.465 (0.074)	60.4 (1.4)	0.864 (0.005)	89.6 (0.6)	53.2 (1.6)	38.4 (0.4)
Post-training alignment						
D2-DPO (Borso et al., 2025)	0.500 (0.051)	66.4 (0.3)	0.909 (0.005)	93.6 (0.8)	61.0 (0.5)	41.7 (7.4)
VRPO (Zhu et al., 2026)	0.548 (0.032)	61.1 (0.1)	0.883 (0.004)	93.5 (0.7)	56.6 (0.3)	39.1 (9.3)
DDPP-IS (Rector-Brooks et al., 2025)	−0.130 (0.047)	46.7 (0.8)	**0.829** (0.008)	89.3 (0.6)	43.3 (0.5)	24.3 (7.6)
DRAKES (Wang et al., 2025)	1.095 (0.026)	86.4 (0.2)	0.918 (0.006)	91.8 (0.5)	<u>78.6</u> (0.7)	33.3 (6.4)
diffu-GRPO (Zhao et al., 2025)	1.286 (0.021)	76.8 (0.3)	1.192 (0.005)	57.1 (0.8)	37.2 (1.4)	40.0 (7.8)
SDPO (Han et al., 2026)	<u>1.400</u> (0.014)	<u>87.1</u> (0.01)	0.938 (0.005)	88.9 (0.3)	75.5 (0.3)	<u>42.3</u> (6.5)
C^3	**1.644** (0.021)	**92.0** (0.3)	1.063 (0.004)	86.4 (0.3)	**79.7** (0.2)	**45.4** (1.3)

Table 3: Fine-tuning LLaDA-8B-Instruct on UltraFeedback from Cui et al. (2024). C^3 improves judged response quality and code generation. We reproduce the fine-tuning baselines and evaluate every row under the same protocol. Single-run percentages; bold marks column-wise best results.

Method	Alpaca WR ↑	Alpaca LC-WR ↑	HumanEval ↑	MBPP ↑	GSM8K ↑	IFEval ↑
Base (Nie et al., 2025)	5.53	10.57	45.73	48.25	78.92	61.38
D2-DPO (Borso et al., 2025)	5.40	9.79	46.34	49.03	**79.91**	61.62
Diffu-GRPO (Zhao et al., 2025)	6.16	11.57	48.17	47.47	77.03	59.25
VRPO (Zhu et al., 2026)	5.56	10.09	46.95	48.64	79.76	62.29
SDPO (Han et al., 2026)	5.70	10.57	46.95	48.25	78.32	60.23
JustGRPO (Ni et al., 2026)	5.14	9.10	45.73	48.25	79.00	**62.33**
C^3	**6.31**	**12.06**	**48.78**	**50.19**	79.45	62.14

5.3 Language modeling

Setup and metrics. We fine-tune LLaDA-8B-Instruct (Nie et al., 2025) on 59,876 UltraFeedback prompts (Cui et al., 2024), keeping each prompt's highest- and lowest-reward responses. Reproduced fine-tuning baselines share the evaluation pipeline. AlpacaEval reports win rate (WR) and length-controlled win rate (LC-WR) (Dubois et al., 2024). HumanEval and MBPP measure code pass@1, GSM8K tests reasoning, and IFEval evaluates verifiable instruction following (Zhou et al., 2023).

Results. C^3 leads four of the six metrics in Table 3. Against the strongest baseline for each metric, WR and LC-WR improve by 0.15 and 0.49 percentage points, while HumanEval and MBPP improve by 0.61 and 1.16 points. GSM8K reaches 79.45%, a 1.13-point gain over SDPO; IFEval reaches 62.14%, up from the base model's 61.38%.

Prompt tokens remain visible throughout training, while each response contributes its own low-confidence tokens to the context. Masked-only scoring trains the unrevealed response tokens conditional on each candidate's supplied choices. Reward-distribution matching then compares completed responses through coordinated conditional scores. Gains on response judging and executable tests are consistent with more effective response-level supervision of conditional predictions.

5.4 Ablation study

Conditioning direction and strength. At fixed full sharing, low-confidence protein priorities ($\zeta = -7.5$) give median Pred-ddG 1.642 and joint success 78.51%, compared with 1.519 and 76.40% for uniform priorities. Reversing the sign at the same magnitude gives 1.611 and 60.42% (Figure 2b). Low-confidence selection improves median stability by 8.1% over uniform and joint success by 18.09 points over reversal. The reveal budget is fixed, isolating the priority choice in Section 4.2. Halving the negative strength to -3.75 retains Pred-ddG 1.612 and 78.14% joint success; the reference strength combines higher median reward with 43.99 versus 35.74 bits of entropy.

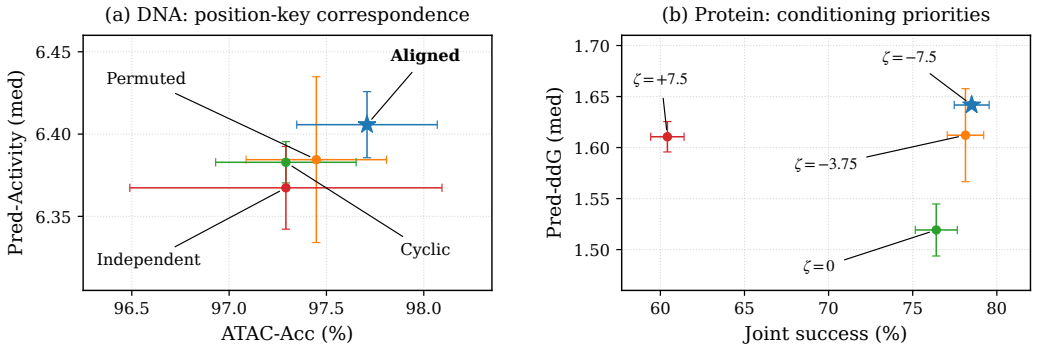


Figure 2: **Conditioning and coupling in sequence design.** (a) DNA key correspondence with low-confidence priorities from the current policy. (b) Protein priority direction and strength with full sharing. Points and error bars show means and sample SDs over three training seeds on both axes. Stars mark the reference configurations; all plotted quantities are final-generation metrics.

Position-wise correspondence. To isolate cross-candidate correspondence, the DNA series uses the same current-policy, low-confidence-first rule across controls and varies only the correspondence of position keys. Aligned shared keys yield activity 6.406 and ATAC-Acc 97.71%, compared with 6.367 and 97.29% for independent keys, 6.384 and 97.45% for candidate-specific permutations, and 6.383 and 97.29% for cyclic shifts (Figure 2a). At a fixed policy, permutations and cyclic shifts retain candidate marginals while changing their correspondence. The higher activity and accessibility with aligned keys identify a coupling benefit beyond marginal preservation under the same priority rule.

Budget and order alignment. At a fixed protein checkpoint and batch, comparison loss falls from 0.4328 with independent contexts to 0.3660 with shared progress, 0.3607 with shared progress and reveal counts, and 0.3494 with full sharing. Shared progress supplies the largest decrease, and shared orders add 3.1% after budget alignment. Every control preserves the candidate score marginals, keeping $\mathcal{L}_{\text{avg}}$ in Equation 3 fixed. The loss differences estimate changes in the comparison gap. Against fully independent contexts, full coupling reduces score-difference variance by 94.67–98.14% in eight groups with identical empirical marginals. Using the gradient of this context-averaged objective as a reference, mean full-group cosine rises from 0.419 to 0.665, improving in all five collections. This diagnostic includes every candidate's gradient contribution and its preference-loss coefficient. These diagnostics hold the policy fixed to isolate the comparison mechanism. The generation results below instead assess the models after training, during which candidate-local priorities evolve with the policy and coupling influences the updates.

Joint yield after training. At fixed protein low-confidence priorities, full sharing achieves 78.51 $(1.04)\%$ joint success, versus 77.48 $(0.37)\%$ with private orders and 76.82 $(0.84)\%$ with independent contexts. Positive-stability rates are 90.69%, 88.04%, and 87.96%, with entropies 43.99, 34.99, and 30.44 bits. Private orders retain shared progress and reveal counts, so this comparison isolates the added effect of order alignment on joint yield and positional entropy.

6 Conclusion

We presented C^3 for coupling candidate-specific contexts in nonlinear preference learning. The construction supplies completed-candidate choices as context and coordinates reveal progress, counts, and orders while preserving each candidate's context distribution. Our analysis links its Kendall-optimal ordering guarantee to a score-sensitive preference-discrepancy bound minimized by the shared sampler under the stated conditions. Experiments demonstrate gains in DNA design, protein inverse folding, and language modeling. Protein ablations support low-confidence conditioning and shared orders. Coupling diagnostics show lower comparison loss and better full-group gradient alignment. Future work will explore task-adaptive priorities and extend the analysis from fixed-update objectives to evolving training dynamics.

