SONAR: SPECTRAL DIVERSITY FOR LOST REASONING MODES IN MASKED DIFFUSION LMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective exploration for complex reasoning problems remains a critical challenge in masked diffusion language models (MDLMs), where standard sampling frequently collapses into redundant failure modes. Raising the sampling temperature can increase diversity but also degrade solution quality, leaving the full reasoning capabilities of these models largely underutilized. To unlock these capabilities, existing approaches are primarily concerned with algorithms for token-reveal decisions, or promoting exploration over the vocabulary space. These methods, though effective to varying degrees, generally overlook the rich representation structure of the MDLMs. In this paper, we thus propose Sonar, a sampling method that first scans for overlap via a spectral objective in RKHS, subsequently transmitting the gradient of this objective throughout the model. Sonar then utilizes this gradient feedback to steer samples to systematically uncover lost reasoning modes of the MDLM. Across HumanEval (coding) and GSM8K (mathematics), Sonar improves fixed-budget reasoning success for LLaDA-8B-Instruct and Dream-7B-Instruct. It also expands solution coverage and reduces algorithmic collapse by uncovering distinct valid solution strategies missed by ordinary sampling, providing empirical evidence for a capability–sampling gap in MDLMs.

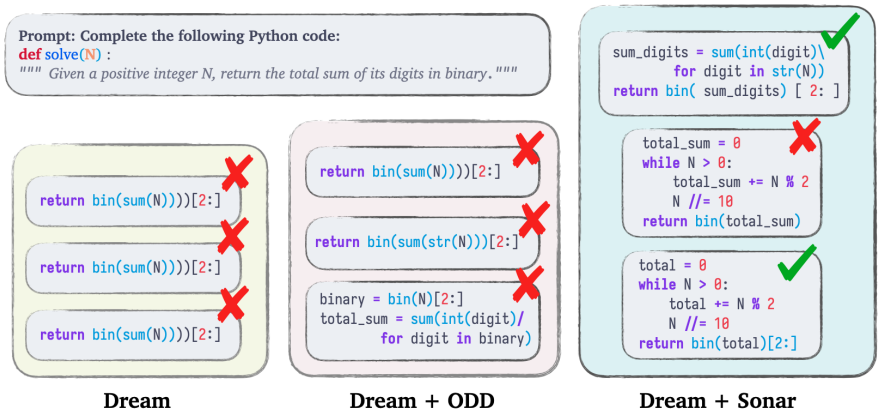

Figure 1: Standard MDLM repeatedly collapses to the same failure mode **(left)**. Vocabulary-space diversity cannot explore enough to solve the problem **(middle)**. Our representation-based diversity method, *Sonar*, explores new ideas that lead to multiple correct solutions **(right)**.

1 INTRODUCTION

The Inference Gap. Masked diffusion language models (MDLMs) (Nie et al., 2025) have recently emerged as a viable alternative to autoregressive language models, offering flexible and parallel generation based on bidirectional masked-token prediction. However, standard MDLM sampling can underexplore the model's solution space, with independent samples often collapsing to redundant reasoning modes (Lamont et al., 2026; Fang et al., 2026) (example in Figure 1). Again, reasoning can be improved without retraining simply by changing inference-time generation order or search strategies (Ni et al., 2026; Shen et al., 2026). Together, these findings suggest a *gap* between the reasoning capabilities supported by MDLMs, and what is exposed by its ordinary sampling dynamics.

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.

Existing approaches for improving DLM exploration primarily manipulate the extrinsic mechanics of DLM inference rather than exploiting the geometry of its contextual representation. For instance, Olausson et al. (2026) adjust sampling stochasticity, Shen et al. (2026) search over token and reveal-order decisions, and Wu et al. (2026) inject time-dependent perturbations into conditioning embeddings. The closest geometric approach for diversity is Orthogonal Diverse Diffusion (ODD) (Lamont et al., 2026), which couples candidates through an explicit geometric repulsion objective. However, ODD defines this geometry over aggregated vocabulary-level output features and logit-level interventions. In this work we instead use the learned geometry of the model's contextual representations as a basis for exploration, a signal overlooked by prior work.

We formulate a representation-based diversity objective to address mode collapse in MDLMs. Using the contextual representation allows us to capture semantic differences in evolving responses rather than lexical or vocabulary-level differences. Further, to measure the overlap between a group of candidates, we do not collapse each candidate into a single embedding. Instead, we retain all of the token representations per candidate, and compare them holistically through their spectral geometry. We define the diversity objective at the final block of the MDLM, immediately before the logits, while intervening at an earlier layer. The gradient of this objective acts as a steering signal that propagates through the model's dynamics, allowing the MDLM itself to transform the intervention into the desired final-layer diversity.

Our goal is therefore not to simply design a better sampling method, but to use structured representation-space exploration as a probe of behaviors that ordinary sampling may leave underutilized. We empirically evaluate this goal on coding (HumanEval (Chen et al., 2021)) and mathematics (GSM8K (Cobbe et al., 2021)) benchmarks, using two MDLMS: LLaDA-8B-Instruct (Nie et al., 2025) and Dream-7B-Instruct (Ye et al., 2025). Across both benchmarks, for both MDLMs, our method, Sonar, improves the empirical pass@K, the solution coverage of the distinct problems in the benchmarks, as well as the algorithmic novelty of the solutions as judged by external LLMs.

Summary of Contributions.

- **Representation Diversity Via Model-Mediated Steering.** We separate where diversity is measured from where it is imposed, turning the diversity gradient into a steering signal through the model's own dynamics. Our layerwise mechanistic analysis motivates this design.
- **Representations as Holistic Geometric Objects.** We retain all contextual token representations and compare candidates through their RKHS spectral overlap, rather than collapsing a response into a single embedding or comparing tokens independently.
- **Empirical Evidence for the Capability–Sampling Gap.** On HumanEval and GSM8K, using LLaDA and Dream, we show that representation-space exploration uncovers solutions missed by ordinary sampling. We quantify this through solved problems under a fixed sample budget, empirical Pass@K, and the novelty of generated solutions measured by an LLM judge.