AI USE STATEMENT

Generative AI was used to assist with writing and language polishing, as well as literature retrieval and the discovery of related work. The authors reviewed and verified AI-assisted content and are responsible for the paper's accuracy and integrity.

ETHICS STATEMENT

All authors have read and are committed to comply with the ICLR Code of Ethics https://iclr.cc/public/CodeOfEthics. We present a principled approach to preference optimization for discrete diffusion models through candidate-specific context coupling. Our goal is to develop a general algorithm for discrete diffusion alignment, and we do not identify major ethical concerns specific to the method. The reported generation results are computational and should be independently validated before use in downstream applications.

REPRODUCIBILITY STATEMENT

Appendix A supplies the proofs and gradient reference. Appendix B specifies the unified sampler and its ablation controls, while Appendices C and D record the task configurations, evaluation protocols, and additional results. The experimental appendices distinguish training seeds from generation repeats and local from source-reported results. Our anonymous code is available at https://anonymous.4open.science/r/C3-C9D7/.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/14f75513f0f1ca01de1e826b52e6b840-Abstract-Conference.html.

Umberto Borso, Davide Paglieri, Jude Wells, and Tim Rocktäschel. Preference-based alignment of discrete diffusion models. *arXiv preprint arXiv:2503.08295*, 2025. URL https://arxiv.org/abs/2503.08295v2.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 5453–5512. PMLR, 2024. URL https://proceedings.mlr.press/v235/campbell24a.html.

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. doi: 10.1093/nar/gkab1113. URL https://doi.org/10.1093/nar/gkab1113.

Yuyuan Chen, Shiyi Wang, Peter Potaptchik, Jaeyeon Kim, and Michael S. Albergo. Discrete tilt matching. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65069.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 9722–9744. PMLR, 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Majid Daliri, Christopher Musco, and Ananda Theertha Suresh. Coupling without communication and drafter-invariant speculative decoding. *arXiv preprint arXiv:2408.07978*, 2024. URL https://arxiv.org/abs/2408.07978.

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 *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64933.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f0d629a734b56a642701bba7bc8bb3ed-Abstract-Conference.html.

Paul Glasserman and David D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992. doi: 10.1287/mnsc.38.6.884. URL https://doi.org/10.1287/mnsc.38.6.884.

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. DiffuCoder: Understanding and improving masked diffusion models for code generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1a5b00e59818e4be1c89855a8b921903-Abstract-Conference.html.

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(8036):1211–1220, 2024. doi: 10.1038/s41586-024-08070-z. URL https://www.nature.com/articles/s41586-024-08070-z.

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*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0922ceebb39f1c76e64ef03d0885fa82-Abstract-Conference.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Chunsan Hong, Seonho An, Min-Soo Kim, and Jong Chul Ye. Improving discrete diffusion unmasking policies beyond explicit reference policies. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b0506debbf49e31d25690fbd1e69cd2f-Abstract-Conference.html.

Zijing Hu, Fengda Zhang, and Kun Kuang. D-Fusion: Direct preference optimization for aligning diffusion models with visually consistent samples. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 24869–24892. PMLR, 2025. URL https://proceedings.mlr.press/v267/hu25ab.html.

Metod Jazbec, Theo X. Olausson, Louis Béthune, Pierre Ablin, Michael Kirchhof, João Monteiro, Victor Guilherme Turrisi da Costa, Jason Ramapuram, and Marco Cuturi. Learning unmasking policies for diffusion language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65277.

Vaibhav Jindal, Hejian Sang, Chun-Mao Lai, Yanning Chen, and Zhipeng Wang. Aligning diffusion language models via unpaired preference optimization. *arXiv preprint arXiv:2510.23658*, 2025. URL https://arxiv.org/abs/2510.23658.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. *arXiv preprint arXiv:2502.06768*, 2025. URL https://arxiv.org/abs/2502.06768.

Hyukhun Koh, Minha Jhang, Dohyung Kim, Sangmook Lee, and Kyomin Jung. Conditional [MASK] discrete diffusion language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8899–8923. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.450. URL https://aclanthology.org/2025.emnlp-main.450/.

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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3499–3508. PMLR, 2019. URL https://proceedings.mlr.press/v97/kool19a.html.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13199–13208, 2025. doi: 10.1109/CVPR52734.2025.01232. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liang_Aesthetic_Post-Training_Diffusion_Models_from_Generic_Preferences_with_Step-by-step_Preference_CVPR_2025_paper.html.

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. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, and Gao Huang. The flexibility trap: Rethinking the value of arbitrary order in diffusion language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61998.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-1689. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/48b383b24230e0e6e649d9c98dae4d8c-Abstract-Conference.html.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/597254dc45be8c166d3ccf0ba2d56325-Abstract-Conference.html.

Seunghyuk Oh, Minjae Lee, Kevin Galim, Minseo Kim, Hyung Il Koo, Wonjun Kang, Hanbaek Lyu, and Kangwook Lee. CATS: Inference-aligned SFT for diffusion LLMs via context-sensitivity aware trajectory sampling. In *Proceedings of the 2nd DeLTa Workshop at ICLR 2026*, 2026. URL https://openreview.net/forum?id=RdMiEf9B0D.

Shubham Parashar, Atharv Chagi, Jacob Helwig, Lakshmi Jotsna, Sushil Vemuri, James Caverlee, Dileep Kalathil, and Shuiwang Ji. Learnability-informed fine-tuning of diffusion language models. *arXiv preprint arXiv:2605.22939*, May 2026. URL https://arxiv.org/abs/2605.22939.

Fred Zhangzhi Peng, Zachary Bezemek, Jarrid Rector-Brooks, Shuibai Zhang, Anru R. Zhang, Michael Bronstein, Alexander Tong, and Avishek Joey Bose. Planner aware path learning in

diffusion language models training. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1AlI5FuIf7.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Jarrid Rector-Brooks, Mohsin Hasan, Zhangzhi Peng, Zachary Quinn, Chenghao Liu, Sarthak Mittal, Nouha Dziri, Michael Bronstein, Yoshua Bengio, Pranam Chatterjee, Alexander Tong, and Avishek Joey Bose. Steering masked discrete diffusion models via discrete denoising posterior prediction. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3f2a9f41dd0cb07cd9b60c058c7b1682-Abstract-Conference.html.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Yair Schiff, Subham Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-Torre, Bernardo Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6cc31b44d88dce8380d36e81485cd07f-Abstract-Conference.html.

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 37*, 2024. doi: 10.52202/079017-3277. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html.

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. doi: 10.1038/s41586-023-06328-6.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Austin Wang, Jiaqi Han, Stefano Ermon, and Yisong Yue. Beyond pairwise preferences: Listwise reward-aware alignment for diffusion models. *arXiv preprint arXiv:2605.26491*, 2026. URL https://arxiv.org/abs/2605.26491.

Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Avantika Lal, Tommi Jaakkola, Sergey Levine, Aviv Regev, Hanchen Wang, and Tommaso Biancalani. Fine-tuning discrete diffusion models via reward optimization with applications to DNA and protein design. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/771e09dd204ea339da0d8114c48afd21-Abstract-Conference.html.

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 36*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63e8bc7bbf1cfea36d1d1b6538aecce5-Abstract-Conference.html.

Oussama Zekri and Nicolas Boullé. Fine-tuning discrete diffusion models with policy gradient methods. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-5112. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e053a2ffa2a7f5156e2b9a852fde6d45-Abstract-Conference.html.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-1898. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/52190a0362148d179f1cbd9080956872-Abstract-Conference.html.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d8e68ddfe22520b45b8fb8d5cbde5a21-Abstract-Conference.html.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. LLaDA 1.5: Variance-reduced preference optimization for large language diffusion models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 11425–11460, 2026. doi: 10.18653/v1/2026.acl-long.524. URL https://aclanthology.org/2026.acl-long.524/.

A Analysis of context coupling and group updates

This appendix proves the comparison-gap and reveal-order results, specifies the training losses and gradient reference, and verifies the candidate-local context law. Section A.6 derives the preference-discrepancy bound and proves its minimization by the shared sampler; Section A.7 extends these results to response-relative indices of unequal lengths. All expectations condition on the candidate group, rewards, optimization history, and model parameters used to compute confidence. For vectors, $\mathrm{Cov}(U, V) = \mathbb{E}[(U - \mathbb{E}U)(V - \mathbb{E}V)^\top]$.

A.1 Comparison objectives under random contexts

This subsection analyzes the context-averaged comparison used in Section 4.1. Fix the candidates, rewards, score-weight rules, and the parameters used to construct contexts. Each candidate has a fixed context law q_k. Only their coupling Π varies. Write $A = (A_1, \dots, A_K)$ and $\bar{A} = \mathbb{E}_\Pi A$; the latter depends only on the marginals. The analysis applies at a fixed policy checkpoint and to the fixed-priority objective used for its update. The nonlinearity of ELBO-based preference estimation is also analyzed by Zhu et al. (2026). Here we separate candidate-wise evidence from the discrepancy introduced by its joint realization.

Exact decomposition and contrast bound. Assume Φ_R is convex and twice continuously differentiable on an open convex domain containing the closed convex hull of the score support. Assume finite score second moments, and $\|\nabla^2 \Phi_R(a)\|_{\mathrm{op}} \le M$ throughout that hull. Define $\delta = A - \bar{A}$. Taylor's theorem in integral form yields the exact identity

$$\Delta_\Pi := \mathbb{E}_\Pi \Phi_R(A) - \Phi_R(\bar{A}) = \mathbb{E}_\Pi \int_0^1 (1-u)\delta^\top \nabla^2 \Phi_R(\bar{A} + u\delta)\delta \, du. \tag{9}$$

The linear term vanishes because $\mathbb{E}_\Pi \delta = 0$. Convexity makes the Hessian positive semidefinite, giving $\Delta_\Pi \ge 0$. Suppose additionally that $\Phi_R(a + b\mathbf{1}) = \Phi_R(a)$ for every real b. Differentiating this identity gives $\nabla^2 \Phi_R(a)\mathbf{1} = 0$. With $\bar{\delta} = K^{-1} \sum_k \delta_k$, the quadratic form in Equation 9 is unchanged by replacing δ with $\delta - \bar{\delta}\mathbf{1}$. The operator-norm bound and integration over u then give

$$\Delta_\Pi \le \frac{M}{2} \mathbb{E}_\Pi \|\delta - \bar{\delta}\mathbf{1}\|_2^2 = \frac{M}{2K} \sum_{i<j} \mathbb{E}_\Pi (\delta_i - \delta_j)^2 = \frac{M}{2K} \sum_{i<j} \mathrm{Var}_\Pi (A_i - A_j). \tag{10}$$

The middle equality follows by expanding $\sum_{i<j} (\delta_i - \delta_j)^2 = K \sum_k (\delta_k - \bar{\delta})^2$. This proves the contrast bound used in Appendix A.6. Both the exact gap and its bound concern the frozen context laws. When the laws are changed, the averaged scores and their comparison generally change as well.

For the normalized all-pairs logistic objective in Appendix A.3, $M = \beta^2/2$ is a valid global bound: each unit-weight term has Hessian norm at most $\beta^2/2$, and the nonnegative pair weights sum to one. The softmax cross-entropy at scale β also has Hessian norm at most $\beta^2/2$. Its Hessian is $\beta^2[\mathrm{diag}(p) - pp^\top]$; the absolute row sum is $2\beta^2 p_k(1 - p_k) \le \beta^2/2$. Both losses belong to the smooth, common-shift-invariant class used by the contrast bound. The contrast argument operates on the scalar candidate scores and permits different sequence lengths and scoring supports; equal-size prefixes enter only the subsequent score-response bound.

Relation to the expected update. When differentiation and expectation can be interchanged with the priorities fixed, $\nabla_\theta \mathcal{L}_\Pi = \nabla_\theta \Phi_R(\bar{A}) + \nabla_\theta \Delta_\Pi$. The reference in Equation 16 is the first term. A smaller loss gap alone does not imply a smaller gradient error. At matched deterministic policy/reference parameters, $A = 0$ for every input, so the loss gap and its first derivative vanish, although the sampled score gradients can still vary. Appendix A.4 gives the complete-gradient reference and covariance decomposition used by the diagnostics.

Loss scope. The discrepancy guarantees below concern pairwise logistic and listwise reward-matching losses. For comparison, a separable, integrable fixed-target loss satisfies $\mathbb{E}_\Pi \sum_k f_k(A_k) = \sum_k \mathbb{E}_{q_k} f_k(A_k)$ for every coupling; its stochastic update covariance can still change. This separates the nonlinear comparison effect studied here from that boundary case.

A.2 Optimal coordination of reveal orders

Fix any finite candidate-specific log-priorities before drawing the fresh ordering keys. They may depend on the candidates and the current optimization history; Equation 4 is one instantiation. The results below hold conditionally on these priorities and their induced PL marginals. We verify the individual permutation distributions and then prove their optimal joint realization.

Preserving the ranking marginals. Let P_k be the Plackett–Luce permutation distribution for candidate k, with $P_k(\pi) = \prod_{r=1}^{|I|} \frac{e^{\eta_{k,\pi(r)}}}{\sum_{s=r}^{|I|} e^{\eta_{k,\pi(s)}}}$. Here $\pi(r)$ is the position at rank r. For independent standard Gumbels G_ℓ, the variables $E_\ell = e^{-G_\ell}$ are independent unit-rate exponentials. Sorting $\eta_{k\ell} + G_\ell$ in decreasing order is equivalent to sorting $E_\ell / e^{\eta_{k\ell}}$ in increasing order. These are independent exponential arrival times with rates $e^{\eta_{k\ell}}$. The first arrival has probability proportional to its rate. The memoryless property yields the same rule among remaining indices, giving the permutation law P_k (Kool et al., 2019). Sharing the Gumbels across candidates retains these marginals because the coordinates remain independent within each candidate.

Let d_{K} denote Kendall distance and $p_{k,\ell m} = \sigma(\eta_{k\ell} - \eta_{km})$, with σ the logistic function. The pairwise minima in Proposition 1 are attained simultaneously. Their sum is

$$\min_{\Pi:\, \pi_k \sim P_k} \mathbb{E}_\Pi \sum_{i<j} d_{\mathrm{K}}(\pi_i, \pi_j) = \sum_{i<j} \sum_{\substack{\ell<m \\ \ell,m \in I}} |p_{i,\ell m} - p_{j,\ell m}|. \tag{11}$$

Proof of Proposition 1. Let $B_{k,\ell m} = \mathbf{1}\{\ell \text{ precedes } m \text{ in } \pi_k\}$ for $\ell < m$. Under marginal P_k, its mean is $p_{k,\ell m} = \sigma(\eta_{k\ell} - \eta_{km})$, obtained by comparing the two exponential arrival times. Every joint law Π preserving the permutation distributions P_k satisfies

$$\Pr_\Pi(B_{i,\ell m} \neq B_{j,\ell m}) = \mathbb{E}_\Pi |B_{i,\ell m} - B_{j,\ell m}| \geq |\mathbb{E}_\Pi B_{i,\ell m} - \mathbb{E}_\Pi B_{j,\ell m}| = |p_{i,\ell m} - p_{j,\ell m}|. \tag{12}$$

Summing over candidate and position pairs gives the lower bound in Equation 11.

Under shared keys, $B_{k,\ell m} = \mathbf{1}\{G_\ell - G_m > \eta_{km} - \eta_{k\ell}\}$. For each fixed position pair, all candidates threshold the same continuous random variable. Their events are nested, so their disagreement probability equals the absolute difference of their probabilities. Equality in Equation 12 holds simultaneously for every candidate pair and every position pair. The marginal argument establishes feasibility and the simultaneous pairwise minima; Appendix A.6 proves Corollary 1 separately. □

The result holds jointly for the complete permutations even though its proof uses binary ordering events. For comparison, independent orders give disagreement $p_i + p_j - 2p_ip_j$ for one position pair, exceeding the shared value by $2\min(p_i, p_j)[1 - \max(p_i, p_j)]$. The shared construction removes this excess without changing either priority law. It also minimizes any nonnegative weighted sum $\mathbb{E}_\Pi \sum_{i<j} w_{ij} d_{\mathrm{K}}(\pi_i, \pi_j)$, since it attains each summand's lower bound. The weights here specify evaluation criteria and require no change to the sampler.

From complete orders to visible sets. For two candidates with the same eligible set I, condition on reveal progress t and the shared count $n = N_i(t) = N_j(t)$. Let S_i, S_j be the length-n prefixes of their permutations and put $r = |S_i \setminus S_j| = |S_j \setminus S_i|$. Every position in $S_i \setminus S_j$ precedes every position in $S_j \setminus S_i$ under π_i, and follows it under π_j. These r^2 distinct pairs contribute to Kendall distance, giving $|S_i \triangle S_j|^2 = 4r^2 \leq 4d_{\mathrm{K}}(\pi_i, \pi_j)$. Count and order draws are independent, so the permutation laws remain P_i, P_j after conditioning on t, n. Taking expectations and applying Proposition 1 yields

$$\mathbb{E}|S_i(t) \triangle S_j(t)|^2 \leq 4\,\mathbb{E} d_{\mathrm{K}}(\pi_i, \pi_j) = 4 \sum_{\substack{\ell<m \\ \ell,m \in I}} |p_{i,\ell m} - p_{j,\ell m}|. \tag{13}$$

The inequality includes $n = 0$ and $n = |I|$, when both visible sets agree, and remains valid after averaging over the training distribution of t. It measures disagreement between visible indices; each candidate supplies its own token values at these indices.