2 PRELIMINARIES

2.1 SAMPLING IN MASKED DIFFUSION LANGUAGE MODELS

Given a prompt q, an MDLM generates a response by progressively unmasking token positions from an initial fully masked response. For candidate i, let $X^{i,t}$ denote its partial response at step t. The contextual representations over its response positions at block l are collected in $\mathbf{H}_l^{i,t} \in \mathbb{R}^{c\times d}$ where c is the number of response tokens and d is the hidden dimension. The representation at position j is denoted by $\mathbf{h}_{l,j}^{i,t} \in \mathbb{R}^d$. We use L to denote the final transformer block, so that $\mathbf{H}_L^{i,t}$ is the final hidden representation before the output head maps it to vocabulary logits. At each denoising step, the model predicts a categorical distribution over the vocabulary for every masked position, $x_j^{i,t} \sim p_{\theta,T}\big(\cdot \mid X^{i,t-1}, q\big)$ where $x_j^{i,t}$ denotes the token proposed for response position j, and T is the sampling temperature, which controls the sharpness of the predictive distribution and consequently the stochasticity of token sampling. The model's token distributions alone do not determine a generation trajectory; at each step, a sampler must determine which token proposals to draw and which masked positions to commit. We denote this ordinary MDLM sampling procedure

by A, with one-step transition

$$X^{i,t} \sim P_A\big(\cdot \mid X^{i,t-1}, q\big) \tag{1}$$

For instance, in LLaDA (Nie et al., 2025), the sampler ranks masked positions by confidence, commits a prescribed number of high-confidence predictions, and leaves the remaining positions masked for subsequent denoising steps. Repeated application of P_A therefore transforms the initially masked response into a complete sequence. To generate multiple responses, the same sampling procedure is applied independently to N candidates. Accordingly, their joint one-step transition factorizes as

$$P_A\big(X^{1,t}, \ldots, X^{N,t} \mid X^{1,t-1}, \ldots, X^{N,t-1}, q\big) = \prod_{i=1}^{N} P_A\big(X^{i,t} \mid X^{i,t-1}, q\big) \tag{2}$$

2.2 Underexploration and the Capability–Sampling Gap in MDLMs

Ordinary MDLM sampling has been shown to be restrictive: confidence-based commitment and generation-order choices can favor a narrow set of trajectories, while i.i.d. responses may collapse to similar reasoning modes (Tang et al., 2026; Hayakawa et al., 2026; Lamont et al., 2026; Fang et al., 2026). At the same time, several works show that modifying only the inference procedure, without changing the model parameters, can recover alternative reasoning paths and improve performance (Ni et al., 2026; Shen et al., 2026; Wu et al., 2026; Lamont et al., 2026). Together, these observations suggest a *capability–sampling gap* in MDLMs: useful reasoning behaviors may be supported by the trained model yet remain underexplored by its ordinary sampling dynamics.

The Problem With Temperature-Based Sampling. A natural way to address underexploration is to increase the sampling temperature T, making lower-probability token choices more likely and thereby broadening the trajectories explored by the sampler. However, this additional exploration is obtained by increasing stochasticity throughout generation, without explicitly controlling where that stochasticity is useful. Suppose a correct solution C requires reaching an unlikely intermediate choice E. Under sampler A,

$$P_A(C, E) = P_A(E)\, P_A(C \mid E) \tag{3}$$

Increasing temperature may raise $P_A(E)$, but the added stochasticity can simultaneously reduce $P_A(C \mid E)$ by disrupting the subsequent decisions needed to complete the reasoning path. This tension compounds when a solution requires several compatible decisions (Wang & Zhou, 2024; Troshin et al., 2025). In contrast, confidence-based sampling can achieve strong single-sample accuracy while exploring a narrower set of solutions (Nie et al., 2025; Olausson et al., 2026). Temperature therefore couples exploration with stochasticity: conservative sampling can repeatedly explore similar reasoning modes, while increasing temperature broadens exploration by injecting stochasticity throughout generation rather than explicitly controlling which trajectories are encouraged.

2.3 From Stochastic to Controlled Diversity

The trade-off above does not directly resolve the intermediate-choice problem in Eq. 3: if E remains unlikely, independent samples may repeatedly follow the same high-probability reasoning modes, while increasing T can also perturb the decisions required after reaching E. Instead, we retain the base sampler at a temperature chosen for reliable generation (Nie et al., 2025; Olausson et al., 2026) and explicitly explore alternatives around the trajectories it naturally visits. Rather than relying on randomness to differentiate K independent samples, we couple the candidates and reduce their overlap before continuing with the same sampler. Given the population at step $t-1$, we construct diversified states

$$\big(Z^{1,t-1}, \ldots, Z^{N,t-1}\big) = G_\tau\big(X^{1,t-1}, \ldots, X^{N,t-1}\big) \quad \text{s.t.} \quad \mathcal{L}_{\text{overlap}}\big(Z^{1,t-1}, \ldots, Z^{N,t-1}\big) \le \tau \tag{4}$$

Here, G_τ is a population-dependent diversification map that discourages candidates from following overlapping trajectories. The original sampler A then continues from these diversified states,

$$Q_\tau\big(X^{1,t}, \ldots, X^{N,t} \mid X^{1,t-1}, \ldots, X^{N,t-1}, q\big) = \prod_{i=1}^{N} P_A\big(X^{i,t} \mid Z^{i,t-1}, q\big) \tag{5}$$

This separates diversity from sampling stochasticity. In Eq. 3, increasing temperature changes both the probability of reaching an alternative, $P_A(E)$, and the probability of completing it correctly,

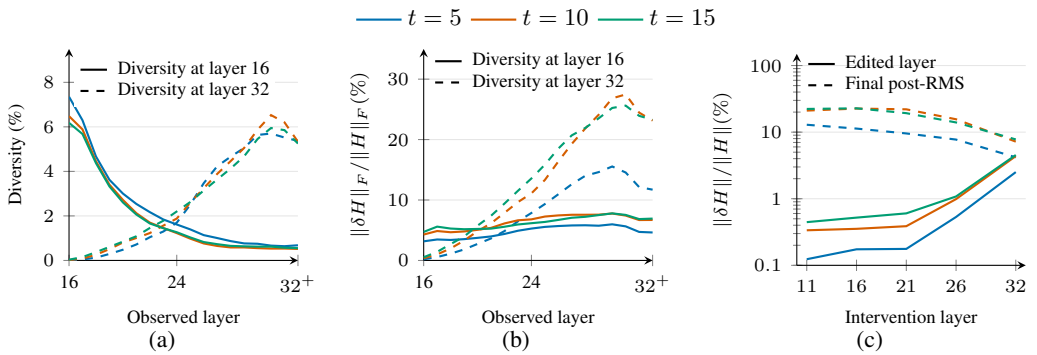

Figure 2: Layerwise analysis averaged over selected HumanEval problems for $K = 16$; colors denote denoising steps. In (a,b), intervention is fixed at layer 16: solid lines optimize diversity locally, while dashed lines optimize final post-RMSNorm. In (c), the intervention depth varies under a matched 5% reduction in final-layer overlap; solid and dashed lines denote movement at the intervention and final post-RMSNorm representations, respectively. All changes are relative to the corresponding unmodified representations.