Unequal eligible sets. For distinct I_i and I_j, restrict the ordering comparison to $I_i \cap I_j$. Every common position pair has the same marginal and nested-event proof, so shared Gumbels minimize the restricted Kendall criterion while preserving the full candidate permutations. Proposition 2 states this result formally for response-relative indices and derives a length-corrected preference bound for the resulting unequal visible counts.

A.3 Preference losses and group weighting

Rank-gap-weighted all-pairs ranking. Let $\mathcal{P}_R = \{(i,j) : R_i > R_j\}$ and assign average ranks r_k in increasing reward order, with ties receiving equal ranks. For a nonempty $\mathcal{P}_R$, set $\omega_{ij} = (r_i - r_j)/\sum_{(u,v)\in\mathcal{P}_R}(r_u - r_v)$. The protein objective is

$$\Phi_R^{\text{pair}}(a) = \sum_{(i,j)\in\mathcal{P}_R} \omega_{ij} \log(1 + e^{-\beta(a_i - a_j)}), \qquad \beta > 0. \tag{14}$$

For an empty strict-pair set, define $\Phi_R^{\text{pair}} = 0$. Equation 8 averages group losses equally. The logistic penalty follows DPO and diffusion preference optimization (Rafailov et al., 2023; Wallace et al., 2024); each candidate score is reused across its eligible comparisons. With e_k the kth coordinate vector, the zero-score derivative is $\alpha = -\frac{\beta}{2}\sum_{(i,j)\in\mathcal{P}_R} \omega_{ij}(e_i - e_j)$.

Reward-distribution matching. Following the normalized reward-matching form of Han et al. (2026), the DNA and language objectives use $q^R = \text{softmax}(R)$ and

$$\Phi_R^{\text{list}}(a) = \text{KL}\big(q^R \,\big\|\, \text{softmax}(\beta a)\big)\,. \tag{15}$$

Its zero-score derivative is $\alpha = \beta(\mathbf{1}/K - q^R)$. Both losses are convex and common-shift invariant, as required by Equation 10; Equation 2 specifies scoring support and time weights.

Application and scaling. Protein uses Equation 14; DNA and language use Equation 15. In both equations, a denotes the unscaled conditional score from Equation 2; implementations that pre-scale scores by β apply no additional scale inside the loss. Convexity is with respect to the score vector, and the listwise KL places the fixed reward distribution in its first argument.

A.4 Complete-group gradients and the diagnostic reference

Let ξ contain the progress, count, and order random variables. At fixed priorities, $\mathcal{T}_k$ maps a completed candidate and ξ to its context, giving $A_k^{\mathcal{T}}(\theta,\xi) = A_k(\theta;\mathcal{T}_k(\xi))$. Under differentiation through the conditional scores, the group gradient is $g_\Pi = \sum_k \partial_k \Phi_R(A^{\mathcal{T}})\nabla_\theta A_k^{\mathcal{T}}$. The context-averaged reference is

$$\bar{g}_\theta = \nabla_\theta \mathcal{L}_{\text{avg}}(\theta) = \sum_k \partial_k \Phi_R(\bar{A}^{\mathcal{T}})\, \mathbb{E}\nabla_\theta A_k^{\mathcal{T}}(\theta,\xi), \qquad \bar{A}_k^{\mathcal{T}} = \mathbb{E}A_k^{\mathcal{T}}(\theta,\xi). \tag{16}$$

The identity assumes integrable gradients and interchange of differentiation and expectation. It first averages each candidate's conditional evidence and then forms the preference comparison. Averaging Equation 16 over groups gives the collection-level reference used in Appendix D.5.

At matched deterministic policy/reference parameters $\theta_0 = \bar{\theta}$, all relative scores vanish. Set $\phi_k = \nabla_\theta A_k^{\mathcal{T}}(\theta_0,\xi)$, $\alpha_k = \partial_k\Phi_R(0)$, and $C_{ij} = \text{Cov}(\phi_i,\phi_j)$ under shared random inputs. With finite gradient second moments, both couplings have mean $\sum_k \alpha_k \mathbb{E}\phi_k$, and

$$\text{Cov}(g_{\text{shared}}) - \text{Cov}(g_{\text{ind}}) = \sum_{i\neq j} \alpha_i\alpha_j C_{ij}. \tag{17}$$

Expanding the covariance of $\sum_k \alpha_k\phi_k$ proves this identity: fixed marginals preserve its diagonal terms, and sharing supplies the off-diagonal covariances. Their effect depends on the preference coefficients and candidate responses. At general checkpoints, the coefficients also depend on the sampled scores. Complete-gradient measurements evaluate this aggregate coupling response.

A.5 CANDIDATE-LOCAL CONTEXT DISTRIBUTION

Fix candidate k, progress t, and priorities η_k. Write $n_k = |I_k|$ and $\mathrm{pref}_r(\pi) = \{\pi(1), \dots, \pi(r)\}$. Equation 5 gives a $\mathrm{Binomial}(n_k, t)$ count independent of the PL permutation. For $S \subseteq I_k$,

$$q_{k,t}(S) = \binom{n_k}{|S|} t^{|S|}(1-t)^{n_k-|S|} \sum_{\pi:\, \mathrm{pref}_{|S|}(\pi)=S} P_k(\pi). \tag{18}$$

The event that the output is S requires count $|S|$ and the indicated unordered prefix. Their independence gives the formula. For a fixed count, prefix events partition the permutations, establishing normalization. Copying y_k on S and averaging over $t \sim \nu$ defines q_k. Shared random arrays leave each candidate's input law unchanged, preserving these marginals under every control.

Finite priorities and $0 < t < 1$ give every eligible subset positive mass. When $\zeta = 0$, a size-r prefix has probability $\binom{n_k}{r}^{-1}$; the binomial coefficient cancels and the absorbing-mask marginal is recovered. The sampler evaluates no subset probabilities explicitly. For unequal eligible sets, uniforms drawn on their union restrict to the same independent-input law within each candidate.

A.6 FROM SCORE RESPONSE TO A COUPLING OBJECTIVE

Fix the model parameters, deterministic scoring procedure, candidate group, score-weight rules, and finite priorities. First assume a common eligible set I of size n. Draw $t \sim \nu$ and $N \mid t \sim \mathrm{Binomial}(n, t)$. Independently draw the permutation vector from $\Lambda \in \mathcal{C}(P_1, \dots, P_K)$, and put $S_k = \mathrm{pref}_N(\pi_k)$. Thus every comparison here has shared progress and count, and only the order coupling varies. This is the theoretical setting of the private-order versus full-sharing control.

Score-response assumption. Let $a_k(S, t)$ be Equation 2 evaluated with candidate k's tokens visible on S, using its prescribed candidate-local score weights. Suppose all involved scores have finite second moments and there are finite constants L_k such that

$$|a_k(S,t) - a_k(S',t)| \le L_k |S \triangle S'| \quad \text{whenever } |S| = |S'|, \quad \text{for } \nu\text{-almost every } t. \tag{19}$$

The constants bound the response of the entire score, including any mask-dependent scoring support. Define $D_{ij}(\Lambda) = \mathbb{E}_\Lambda d_{\mathrm{K}}(\pi_i, \pi_j)$ and

$$V_{ij} = \mathrm{Var}_{t,N,\pi_i \sim P_i} \left[a_i(\mathrm{pref}_N(\pi_i), t) - a_j(\mathrm{pref}_N(\pi_i), t) \right]. \tag{20}$$

Here candidate j uses the same visible indices as i but supplies its own token values. This common-index evaluation defines the reference term of the proof. Its distribution depends only on P_i and the fixed stage/count law; V_{ij} therefore does not change with Λ.

Bounding the actual pairwise margin. Write $M_{ij} = a_i(S_i, t) - a_j(S_j, t)$, and split it into $M^0_{ij} = a_i(S_i, t) - a_j(S_i, t)$ and $E_{ij} = a_j(S_i, t) - a_j(S_j, t)$. The equal-prefix inequality proved in Appendix A.2 is pathwise. Consequently,

$$\sqrt{\mathrm{Var}_\Lambda M_{ij}} \le \sqrt{V_{ij}} + \sqrt{\mathbb{E}_\Lambda E_{ij}^2} \le \sqrt{V_{ij}} + 2L_j \sqrt{D_{ij}(\Lambda)}. \tag{21}$$

The first inequality applies the triangle inequality in centered L^2 and uses $\mathrm{Var}(E_{ij}) \le \mathbb{E}E_{ij}^2$. The second uses Equation 19 and $|S_i \triangle S_j|^2 \le 4 d_{\mathrm{K}}(\pi_i, \pi_j)$. No zero-mean assumption on E_{ij} is needed. Fixed score marginals give $\mathbb{E}M_{ij} = \bar{A}_i - \bar{A}_j$ for every coupling.

For $\ell_\beta(d) = \log(1 + e^{-\beta d})$, convexity and $0 \le \ell''_\beta \le \beta^2/4$ imply $0 \le \mathbb{E}\ell_\beta(M_{ij}) - \ell_\beta(\mathbb{E}M_{ij}) \le (\beta^2/8)\,\mathrm{Var}(M_{ij})$. This follows by integrating the second derivative between M_{ij} and its mean; the linear term has expectation zero. Substitution of Equation 21 and summation with the fixed nonnegative pair weights prove the inequality in Corollary 1. An empty strict-pair set has zero loss and satisfies the bound trivially.

Why this selects shared Gumbels. All terms in $\mathcal{U}(\Lambda)$ apart from $D_{ij}(\Lambda)$ depend only on fixed candidate marginals, scoring responses, and rewards. Each term is nondecreasing in its corresponding

D_{ij}. Proposition 1 supplies a single feasible coupling attaining the minimum of every D_{ij} simultaneously, including all reward-ordered pairs. This proves the minimization statement in Corollary 1:

$$\Lambda_{\mathrm{Gumbel}} \in \operatorname*{arg\,min}_{\Lambda \in \mathcal{C}(P_1,\dots,P_K)} \mathcal{U}(\Lambda). \tag{22}$$

This result selects an order coupling with fixed candidate PL marginals and a fixed shared stage/count process. It holds for every admissible set of response constants and anchor variances, so selecting the minimizer requires no estimates of these quantities. Numerical tightness depends on their magnitudes and on the intermediate inequalities. The experiments report losses, score fluctuations, and gradient alignment; they do not estimate L_j, V_{ij}, or the gap-to-bound ratio. Minimization of this upper bound is distinct from a universal reduction of the actual loss or gradient variance.

Reward-distribution matching. For $\Phi_R^{\mathrm{list}}(A) = \mathrm{KL}(q^R \| \operatorname{softmax}(\beta A))$, the common-shift-invariant Hessian has operator norm at most $\beta^2/2$. Appendix A.1 and Equation 21 give

$$0 \le \Delta_\Lambda^{\mathrm{list}} \le \frac{\beta^2}{4K} \sum_{i<j} \left(\sqrt{V_{ij}} + 2L_j \sqrt{D_{ij}(\Lambda)} \right)^2. \tag{23}$$

The same shared coupling minimizes this upper bound for the fixed-target listwise loss used in DNA and language. The equal-set result applies directly to fixed-length DNA groups. For variable-length responses, Proposition 2 gives the corresponding bound with a length-mismatch term, using the same response-relative count and order arrays.

Additional scoring randomness. The same argument applies to scores $a_k(S, t, \varepsilon_k)$ when the scorer-noise vector ε has a fixed joint law independent of progress, counts, and orders. Assume the bound in Equation 19 holds with the same finite L_k for almost every ε_k, and all scores have finite second moments. Include ε in the variance defining V_{ij}. In the margin decomposition, evaluate both occurrences of candidate j using the same ε_j. The centered-L^2 argument then holds unchanged, and the noise-averaged V_{ij} remains independent of Λ. Both preference bounds and their shared-Gumbel minimizers follow. This extension requires these response and independence conditions; the deterministic fixed-checkpoint setting is the special case of constant ε.

Connection to the implemented parameter update. At optimizer step s, condition on the history, candidate group, rewards, and detached priorities computed before the fresh context draws. These quantities induce fixed marginals $q_{s,k}$ and a joint law Π_s. Under integrable gradients and interchange of differentiation and expectation, the raw gradient of Equation 8, before clipping or optimizer transformations, is unbiased for $\nabla_\theta \mathbb{E}_{\Pi_s} \Phi_R(A(\theta))$ with Π_s held fixed. To distinguish this partial derivative from adaptive-law differentiation, write $F(\vartheta, \psi) = \mathbb{E}_{Z \sim \Pi_\psi} \Phi_R(A(\vartheta; Z))$, where ψ determines detached priorities and ϑ determines the scored policy. The conditional expected gradient at step s is $\nabla_1 F(\theta_s, \theta_s)$. Wherever both partial derivatives exist, the total derivative of $F(\theta, \theta)$ is $\nabla_1 F(\theta, \theta) + \nabla_2 F(\theta, \theta)$; the second term, from changing the sampling law, is omitted by stop-gradient. The conditional guarantees thus apply to each frozen-law update and supply no stationary-objective or convergence result for the adaptive sequence. Appendix A.4 defines the separate context-averaged gradient reference used in the experiments.

A.7 Response-relative coupling for unequal lengths

Language candidates have response-relative eligible sets $I_k = \{1, \dots, n_k\}$, with $n_k \ge 1$. Fix their finite priorities and full PL laws P_k. Draw the count array on $\{1, \dots, \max_k n_k\}$ as in Equation 5; its joint distribution with progress t is held fixed. Vary only the permutation coupling $\Lambda \in \mathcal{C}(P_1, \dots, P_K)$, independently of these draws. Write $S_k = \mathrm{pref}_{N_k(t)}(\pi_k)$ and let $\pi|_C$ denote the relative order of the indices in C. For an ordered pair (i, j), let $(s, h) = (i, j)$ when $n_i \le n_j$, and $(s, h) = (j, i)$ otherwise. Thus $I_s \subseteq I_h$, with $m_{ij} = n_h - n_s$ unmatched positions and $C_{ij} = I_s$. Define

$$D_{ij}^{\cap}(\Lambda) = \mathbb{E}_\Lambda d_{\mathrm{K}}(\pi_i|_{C_{ij}}, \pi_j|_{C_{ij}}). \tag{24}$$

This compares common-position orders while retaining the full candidate permutation laws.

For the loss bound, retain the equal-cardinality score-response condition in Equation 19, separately on each candidate's own eligible set. Define an equal-count context for the longer candidate by

$T_{h\leftarrow s} = S_s \cup E_{sh}(t)$, where $E_{sh}(t) = \{\ell \in I_h \setminus I_s : u_\ell^{\mathrm{cnt}} \le t\}$. Its cardinality is $N_h(t)$, since $|E_{sh}(t)| = N_h(t) - N_s(t)$. Set $V_{ij}^{\mathrm{len}} = \mathrm{Var}[a_s(S_s, t) - a_h(T_{h\leftarrow s}, t)]$, with each score using its own candidate's tokens. This variance depends on the shorter candidate's marginal and the fixed count process, so it is constant as Λ varies. Assume it and all score second moments are finite.

Proposition 2 (Unequal-length order and preference guarantees)**.** *Under the setup above, shared Gumbels preserve every full candidate context law q_k and simultaneously attain*

$$\min_{\Lambda\in\mathcal{C}(P_1,\dots,P_K)} D_{ij}^{\cap}(\Lambda) = \sum_{\substack{\ell<r\\ \ell,r\in C_{ij}}} |\sigma(\eta_{i\ell} - \eta_{ir}) - \sigma(\eta_{j\ell} - \eta_{jr})| \quad \textit{for every } i \neq j. \tag{25}$$

Define $B_{ij}(\Lambda) = \sqrt{V_{ij}^{\mathrm{len}}} + 2L_h\big(\sqrt{D_{ij}^{\cap}(\Lambda)} + m_{ij}\big)$. The pairwise logistic ranking and listwise reward-distribution matching losses in Appendix A.3 satisfy

$$\begin{aligned} 0 \le \Delta_\Lambda^{\mathrm{pair}} &\le \mathcal{U}_{\mathrm{len}}^{\mathrm{pair}}(\Lambda) := \frac{\beta^2}{8} \sum_{(i,j)\in\mathcal{P}_R} \omega_{ij} B_{ij}(\Lambda)^2, \\ 0 \le \Delta_\Lambda^{\mathrm{list}} &\le \mathcal{U}_{\mathrm{len}}^{\mathrm{list}}(\Lambda) := \frac{\beta^2}{4K} \sum_{i<j} B_{ij}(\Lambda)^2. \end{aligned} \tag{26}$$

Shared Gumbels minimize both upper bounds over the specified class of order couplings.

Proof. Marginals and overlap orders. Each candidate reads independent Gumbel coordinates on its own I_k, giving its full PL law, independently of its binomial count. Shared inputs thus preserve q_k from Equation 18. For each pair of positions in C_{ij}, marginal order probabilities remain logistic in their priority difference. The Bernoulli lower bound and nested-event construction from Appendix A.2 apply to every such pair simultaneously. The complete permutations remain feasible on their original eligible sets, proving Equation 25. The shared count construction also gives $N_h(t) - N_s(t) \mid t \sim \mathrm{Binomial}(m_{ij}, t)$, allowing different realized reveal counts.