$P_A(C \mid E)$. Our diversification instead changes which intermediate states are explored across the population, while subsequent generation remains governed by the same base sampler P_A.

3 Representation Diversity via Model-Mediated Steering

3.1 Choosing the Space for Diversity

Explicit diversity requires a notion of distance or overlap, and the chosen space determines what differences are encouraged. Token-level measures such as edit distance, n-gram overlap, or Self-BLEU primarily capture lexical variation (Zhu et al., 2018; Alihosseini et al., 2019; Tevet & Berant, 2021), and are further complicated in MDLMs because candidates may reveal different token positions at the same denoising step. Diversity can instead be defined over output probabilities, as in ODD (Lamont et al., 2026), but its coordinates correspond to categorical vocabulary entries and do not themselves encode semantic proximity between tokens.

Moving to a learned representation space provides a richer basis for diversity, allowing candidates to be compared through learned contextual and semantic relations rather than only token identity or probability mass. Representation-space diversity has also been successfully explored in autoregressive LLMs (Chen et al., 2026). For MDLMs, we deliberately use the model's own contextual representations. Probing studies show that these representations already encode task-relevant information, including functional correctness and answer information before decoding is complete (Chiu & Liu, 2026; Li et al., 2026). We therefore hypothesize that ordinary sampling reaches representation neighborhoods containing useful reasoning structure but explores them incompletely; defining diversity in this space lets us explicitly search these neighborhoods.

3.2 Separating Diversity Measurement and Intervention

Once diversity is defined over contextual representations, two distinct choices remain: *where diversity is measured* and *where the resulting intervention is applied*. In principle, these can be any two transformer blocks. However, representation geometry can change substantially across depth, so diversity imposed at an intermediate layer need not survive to the output. Since the final contextual representation directly determines the vocabulary logits, we measure the diversity we ultimately seek at the final block. Our layerwise analysis supports this choice: diversity imposed locally at layer 16 progressively attenuates toward the output, whereas a final-layer diversity objective optimized through the same intervention layer increasingly separates the representations as they propagate through the remaining blocks (Fig. 2a). The intervention site introduces a different geometric effect. Let $\mathbf{g}_L = \nabla_{\mathbf{H}_L} \mathcal{D}$ denote the gradient of a final-layer diversity objective. Intervening directly at the final representation locally gives

$$\min_{\delta \mathbf{H}_L} \langle \mathbf{g}_L, \delta \mathbf{H}_L \rangle_F \quad \text{s.t.} \quad \|\delta \mathbf{H}_L\|_F^2 \leq \epsilon \quad \rightarrow \quad \delta \mathbf{H}_L \propto -\mathbf{g}_L \tag{6}$$

Now consider intervening at an earlier block l, with $\mathbf{H}_L = F_{l\to L}(\mathbf{H}_l)$ and downstream Jacobian $\mathbf{J}_{l\to L}$. Under the local linearization $\delta\mathbf{H}_L \approx \mathbf{J}_{l\to L}\delta\mathbf{H}_l$, the corresponding optimization problem in the final-layer space is therefore

$$\min_{\delta\mathbf{H}_L \in \mathrm{Col}(\mathbf{J}_{l\to L})} \langle \mathbf{g}_L, \delta\mathbf{H}_L\rangle_F \quad \text{s.t.} \quad \left\|\mathbf{J}_{l\to L}^{\dagger}\delta\mathbf{H}_L\right\|_F^2 \le \epsilon \quad \longrightarrow \quad \delta\mathbf{H}_L \propto -\mathbf{J}_{l\to L}\mathbf{J}_{l\to L}^{\top}\mathbf{g}_L \tag{7}$$

where $\mathbf{J}_{l\to L}^{\dagger}$ denotes its pseudoinverse.[1]. Thus, direct intervention searches an isotropic local neighborhood at the final representation, whereas earlier intervention induces an ellipsoidal search geometry shaped by the model's own downstream transformation. This gives a model-mediated view of diversity steering: $\mathbf{J}_{l\to L}^{\top}\mathbf{g}_L$ acts as a steering signal at the earlier layer, while the remaining transformer blocks determine how that intervention is realized at the final representation. This propagation-aware perspective is closely related to model-based activation steering, where accounting for downstream model dynamics provides more systematic control over an intervention (Skifstad et al., 2026). Our mechanistic analysis further supports this design: the downstream blocks can substantially reshape a small earlier intervention (Fig. 2b), and for a matched 5% reduction in final-layer overlap, earlier intervention sites require substantially smaller local edits than direct final-layer intervention (Fig. 2c). We therefore measure diversity at the final representation while using an intermediate layer as the steering site through which that diversity is realized.

4 SONAR: DIVERSITY VIA SPECTRAL FEEDBACK

A Functional View of Representation Geometry. Let $\hat{\mathbf{H}}_L^i = \mathrm{RMSNorm}(\mathbf{H}_L^i)$ denote the final post-RMSNorm representation of candidate i[2]. We view these representations as a point cloud in the model's feature space and seek a rich description of its geometry without collapsing the response through pooling operations such as mean or max. Our construction, illustrated in Fig. 3, describes this geometry through an RBF kernel: each representation point induces a localized Gaussian function over the feature space, with nearby points producing strongly overlapping functions and distant points interacting weakly. Together, the point cloud generates a family of functions

$$f_{\boldsymbol{\alpha}}(\cdot) = \sum_{n=1}^{c} \alpha_n\, \phi(\hat{\mathbf{h}}_{L,n}^i)(\cdot)$$

where $\boldsymbol{\alpha}$ specifies a weighting over the representation points, producing different smooth profiles supported by the same point cloud. Varying $\boldsymbol{\alpha}$ therefore produces different smooth profiles of the same point cloud, ranging from localized regions to broader combinations of its structure. Formally, consider an RKHS $\mathcal{H}$ with feature map $\phi : \mathbb{R}^d \to \mathcal{H}$. The function family induced by candidate i spans $\mathcal{S}^i = \mathrm{span}\{\phi(\hat{\mathbf{h}}_{L,1}^i), \ldots, \phi(\hat{\mathbf{h}}_{L,c}^i)\}$. We obtain a compact spectral description of this family through the unnormalized second-moment operator

$$\boldsymbol{C}^i = \sum_{n=1}^{c} \phi(\hat{\mathbf{h}}_{L,n}^i) \otimes \phi(\hat{\mathbf{h}}_{L,n}^i) \approx \sum_{k=1}^{r} \lambda_k^i u_k^i \otimes u_k^i \tag{8}$$

where $(a \otimes b)f = a\langle b, f\rangle_{\mathcal{H}}$ and $\{u_k^i\}_{k=1}^r \subset \mathcal{H}$ are the retained dominant spectral functions, giving a rank-r approximation of the operator. As illustrated in Fig. 3, these functions provide a compact basis for the geometry of the representation point cloud.