Visible contexts at unequal lengths. Put $Q = S_h \cap C_{ij}$ and $e = |S_h \setminus C_{ij}|$. Then Q is a prefix of $\pi_h|_{C_{ij}}$ of length $N_h - e$, whereas S_s is a prefix of length N_s. With $u = |S_s \setminus Q|$ and $v = |Q \setminus S_s|$, all uv cross pairs reverse order, so $uv \le d_{\mathrm{K}}(\pi_s, \pi_h|_{C_{ij}})$. Since $u - v = N_s - N_h + e$ and both $N_h - N_s$ and e belong to $\{0, \dots, m_{ij}\}$, we obtain $|S_s \triangle Q| \le m_{ij} + 2\sqrt{d_{\mathrm{K}}(\pi_s, \pi_h|_{C_{ij}})}$. The exclusive-position sets $E_{sh}(t)$ and $S_h \setminus C_{ij}$ differ at at most m_{ij} indices. Hence

$$|T_{h\leftarrow s} \triangle S_h| \le 2\sqrt{d_{\mathrm{K}}(\pi_s, \pi_h|_{C_{ij}})} + 2m_{ij}, \qquad |T_{h\leftarrow s}| = |S_h| = N_h. \tag{27}$$

Apply the response bound to the longer candidate's equal-sized contexts. Decompose its margin as

$$a_s(S_s, t) - a_h(S_h, t) = \big[a_s(S_s, t) - a_h(T_{h\leftarrow s}, t)\big] + \big[a_h(T_{h\leftarrow s}, t) - a_h(S_h, t)\big]$$

and apply the centered-L^2 triangle inequality to obtain $\sqrt{\mathrm{Var}(A_i - A_j)} \le B_{ij}(\Lambda)$. The sign of the margin is immaterial to its variance. Logistic curvature gives the first bound in Equation 26; the listwise contrast bound from Appendix A.1 gives the second.

Optimality of the bounds. The anchor variances, response constants, lengths, and reward weights are fixed. Both bounds are nondecreasing in every $D_{ij}^{\cap}(\Lambda)$. Equation 25 supplies one feasible coupling attaining all their minima, proving the claim. □

For equal lengths, $m_{ij} = 0$, $T_{j\leftarrow i} = S_i$, and these bounds reduce to Corollary 1 and Equation 23. At unequal lengths, the extra term accounts for unmatched response positions; the optimized term remains ordering disagreement on the shared index range. The contexts $T_{h\leftarrow s}$ and variances are proof quantities. Algorithm 1 retains every response at its original length and uses its existing count and order arrays. The scoring-noise extension in Appendix A.6 carries over by including the fixed noise law in V_{ij}^{len} under the same response and independence conditions.

B IMPLEMENTATION OF THE CONTEXT CONSTRUCTION

B.1 CONFIDENCE AND ORDERED MASKING

At each optimizer step, evaluate the current policy on $(y_k, 1, c)$ with dropout, input augmentation, and gradients disabled. Equation 4 uses raw non-mask probabilities before visible-token copying. For a nonempty eligible set, the normalization statistics are $\bar{\gamma}_k = |I_k|^{-1} \sum_{\ell \in I_k} \gamma_{k\ell}$ and $s_k^2 = |I_k|^{-1} \sum_{\ell \in I_k} (\gamma_{k\ell} - \bar{\gamma}_k)^2$. For an empty set, define $\mathrm{StdClip}(\gamma; \varnothing) = \varnothing$ and $\mathrm{Top}_0(\varnothing) = \varnothing$. Open-interval uniforms avoid the endpoints of the Gumbel transform.

Equivalent reveal times. Let $r_{k\ell}$ be the rank of position ℓ in the perturbed priority order, and $u^{\mathrm{cnt}}_{(r;I_k)}$ the rth smallest count uniform on I_k. Set $w_{k\ell} = u^{\mathrm{cnt}}_{(r_{k\ell};I_k)}$. The context can be written as

$$x_{k\ell} = \begin{cases} y_{k\ell}, & w_{k\ell} \le t, \\ \mathtt{m}, & w_{k\ell} > t, \end{cases} \qquad \ell \in I_k. \tag{28}$$

Exactly $N_k(t)$ count uniforms lie below t, proving equivalence with Equation 5. Prompt or other conditioning tokens remain visible, and padding is excluded.

Uniform priorities. With $\zeta = 0$, the Gumbel permutation is uniform and independent of $N_k(t) \sim \mathrm{Binomial}(|I_k|, t)$. Conditional on $N_k(t) = n$, every size-n subset is equally likely. A particular $S \subseteq I_k$ has probability $t^{|S|}(1-t)^{|I_k|-|S|}$, recovering the standard absorbing-mask marginal. Changing the sign or magnitude of ζ changes the subset priorities while preserving this count law.

B.2 SHARED INPUTS AND SCORE EVALUATION

For the layered controls, candidate priorities and token values are fixed. The `None` control uses candidate-private progress, count uniforms, and Gumbels. P shares progress while count and order arrays remain candidate-private. P+C additionally shares count uniforms, retaining private Gumbels. P+C+O shares the Gumbels as well. P aligns reveal progress; adding C matches counts for equal eligible sets. P+C versus P+C+O isolates order alignment at that shared budget. These constructions give every candidate identically distributed random inputs to its fixed local map, preserving its context and score marginals at a fixed model. Groups use independent random draws. For unequal response lengths n_k, use response-relative indices $I_k = \{1, \dots, n_k\}$ and draw shared arrays on $\{1, \dots, \max_k n_k\}$. Candidate k reads the first n_k entries; prompt and padding positions remain outside I_k. Counts share uniforms and need not have equal realizations. Proposition 2 analyzes this indexing rule without changing candidate lengths.

Scoring support. For masked and visible eligible sets M_k, V_k, weights can be written as $v_{k\ell} = b_{k\ell}(\mathbf{1}\{\ell \in M_k\} + \lambda \mathbf{1}\{\ell \in V_k\})$, where $b_{k\ell}$ contains the specified normalization and time weights. Masked-only scoring uses $\lambda = 0$. All-valid scoring includes visible positions with the weight specified for each training stage. Masked averages use denominator $\max(1, |M_k|)$, giving zero score for an empty masked set. Policy and reference receive the same context and support; raw probabilities are used for visible reconstruction terms. Contexts, priorities, weights, and reference parameters are fixed during differentiation. The next step recomputes priorities from the updated policy. Each candidate needs one detached policy evaluation and one sort; comparisons reuse its score.

C EXPERIMENTAL DETAILS

This appendix gives the implementation and evaluation details for the three main experiments. Unless stated otherwise, the ablations in Appendix D use the configurations below. DNA and protein results report the mean and sample standard deviation across training seeds; language results report percentages from one training run.

C.1 DNA SEQUENCE DESIGN

Setup. We use the released 200-bp masked-diffusion DNA model and the GOSAI enhancer candidate pool. The denoiser has four convolutional stacks, 128 hidden channels, and no dropout. The

candidate pool is divided into preference groups of $K = 500$. The frozen copy of the initialization is the reference model; the trainable copy is updated with AdamW (learning rate 1.48×10^{-5}, zero weight decay, gradient accumulation 2, and 50 optimizer updates). The listwise scale is $\beta = 0.2$.

At each training update, the context map evaluates every completed candidate once with the current policy, extracts the raw non-mask maximum probability at each eligible position, population-standardizes and clips these values to $[-2, 2]$, and multiplies them by $\zeta = -7.5$. Thus lower-confidence positions receive larger reveal priorities. One progress draw in $[0.01, 0.99]$, one array of count uniforms, and one array of ordering uniforms are shared by the candidates in a group; priorities and nucleotide values remain candidate-local. Gumbel top-N selects the visible positions.

The DNA score A_k is the policy–reference log-probability difference on eligible masked (unrevealed) positions, averaged over that support as in Equation 2. The alignment loss is $\mathrm{KL}(\mathrm{softmax}(R) \| \mathrm{softmax}(\beta A))$ from Equation 15: the reward distribution is the first argument, and β scales the score once. This loss is convex in A. The candidate reward follows the HepG2 activity definition used by SDPO (Han et al., 2026).

Pretrained and post-training baselines use released SDPO/DRAKES data and checkpoints; test-time guidance uses the published samplers. GILC-DB uses 128 reverse steps, guidance scale 11,000, 10 Monte Carlo proposals, Gumbel temperature 1, and 640 unfiltered sequences per sampling seed.

Evaluation. Each checkpoint uses the native DNA decoder to produce ten batches of 64 unfiltered sequences with 128 denoising steps. Generation and likelihood estimation use seed 2001. For each generated set, Pred-Activity is the median HepG2 activity-oracle output. ATAC-Acc is 100 times the fraction whose binary accessibility-oracle output exceeds 0.5. The 3-mer correlation is the Pearson correlation between generated and top-0.1%-activity 3-mer count vectors. JASPAR Corr is the Spearman correlation between their motif-frequency vectors. App-Log-Lik is the median of the absorbing-state ELBO estimate returned by the pretrained denoiser's 128-step likelihood routine. The sum of position-wise nucleotide entropies gives the reported diversity metric:

$$-\sum_{\ell=1}^{200} \sum_{a \in \{A,C,G,T\}} \hat{p}_\ell(a) \log_2 \hat{p}_\ell(a),$$

where $\hat{p}_\ell(a)$ is the empirical base frequency at position ℓ in the generated set. The six metrics are first computed for each training seed and then summarized across seeds. Correlations compare sequence statistics; activity and accessibility directly evaluate the task oracles.