RKHS Spectral Overlap. Having obtained a spectral basis for each candidate, we measure how well the basis of one candidate explains the function family induced by another. Specifically, we evaluate candidate j against the dominant spectral span of candidate i as

$$\mathcal{L}_U^{i\leftarrow j} = \frac{1}{c}\sum_{n=1}^{c} \left\|\mathbf{P}_U^i\, \phi(\hat{\mathbf{h}}_{L,n}^j)\right\|_{\mathcal{H}} \quad \to \quad \mathcal{L}_U = \frac{1}{N(N-1)}\sum_{i\neq j} \mathcal{L}_U^{i\leftarrow j} \tag{9}$$

where $\mathbf{P}_U^i = \sum_{k=1}^r u_k^i \otimes u_k^i$ is the orthogonal projector onto the span of the retained RKHS spectral functions of candidate i. A large overlap indicates that the representation points of candidate j are well explained by the dominant spectral functions of candidate i. We aggregate this overlap over all ordered pairs and minimizing $\mathcal{L}_U$ therefore reduces the mutual spectral explainability of the candidates, encouraging their representation point clouds to induce distinct function families.

[1] See Appendix B.1 for derivation

[2] We omit t for brevity, but the quantities are recomputed across denoising steps.

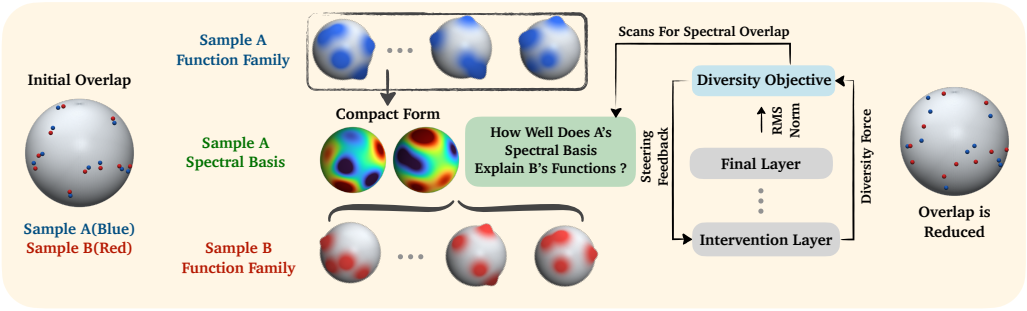

Figure 3: Overview of Sonar. The spectral basis of one sample's function family measures how well it explains another sample, defining a diversity objective at the final MDLM representation. Its gradient provides steering feedback at the intervention layer, while the model's dynamics realize the resulting reduction in overlap.

Token-Coordinate Spectral Structure. The RKHS spectral basis characterizes the geometry of a candidate's representation point cloud, but does not retain how that geometry is organized across the response positions. Two candidates may therefore induce similar spectral functions while their token representations contribute differently to those modes. This information can be examined through the Gram matrix whose rank-r approximation is given by:

$$\mathbf{G}^i \approx \sum_{k=1}^{r} \lambda_k^i \mathbf{v}_k^i (\mathbf{v}_k^i)^\top \qquad \rho^i = \sum_{k=1}^{r} \sqrt{\lambda_k^i} \tag{10}$$

where the coordinates of $\mathbf{v}_k^i$ describe how pairwise similarities among token representations are distributed across the response positions for spectral mode k. We can therefore compare this token-coordinate structure by measuring how strongly the retained eigenbasis of candidate j is supported by the kernel geometry of candidate i:

$$\mathcal{L}_V^{i \leftarrow j} = \frac{1}{\rho^i} \sum_{k=1}^{r} \sqrt{(\mathbf{v}_k^j)^\top \mathbf{G}^i \mathbf{v}_k^j} \quad \rightarrow \quad \mathcal{L}_V = \frac{1}{N(N-1)} \sum_{i \neq j} \mathcal{L}_V^{i \leftarrow j} \tag{11}$$

This objective measures the alignment between the retained dominant eigenspace of candidate j and the token-coordinate structure encoded by $\mathbf{G}^i$. Thus, while $\mathcal{L}_U$ compares the dominant RKHS functions induced by each point cloud, $\mathcal{L}_V$ complements it by comparing the token-coordinate spectral structure underlying those functions.

Optimizing the Diversity Objective. Eq. 11 is already expressed through Gram matrices, the kernelized form of the pairwise RKHS overlap in Eq. 9 is

$$\mathcal{L}_U^{i \leftarrow j} = \frac{1}{c} \sum_{n=1}^{c} \sqrt{(\mathbf{G}_{:,n}^{ij})^\top \mathbf{V}_r^i (\mathbf{\Lambda}_r^i)^{-1} (\mathbf{V}_r^i)^\top \mathbf{G}_{:,n}^{ij}} \tag{12}$$

where $\mathbf{V}_r^i$ and $\mathbf{\Lambda}_r^i$ denote the retained eigenbasis and eigenvalues of the Gram matrix of sample i, and $\mathbf{G}_{:,n}^{ij}$ is the n-th column of the cross-Gram matrix between samples j and i. The derivation of Eq. 12 is provided in Appendix B.2. Rather than directly summing their gradients, we treat $\mathcal{L}_U$ as the primary geometric signal and retain only the component of $\mathcal{L}_V$ that provides a complementary direction. Although both objectives are evaluated post-RMSNorm, their gradients are first propagated through RMSNorm, and gradient surgery is performed at the final transformer representation $\mathbf{H}_L$. Thus,

$$\mathbf{g}_U^i = \nabla_{\mathbf{H}_L^i} \mathcal{L}_U, \qquad \mathbf{g}_V^i = \nabla_{\mathbf{H}_L^i} \mathcal{L}_V \tag{13}$$

We remove from $\mathbf{g}_V^i$ its component parallel to $\mathbf{g}_U^i$ and combine the resulting directions as

$$\mathbf{g}_{V,\perp}^i = \mathbf{g}_V^i - \frac{\langle \mathbf{g}_V^i, \mathbf{g}_U^i \rangle_F}{\|\mathbf{g}_U^i\|_F^2 + \epsilon} \mathbf{g}_U^i \quad \longrightarrow \quad \mathbf{g}_D^i = (1-\beta)\mathbf{g}_U^i + \beta \mathbf{g}_{V,\perp}^i \qquad 0 \leq \beta \leq 1 \tag{14}$$

The corresponding steering direction at intervention layer l is $-\mathbf{J}_{l \rightarrow L}^\top \mathbf{g}_D^i$, while the linearized diversity force induced at the final transformer representation is proportional to $-\mathbf{J}_{l \rightarrow L} \mathbf{J}_{l \rightarrow L}^\top \mathbf{g}_D^i$.