C.2 Protein inverse folding

Setup. We initialize the trainable and frozen reference models from the released FMIF inverse-folding checkpoint. The denoiser is a 128-dimensional ProteinMPNN with three encoder and three decoder layers, 30 neighbors, dropout 0.1, backbone augmentation 0.1, and maximum length 75. The training set contains 438,540 reward-labeled candidates associated with 311 Megascale backbones. A minibatch packs up to 200 rows into eight groups of 25 candidates, grouped by backbone.

For a candidate, the training reward is $dG_{\mathrm{ML}} - dG_{\mathrm{ML,wt}}$. The training context map evaluates each completed candidate with the current policy and uses raw non-mask confidence, population standardization with $\epsilon = 10^{-8}$, clipping at 2, and $\zeta = -7.5$. Every group shares its progress, count uniforms, and ordering uniforms, while each candidate retains its own priorities. Candidate scores are the $\beta = 0.15$-scaled policy–reference log-probability differences on valid residues; visible residues have weight 1 in the first stage and 0.85 in the continuation stage. The preference loss is the rank-gap-weighted all-pairs softplus loss, with strict reward-ordered pairs and equal weighting of groups. Training uses nine first-stage epochs at learning rate 1.61×10^{-5}, followed by a two-epoch optimizer-reset continuation at 4×10^{-6}. The scoring reference is frozen. Priorities are recomputed from the current policy at each update and held fixed during differentiation.

Generation uses the native FMIF decoder with 50 denoising steps at temperature 0.1. The DRAKES and SDPO baselines use their released checkpoints and samplers. GILC-DB uses frozen FMIF and guidance oracles, 50 transitions, temperature 0.1, guidance scale 1000, 20 Monte Carlo proposals per step, and Gumbel temperature 1.

Evaluation. For each test backbone, we draw 128 sequences per generation repeat, using three repeats for reward metrics and two for structural metrics. Pred-ddG is the median output of the fixed reward oracle, and Positive is 100 times the fraction with Pred-ddG > 0. For structural scoring, ESMFold predicts a structure for each generated sequence; both the native backbone and the prediction are relaxed with the fixed PyRosetta procedure, and scRMSD is their backbone RMSD in Å. Structure is the percentage with scRMSD < 2 Å. Joint success is

$$100\,\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\text{Pred-ddG}_i > 0\}\mathbf{1}\{\text{scRMSD}_i < 2\},$$

computed on the paired structural samples. Positional entropy is the sum of per-residue amino-acid entropies for each backbone, summarized by the median across backbones. Unlike either marginal success rate, joint success requires both stability and backbone consistency.

C.3 Language modeling

Setup. We fine-tune LLaDA-8B-Instruct on 59,876 UltraFeedback prompts, retaining the highest- and lowest-reward response for each prompt while keeping prompt tokens visible. Maximum sequence and prompt lengths are 2048 and 1800. Training uses a global prompt batch size of 256, per-device batch size 2, gradient accumulation 16, two epochs, AdamW, cosine decay, warmup fraction 0.1, and seed 42. The preference scale is $\beta = 1$.

The completed response is evaluated once by the current policy. The MASK logit is excluded before forming the raw non-mask maximum probability; prompt and padding positions are ineligible. Candidate-local standardized priorities use scale -0.5, clipping 2, and $\epsilon = 10^{-8}$, so low-confidence response tokens are preferred. Each prompt group shares one progress draw and count/order arrays indexed by response-relative position. Response k uses entries $1, \ldots, n_k$, preserving its original length and binomial count law; Proposition 2 gives the corresponding order and preference-bound guarantees. For masked response positions M_k, the score uses the mean log-probability difference over M_k, with denominator $\max(1, |M_k|)$; visible prompt tokens never enter the target. With A_k denoting the policy–reference score, the loss is

$$\mathrm{KL}(\mathrm{softmax}(R) \parallel \mathrm{softmax}(\beta A))\,,$$

where R denotes rewards; the reference is frozen, with no SFT, length, or reference-KL penalty.

We reproduce the fine-tuning baselines on the same UltraFeedback data and evaluate every row with the common tokenizer, sequence limits, generation settings, and evaluation pipeline. Baseline learning rates follow their implementations; C3 uses 6×10^{-7} and SDPO 10^{-6}.

Evaluation. Generation uses the native LLaDA decoder with temperature and guidance scale zero and seed 42. The main table evaluates AlpacaEval on 805 prompts, HumanEval on 164 problems, sanitized MBPP on 257 problems, and GSM8K on 1,319 problems. GSM8K uses a 256-token/step budget with block length 8 and four demonstrations; AlpacaEval uses 256 tokens and block length 32; HumanEval uses 512 tokens and block length 512 with EOS-logit suppression; sanitized MBPP uses 256 tokens and block length 256 with EOS/EOT prioritized in its evaluator. Failed generations count as failures, and all benchmarks use fixed extraction rules.

Alpaca WR is the fraction of pairwise judge comparisons won by the model, and LC-WR is the same comparison after the official length-control correction. For each prompt, an external judge compares the model response with a reference response; we use GPT-4.1 for these pairwise judgments, and AlpacaEval aggregates them into WR and LC-WR. HumanEval and MBPP are pass@1, the fraction of generated programs passing the task tests; GSM8K is exact-answer accuracy. IFEval (Zhou et al., 2023) evaluates compliance with automatically verifiable instructions, complementing the pairwise response-quality judgments. Table 3 reports all results as percentages, with IFEval in the last column.

D Additional Results

Tables 4 and 5 are matched component series: within each series, the named priority or coupling component is varied while the remaining choices are held fixed. Their full/Low or aligned/Low rows are reference arms for the corresponding ablation series.

D.1 PROTEIN PRIORITIES AND COUPLING

The component table changes the priority strength/direction or the shared random inputs. "Low" denotes the negative-ζ rule, "High" reverses its sign, and "Uniform" removes confidence dependence. "Private orders" keeps progress and counts shared but gives each candidate its own order keys; "Independent" makes all three random inputs candidate-private. The Low/full row anchors this nine-epoch series. The main C^3 result includes two continuation epochs (Table 2; Appendix C.2).

Table 4: **Protein component ablations.** Parentheses give the standard deviation under the aggregation defined in Experimental Details. Pred-ddG and scRMSD are medians; Positive, Structure, and Joint are percentages; Entropy is in bits.

Configuration	Pred-ddG (med) ↑	Positive (%) ↑	scRMSD (med) ↓	Structure (%) ↑	Joint (%) ↑	Entropy (bits) ↑
Low / full, -7.5	1.6417 (0.0031)	90.69 (0.45)	1.0597 (0.0089)	86.36 (0.79)	78.51 (1.04)	43.99 (0.78)
Uniform / full, 0	1.5192 (0.0255)	91.75 (0.43)	1.0587 (0.0035)	81.73 (1.42)	76.40 (1.25)	48.20 (0.58)
High / full, $+7.5$	1.6106 (0.0149)	95.72 (0.22)	1.4992 (0.0406)	62.06 (0.91)	60.42 (0.99)	49.01 (0.44)
Low / full, -3.75	1.6121 (0.0456)	90.44 (0.79)	0.9964 (0.0096)	87.07 (0.10)	78.14 (1.08)	35.74 (0.67)
Low / private orders	1.6534 (0.0254)	88.04 (0.20)	0.9724 (0.0057)	88.94 (0.02)	77.48 (0.37)	34.99 (0.62)
Low / independent	1.7179 (0.0246)	87.96 (0.23)	0.9217 (0.0058)	88.22 (0.66)	76.82 (0.84)	30.44 (0.09)

Changing the direction from Uniform to Low raises median Pred-ddG by 0.1225 and Joint by 2.11 percentage points. Reversing the direction to High collapses Structure by 24.31 points and Joint by 18.09 points, despite its higher Positive rate. These results favor negative priorities for the simultaneous reward–structure criterion. Reducing $|\zeta|$ to 3.75 gives a lower scRMSD but loses reward and entropy, showing a reward–structure–diversity trade-off within the negative direction.

Full coupling gives the best Positive, Joint, and Entropy among the three coupling choices. Private orders and Independent contexts obtain higher median Pred-ddG and lower scRMSD, but their Joint rates fall to 77.48% and 76.82%, and their entropies fall to 34.99 and 30.44 bits.

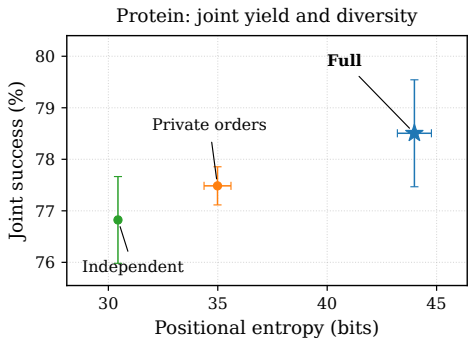


Figure 3: **Protein coupling ablations.** The axes show joint success and positional entropy for full sharing, private orders, and independent contexts.