Table 1: Comparison of Sonar and baselines in terms of empirical pass@16 (EP@16) and pass@8. Results are averaged over 8 random seeds and reported as mean ± std. dev. Bold indicates the best result for each model, benchmark, and metric across all temperatures.

		EP@16						Pass@8					
		HumanEval			GSM8K			HumanEval			GSM8K		
	Method	$T = 0.5$	1.0	1.5	0.5	1.0	1.5	0.5	1.0	1.5	0.5	1.0	1.5
LLaDA	Base	40.4 ±1.5	47.6 ±1.4	45.0 ±1.8	76.2 ±1.7	81.5 ±1.4	82.6 ±1.9	35.3 ±1.0	41.6 ±1.2	37.2 ±1.1	72.2 ±1.3	76.7 ±1.0	76.7 ±1.0
	ODD	49.9 ±1.7	50.5 ±2.6	48.4 ±2.4	87.6 ±1.3	87.4 ±1.5	**88.2** ±1.2	41.0 ±1.4	42.9 ±1.8	40.3 ±1.3	82.1 ±1.2	82.0 ±0.8	**82.3** ±1.3
	Sonar	**54.3** ±2.2	51.5 ±1.4	50.9 ±2.1	86.3 ±1.0	86.6 ±1.9	86.1 ±2.3	**46.3** ±1.5	43.2 ±2	42.8 ±1.8	80.1 ±0.9	80.1 ±1.3	80.0 ±1.4
	Method	$T = 0.5$	0.75	1.0	0.5	0.75	1.0	0.5	0.75	1.0	0.5	0.75	1.0
Dream	Base	41.8 ±1.7	51.4 ±1.3	49.3 ±3.0	57.1 ±0.9	70.1 ±1.4	74.1 ±1.9	38.4 ±1.3	45.2 ±0.8	38.7 ±2.0	55.2 ±0.6	64.8 ±1.0	65.3 ±1.3
	ODD	61.6 ±0.9	65.4 ±1.0	64.9 ±1.3	84.8 ±0.8	86.2 ±0.8	**89.8** ±1.3	53.1 ±0.5	56.7 ±0.9	55.4 ±1.1	76.9 ±0.4	79.2 ±0.6	82.1 ±0.9
	Sonar	**69.4** ±2.0	67.8 ±1.7	63.5 ±1.8	88.2 ±0.9	85.8 ±1.0	84.2 ±1.5	**61.1** ±1.2	58.5 ±1.1	53.5 ±1.4	**83.1** ±0.5	80.1 ±0.8	77.7 ±1.0

5 Experimental Evaluation

We empirically evaluate our method using LLaDA-8B-Instruct (Nie et al., 2025) and Dream-7B-Instruct (Ye et al., 2025) on two benchmarks: HumanEval (Chen et al., 2021) (coding), and GSM8K (Cobbe et al., 2021) (mathematics). We compare Sonar against two baselines: (i) temperature-based sampling from the base MDLM, and (ii) Orthogonal Diverse Diffusion (ODD) (Lamont et al., 2026). Following the settings of ODD, we use the first 200 problems in GSM8K. We also sample a set of $N = 16$ solutions per problem, using 64 response tokens and 32 denoising steps for both MDLMs.

5.1 Improving Pass@K Reasoning With Sonar

Each method is evaluated across a range of three temperatures: $T = 0.5, 1.0, 1.5$ for LLaDA-8B-Instruct, and $T = 0.5, 0.75, 1.0$ for Dream-7B-Instruct. We first measure the **empirical pass@K** metric, for $K = 16$ (Lamont et al., 2026), which measures whether a method solves a problem at least once in the *first* K solutions. Given $N = 16$ solutions per problem, EP@16 simply reduces to the fraction of problems with at least one correct solution. We repeat this for $R = 8$ random seeds and report the average in Table 1. To get a more robust estimate, for $K = 8$, we also report the actual **pass@K** metric which computes $1 - \binom{N-C}{K} / \binom{N}{K}$ (where C is the number of correct solutions). For both Sonar and the baseline ODD, we evaluate three levels of diversity strength (refer to table 3, table 2) at each temperature, and report the best.

EP@16 and pass@8 measure how well the reasoning capabilities of the MDLM are revealed over multiple passes. From Table 1, through the lens of both metrics, we observe that Sonar strongly improves the capabilities of the underlying MDLMs. For HumanEval, using LLaDA and Dream, Sonar improves pass@8 by +4.7pp and +15.9pp respectively, and on GSM8K, this improvement is +3.4pp and +17.8pp respectively. Our results suggest that **Dream-7B-Instruct benefits more from diversity compared to LLaDA-8B-Instruct.**

When comparing Sonar to ODD, we find that Sonar outperforms ODD on the more difficult HumanEval benchmark, for both LLaDA and Dream (+3.4pp and +4.4pp pass@8 respectively). For GSM8K, Sonar is slightly weaker on LLaDA (-2.2pp pass@8), but slightly stronger on Dream (+1.0pp pass@8). These results tell us that, though our method has similar results to ODD on the GSM8K benchmark, **Sonar shows a strong improvement in the coding-based HumanEval**.

Finally, we make an observation regarding temperature. **Sonar generally shows strong performance at the lowest temperature we test** ($T = 0.5$), suggesting that the method performs exploration in a controlled way, rather than relying on the temperature's stochasticity.

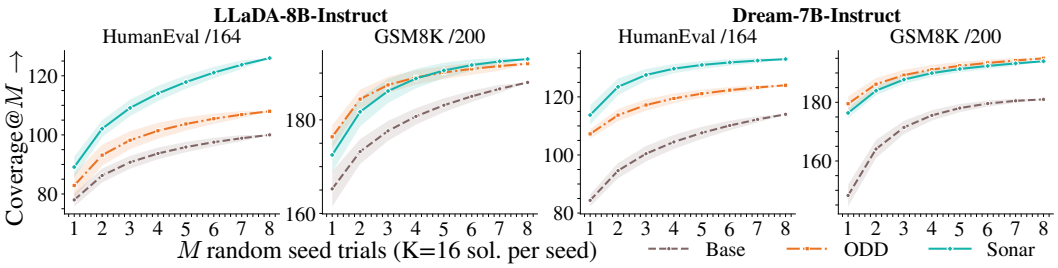

Figure 4: How increasing the number of random seeds, M, improves coverage (unique problems solved) for Sonar and baselines across math and coding (shaded area shows std. dev.).

5.2 The Exploration Capabilities of Sonar

While pass@K captures how many problems the MDLM can be expected to solve given a budget of K candidates, it doesn't sufficiently capture the full picture in terms of exploration. For both ODD and Sonar, a set of 16 solutions do not constitute independent trials, as the samples are dependent on each other. We thus look at exploration across the random seeds. Given that we average EP@16 over $R = 8$ seeds, we ask: *are the models solving the same problems over and over again across seeds?* Or do they solve novel problems?

In order to measure this, we introduce the notion of **Coverage@M**. From $R = 8$ random seeds, we consider subsets of $M \leq R$ seeds, and find the total unique problems solved over the union of those seeds. We further compute the avg. and std. deviation over the $\binom{R}{M}$ ways M seeds can be selected. In Figure 4, we show how coverage@M scales for $M \in \{1, \ldots, 8\}$, using the *best EP@16 result* of each method for every (model, benchmark) pair in Table 1. From Figure 4, for GSM8K, both Sonar and ODD strongly saturate the benchmark, converging to near-identical coverage@8 of ~ 194 on average; on LLaDA and Dream respectively, Sonar and ODD solve just 1 additional problem compared to the other. In contrast, for HumanEval, we observe a much more prominent coverage gap. For LLaDA and Dream respectively, Sonar solves 26 and 19 more problems over the base MDLMs, and 18 and 9 more problems than ODD. This shows that at the same exploration budget, **Sonar is more efficient in covering the solution space compared to baselines**.

5.3 Is Sonar Really Discovering Hidden Reasoning?

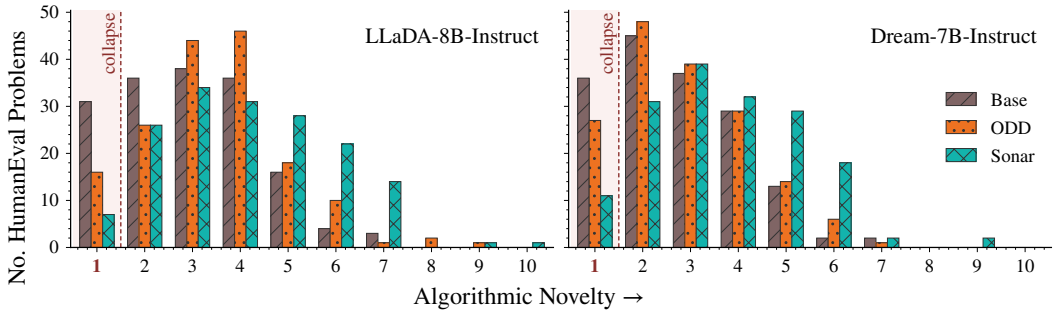

Figure 5: Distribution of novel algorithmic approaches attempted on HumanEval problems. We show that Sonar reduces *algorithmic collapse*, where only a single algorithm is attempted over all eight seeds.

EP@16 and coverage@M are both statistical metrics that are tied to the *correctness* of the results. Even though Sonar improves these metrics, *is it genuinely exploring new reasoning ideas*? This is particularly relevant in HumanEval, where a single coding problem may be solved through multiple algorithms. Therefore, we introduce the **Algorithmic Novelty** metric (AN), which measures the number of novel problem-solving approaches attempted, irrespective of pass or fail. We use an external LLM judge to measure AN in a 'method-blind' manner. For each problem in HumanEval, the LLM judge receives three sets of anonymized solutions in random order, each containing eight

seeds. The judge holistically evaluates all the submitted solutions across all seeds, and for each anonymous method, reports the total number of unique ideas over the union of all seeds.

From Figure 5, we observe a surprising result: even after generating a total of 128 solutions ($K = 16$ candidates $\times$ 8 random seeds), the MDLMs explore only a single algorithmic idea (AN=1) for many HumanEval problems (31 and 36 problems, for LLaDA and Dream respectively). We refer to this as *algorithmic collapse*, where, even if the solutions appear different on the surface, they effectively explore the same redundant idea. We observe both the base MDLMs and ODD admit heavy-tailed AN distributions biased towards AN=1 in Figure 5. By formulating diverse exploration from a semantic rather than vocabulary space, **Sonar substantially reduces algorithmic collapse**.

6 Related Work

Inference-time diversity in language models has been explored through diverse beam search, explicit repulsion between decoding hypotheses, determinant-based selection, and semantic guidance (Vijayakumar et al., 2016; Li et al., 2016; Meister et al., 2021; Shi et al., 2025). A complementary line incorporates diversity into training or post-training, including diverse preference optimization, entropy-based reasoning objectives, diversity-preserving supervised fine-tuning, and joint quality–diversity optimization (Lanchantin et al., 2025; Yao et al., 2025; Li et al., 2025b;a; Chen et al., 2026). These works collectively motivate diversity as a tool for increasing coverage of desired behaviors.

Diversity has also been broadly studied as geometric coverage, for instance, through the lens of the determinant or volume (Kulesza & Taskar, 2012), or the effective rank (Friedman & Dieng, 2022). Several methods incorporate these priors into **text-to-image diffusion models** (Corso et al., 2024; Morshed & Boddeti, 2025; Askari Hemmat et al., 2024) to improve sample diversity and mode coverage. Another growing line of work in language generation defines diversity in learned representation spaces, where geometric separation can capture distinctions beyond lexical or token-level variation. Representation-based exploration uses learned features to encourage coverage of different behaviors (Tuyls et al., 2026), while semantic-guided decoding and steering use representation geometry to expose distinct reasoning trajectories (Shi et al., 2025; Khanh et al., 2026).

Diversity in diffusion language models is a growing topic of interest in recent work. TAPS (Wu et al., 2026) introduces time-dependent perturbations during denoising, while ODD (Lamont et al., 2026) directly coordinates multiple candidates to reduce redundancy and improve Pass@k. ODD is the closest approach to ours: at each step it compresses the response into a vocabulary-level feature vector and sequentially pushes candidate i away from the subspace spanned by candidates $1, \ldots, i-1$. This sequential construction makes the diversity imposed on a candidate depend on its order in the group, and discards the within-response geometry when the token sequence is collapsed to one feature. In contrast, Sonar retains all contextual token representations and compares their spectral geometry, while coupling the population through directed non-self pairs. Our objective is thus symmetric at the population level, not privileging an ordering of the candidates.