D.2 DNA PRIORITY DIRECTION AND KEY CORRESPONDENCE

The DNA table changes only the priority direction (Low, High, or Uniform) and the correspondence of ordering keys (aligned, permuted, cyclic, or independent). Here, "Low" denotes the negative-ζ low-confidence-first rule, "High" denotes its sign-reversed counterpart, and "Uniform" removes confidence dependence. Candidate-specific permutations and cyclic shifts preserve each candidate's local state distribution while changing the cross-candidate dependence. The Low/aligned row is the reference arm for this matched DNA series. Priorities are recomputed from the current policy as in Section 4.2. The key-correspondence controls retain shared progress and count uniforms and vary only the ordering keys. Here, "independent" refers to candidate-private ordering keys.

Table 5: **DNA priority and key-coupling ablations.** Parentheses give the standard deviation under the aggregation defined in Experimental Details. Activity and App-Log-Lik are medians within each generated set; ATAC is a percentage.

Configuration	Pred-Activity (med) ↑	ATAC-Acc (%) ↑	3-mer Corr ↑	JASPAR Corr ↑	App-Log-Lik (med) ↑	Entropy (bits) ↑
Low / aligned	6.406 (0.020)	97.71 (0.36)	0.911 (0.005)	0.956 (0.002)	−249.09 (0.25)	355.75 (0.35)
High / aligned	6.423 (0.011)	96.41 (0.41)	0.866 (0.003)	0.935 (0.001)	−250.17 (0.21)	349.81 (0.42)
Uniform / aligned	6.404 (0.038)	97.76 (0.39)	0.886 (0.003)	0.942 (0.002)	−248.66 (0.48)	351.57 (0.54)
Low / permuted	6.384 (0.050)	97.45 (0.36)	0.905 (0.002)	0.956 (0.001)	−249.29 (0.85)	355.23 (0.62)
Low / cyclic	6.383 (0.013)	97.29 (0.36)	0.908 (0.003)	0.957 (0.001)	−248.94 (0.31)	355.39 (0.19)
Low / independent	6.367 (0.025)	97.29 (0.80)	0.910 (0.004)	0.956 (0.003)	−249.40 (0.56)	355.95 (0.85)
High / independent	6.415 (0.012)	95.16 (0.41)	0.861 (0.003)	0.937 (0.001)	−249.78 (0.31)	349.27 (0.01)

Aligned keys outperform the three key-scrambling controls at the same Low priority: activity exceeds the independent control by 0.0384, and ATAC-Acc by 0.42 points; permutation and cyclic shifts reduce activity by 0.0213 and 0.0229. Because these controls preserve candidate-local marginals, the gain isolates the cross-candidate correspondence predicted by the coupling construction. Among aligned arms, Low gives the strongest 3-mer correlation, JASPAR correlation, and positional entropy, with activity and ATAC-Acc close to Uniform (6.406 versus 6.404 and 97.71% versus 97.76%). High gives higher Pred-Activity but weaker accessibility and sequence-statistic results. The coupling benefit holds in both directions: aligned sharing raises activity by 0.0384 under Low and 0.0078 under High, and raises ATAC-Acc by 0.42 and 1.25 points, respectively.

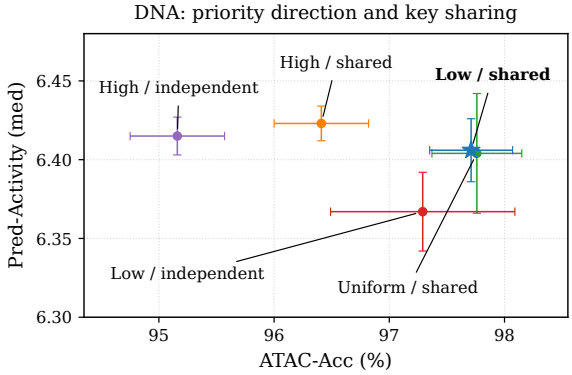


Figure 4: **DNA priority direction and key correspondence.** The axes report predicted activity and ATAC-Acc under the named priority and coupling controls.

D.3 Incremental sharing of random inputs

At the same protein checkpoint and batch, the controls in Appendix B.2 successively share progress (P), count uniforms (C), and order keys (O), holding candidate priorities and values fixed. Table 6 gives consecutive loss reductions of 0.0668, 0.0053, and 0.0113, or 15.4%, 1.4%, and 3.1% relative to the preceding configuration. Shared progress supplies the largest reduction in this sequence. Count sharing completes budget alignment for equal eligible sets, and order alignment further lowers loss from 0.3607 to 0.3494. The final comparison directly tests shared orders after fixing the budget process; every step preserves the candidate-local marginals. At fixed score marginals, the context-averaged loss is common to the controls, so a sampled-loss difference estimates the corresponding difference in comparison discrepancy. This tests objective behavior separately from numerical calibration of the bound in Corollary 1.

Table 6: **Incremental sharing of random inputs.** Only the indicated shared input is changed; values use the comparison-loss definition above.

Shared inputs	None	P	P+C	P+C+O
Comparison loss ↓	0.4328	0.3660	0.3607	0.3494

D.4 Fixed-marginal score comparisons

We compare the full P+C+O construction with the exact product of the same empirical candidate score marginals. Sharing reduces the score-difference variance by 94.67–98.14% and lowers the pairwise loss in every evaluated group, while the context-averaged comparison remains unchanged. Because the empirical marginal means are identical, the measured loss decrease equals the decrease in the empirical Jensen gap. This isolates the nonlinear comparison effect from changes to average candidate scores or rewards; it does not require estimating a response constant.

In three groups where the mean score difference already disagrees with the reward order over 59.00–68.19% of pair weight, sharing increases the fraction of nonpositive reward-ordered margins. The fixed-marginal comparison keeps these means unchanged while changing how often sampled margins reverse sign. Loss reduction and reward-order agreement capture different properties.

D.5 Complete-group gradient alignment

The gradient diagnostic compares the sampled group gradient with the reference obtained by averaging each candidate's conditional score before applying the preference loss. Full sharing (P+C+O) improves cosine alignment in every collection, from a mean of 0.41911 under independent inputs to 0.66504. The two independent reference halves have cosines between 0.94652 and 0.96463, indicating a stable reference direction in these measurements. Unlike the score-only diagnostic, this calculation includes every candidate-gradient contribution and every loss coefficient. It evaluates the observed learning direction alongside the separate score-variance diagnostic.

Table 7: **Complete-group gradient alignment.** Cosines are measured against the context-averaged gradient; the last column checks the reference estimate.

Collection	Independent	Shared	Reference halves
1	0.43153	0.65356	0.95236
2	0.42619	0.70715	0.96463
3	0.41251	0.63106	0.94652
4	0.42446	0.62420	0.94816
5	0.40087	0.70925	0.94955
Mean	0.41911	0.66504	0.95224

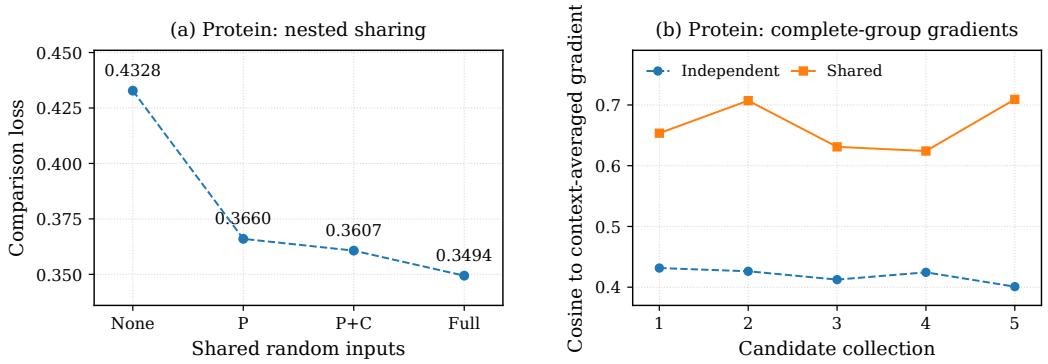


Figure 5: **Protein mechanism diagnostics.** (a) Comparison loss as progressively more random inputs are shared. (b) Alignment of the complete sampled gradient with the context-averaged reference.

D.6 Priority strength

Changing only the negative priority strength from $\zeta = -7.5$ to -3.75 changes median Pred-ddG from 1.6417 to 1.6121 and Joint from 78.505% to 78.136%, while reducing scRMSD from 1.0597 to 0.9964 Å and entropy from 43.99 to 35.74 bits. The two negative strengths retain higher joint success than the sign-reversed control, with different reward–structure–diversity trade-offs.