7 Concluding Remarks

In this paper, we presented Sonar, a diversity-based approach for controlled sampling in masked diffusion language models. Sonar treats the contextual token representations of each candidate as a geometric object and uses RKHS spectral functions to scan which structure is already occupied by other candidates. The resulting diversity feedback is propagated to an earlier layer, where it steers the model toward less explored directions while allowing the remaining transformer blocks to realize the change at the output-facing representation.

Our experiments show that this geometric steering translates into more effective exploration of the model's solution space. Sonar not only increases the chance of finding a correct solution within a fixed sampling budget, but continues to uncover new solvable problems as additional samples are generated, indicating that its gains do not quickly saturate around the same set of behaviors. This effect is especially clear on HumanEval, where ordinary sampling and ODD often produce many superficially different solutions while repeatedly relying on only one or a small number of underlying algorithms. Sonar breaks this *algorithmic collapse*, recovering multiple distinct solution strategies for the same problem. Together, these findings suggest that the geometric bias introduced by Sonar is not merely creating variation: it directs sampling toward useful regions of the model's representation space that conventional sampling repeatedly fails to explore.

REFERENCES

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 90–98. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-2311.

Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Yilei Chen, Souradip Chakraborty, Lorenz Wolf, Ioannis Paschalidis, and Aldo Pacchiano. Post-training large language models for diverse high-quality responses. In *International Conference on Learning Representations*, 2026.

Guan-Ming Chiu and Jeng-Yue Liu. Probing functional correctness in diffusion language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 163–172. Association for Computational Linguistics, 2026.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, 2021.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 22480–22507, 2024.

Liancheng Fang, Aiwei Liu, Henry Peng Zou, Yankai Chen, Enze Ma, Leyi Pan, Chunyu Miao, Wei-Chieh Huang, Xue Liu, and Philip S. Yu. Locally confident, globally stuck: The quality-exploration dilemma in diffusion language models. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=yRzYJwEOYk.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. Demystifying maskGIT sampler and beyond: Adaptive order selection in masked diffusion. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=mKlW68i2Ig. Featured Certification.

Ly Tran Ho Khanh, Dongxuan Zhu, Man-Chung Yue, and Viet Anh Nguyen. Test-time diverse reasoning by riemannian activation steering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31429–31437, 2026. doi: 10.1609/aaai.v40i37.40407.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

Sean Lamont, Christian Walder, Paul Montague, Amir Dezfouli, and Michael Norrish. Free lunch for pass@ k? low cost diverse sampling for diffusion language models. *arXiv preprint arXiv:2603.04893*, 2026.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*, 2016.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Yi Liang, Soroush Vosoughi, and Shiwei Liu. Diffusion language model knows the answer before it decodes. In *International Conference on Learning Representations*, 2026.

T. Li, Y. Zhang, P. Yu, S. Saha, D. Khashabi, J. Weston, J. Lanchantin, and T. Wang. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025a.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *International Conference on Learning Representations*, 2025b.

Clara Meister, Martina Forster, and Ryan Cotterell. Determinantal beam search. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6551–6562, 2021.

M. M. Morshed and Vishnu N. Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23303–23312, 2025.

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.

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*, 2025.

Theo X. Olausson, Metod Jazbec, Xi Wang, Armando Solar-Lezama, Christian A. Naesseth, Stephan Mandt, and Eric Nalisnick. A tale of two temperatures: Simple, efficient, and diverse sampling from diffusion language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Yangyi Shen, Tianjian Feng, Jiaqi Han, Wen Wang, Tianlang Chen, Chunhua Shen, Jure Leskovec, and Stefano Ermon. Improving diffusion language model decoding through joint search in generation order and token space. *arXiv preprint arXiv:2601.20339*, 2026.

Weijia Shi, Yiming Cui, Yiquan Wu, Jiajun Fang, Shuo Zhang, Ming Li, Shijie Han, Junnan Zhu, Jiaming Xu, and Xiaobo Zhou. Semantic-guided diverse decoding for large language model. In *Advances in Neural Information Processing Systems*, 2025.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local linearity of llms enables activation steering via model-based linear optimal control. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Luhan Tang, Longxuan Yu, Shaorong Zhang, and Greg Ver Steeg. Is your diffusion sampler actually correct? a sampler-centric evaluation of discrete diffusion language models. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=oYCCOl7ahf`.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.25.

Sergey Troshin, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. Control the temperature: Selective sampling for diverse and high-quality LLM outputs. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=IyOC5GCzv4`.

Jens Tuyls, Dylan J. Foster, Akshay Krishnamurthy, and Jordan T. Ash. Representation-based exploration for language models: From test-time to post-training. In *International Conference on Learning Representations*, 2026.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems*, 2024.

Jingxuan Wu, Zhenglin Wan, Xingrui Yu, Yuzhe Yang, Yiqiao Huang, Ivor W. Tsang, and Yang You. Time-annealed perturbation sampling: Diverse generation for diffusion language models. *arXiv preprint arXiv:2601.22629*, 2026.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-aware policy optimization for large language model reasoning. *arXiv preprint arXiv:2505.23433*, 2025.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1097–1100. ACM, 2018. doi: 10.1145/3209978.3210080.

A Appendix

A.1 Practical Implementation

We describe the optimization used to realize the spectral diversity objective at a selected denoising step t. Let $\{\mathbf{H}_{l,0}^i\}_{i=1}^K$ denote the clean, unguided representations at the intervention layer. We forward these states through the remaining MDLM blocks and final RMSNorm to obtain the representations on which $\mathcal{L}_U$ and $\mathcal{L}_V$ are evaluated.

Clean-Relative Diversity Target. The amount of diversity required at a denoising step is defined relative to the overlap naturally produced by the unguided model. From the clean states, we compute

$$\mathcal{L}_{\text{clean}}^t = (1-\beta)\mathcal{L}_{U,\text{clean}}^t + \beta\mathcal{L}_{V,\text{clean}}^t \tag{15}$$

Rather than using a fixed absolute overlap threshold, we require a fractional reduction δ_t from this clean reference,

$$\tau_t = (1-\delta_t)\mathcal{L}_{\text{clean}}^t \tag{16}$$

We decrease δ_t linearly over the intervention window,

$$\delta_t = \delta_{\max} + s_t(\delta_{\min} - \delta_{\max}), \qquad s_t \in [0,1], \tag{17}$$

and gradually relaxing the required reduction in overlap as denoising progresses. (see Fig. 6). The same combined scalar

$$\mathcal{L} = (1-\beta)\mathcal{L}_U + \beta\mathcal{L}_V \tag{18}$$

is used both for the stopping criterion and for comparing candidate steps. Gradient surgery changes the applied gradient, but not this reported overlap scalar.

Frozen Spectral Geometry. For each candidate, we eigendecompose its initial self-Gram matrix and choose the rank required to retain the prescribed spectral energy. We then use the median of these per-sample ranks as a shared rank r for the population. This rank remains fixed throughout the local optimization.

The retained token-coordinate bases and spectral values are initialized from the self-Gram matrices. During each backward pass and during the subsequent candidate evaluations, these spectral quantities are detached and treated as fixed. Thus, gradients modify the current representation relative to a frozen spectral description rather than reducing overlap by rotating the basis during the same optimization step. For every directed comparison, the peer basis and peer anchor are likewise detached; gradients flow through the current target candidate.

After an accepted representation movement, the bases are refreshed from the new self-Gram matrices. Starting from the previous retained basis $\mathbf{V}_r^i$, we apply a small number of warm-started subspace iterations,

$$\mathbf{Y}^i = \mathbf{G}^i\mathbf{V}_r^i, \qquad \mathbf{Y}^i = \mathbf{Q}^i\mathbf{R}^i, \qquad \mathbf{V}_r^i \leftarrow \mathbf{Q}^i \tag{19}$$

The retained spectral values are then updated from the current kernel within this refreshed r-dimensional subspace. No gradient is propagated through this spectral-refresh procedure.

Local Gradient Step. With the spectral geometry frozen, we evaluate $\mathcal{L}_U$ and $\mathcal{L}_V$ and differentiate them separately through the final RMSNorm to the final transformer representation. Their gradients are combined using the surgery in Eq. 14, giving $\mathbf{g}_D^i$. This force is then backpropagated through the remaining transformer blocks to the intervention layer,

$$\mathbf{g}_l^i = \mathbf{J}_{l\to L}^\top \mathbf{g}_D^i \tag{20}$$

The resulting direction is normalized before determining the movement,

$$\mathbf{d}_k = -\frac{\mathbf{g}_{l,k}}{\|\mathbf{g}_{l,k}\|_F + \epsilon_g} \tag{21}$$

Algorithm 1: Sonar representation-diversity optimization

Input: clean intervention states $\{\mathbf{H}^i_{l,0}\}_{i=1}^K$; mixing weight β; backward budget B

Output: applied states $\{\mathbf{H}^i_{l,\text{applied}}\}_{i=1}^K$

$(\mathcal{L}_{\text{clean}}, \{p_i^0\}) \leftarrow \textsc{EvaluateClean}(\{\mathbf{H}^i_{l,0}\})$;
$\delta_t \leftarrow \textsc{LinearSchedule}(t)$; $\tau_t \leftarrow (1-\delta_t)\mathcal{L}_{\text{clean}}$;
$(r, \{\mathbf{V}^i_r, \mathbf{\Sigma}^i_r\}) \leftarrow \textsc{InitializeSubspaces}(\{\mathbf{H}^i_{l,0}\})$;
$\mathbf{H}_l \leftarrow \mathbf{H}_{l,0}$;
for $k \leftarrow 1$ **to** B **do**
  $(\mathcal{L}_U, \mathcal{L}_V) \leftarrow \textsc{PairwiseLSQ}(\mathbf{H}_l; \mathbf{V}_r, \mathbf{\Sigma}_r)$;
  $\mathcal{L}_k \leftarrow (1-\beta)\mathcal{L}_U + \beta\mathcal{L}_V$;
  if $\mathcal{L}_k \le \tau_t$ **then**
    break;
  $(\mathbf{g}_U, \mathbf{g}_V) \leftarrow \textsc{Backward}(\mathcal{L}_U, \mathcal{L}_V)$;
  $\mathbf{g}_D \leftarrow \textsc{Surgery}(\mathbf{g}_U, \mathbf{g}_V, \beta)$; // Eq. 14
  $\mathbf{d}_k \leftarrow \textsc{Normalize}(-\text{VJP}_{L\to l}(\mathbf{g}_D))$;
  $s_k \leftarrow \textsc{PredictStep}(\mathcal{L}_k, \tau_t, \mathbf{d}_k)$;
  $(\mathbf{H}^{1/2}, \mathbf{H}^1) \leftarrow \textsc{Propose}(\mathbf{H}_l, \mathbf{d}_k, s_k)$;
  $(\mathcal{L}_{1/2}, \mathcal{L}_1) \leftarrow \textsc{ForwardLoss}(\mathbf{H}^{1/2}, \mathbf{H}^1; \mathbf{V}_r, \mathbf{\Sigma}_r)$;
  $\mathbf{H}_l \leftarrow \textsc{Select}(\mathbf{H}^{1/2}, \mathbf{H}^1, \mathcal{L}_{1/2}, \mathcal{L}_1, \tau_t)$;
  $(\mathbf{V}_r, \mathbf{\Sigma}_r) \leftarrow \textsc{SubspaceRefresh}(\mathbf{H}_l, r)$; // power iteration + QR
$\mathbf{H}_{l,\star} \leftarrow \mathbf{H}_l$;
$\{p_i^\star\} \leftarrow \textsc{OutputDistributions}(\mathbf{H}_{l,\star})$;
$\{Q^i\} \leftarrow \textsc{UncertaintyScale}(\{p_i^0\}, \{p_i^\star\})$; // Eqs. 26--28
$\mathbf{H}^i_{l,\text{applied}} \leftarrow \mathbf{H}^i_{l,0} + Q^i(\mathbf{H}^i_{l,\star} - \mathbf{H}^i_{l,0}), \quad \forall i$;
return $\{\mathbf{H}^i_{l,\text{applied}}\}_{i=1}^K$;

Adaptive Step Selection. Let $\mathcal{L}_k$ be the current combined overlap scalar at local optimization iteration k. For the actual normalized direction $\mathbf{d}_k$, the predicted decrease per unit displacement is

$$r_k = -\langle \mathbf{g}_{l,k}, \mathbf{d}_k \rangle_F \tag{22}$$

We estimate the displacement required to reach the clean-relative target as

$$s_k = \frac{\mathcal{L}_k - \tau_t}{\max(r_k, \epsilon_r)} \tag{23}$$

subject to the intervention movement cap.

From the same current state $\mathbf{H}_{l,k}$, we evaluate two forward-only proposals,

$$\mathbf{H}^{1/2}_{l,k} = \mathbf{H}_{l,k} + \frac{s_k}{2}\mathbf{d}_k \qquad \mathbf{H}^1_{l,k} = \mathbf{H}_{l,k} + s_k\mathbf{d}_k \tag{24}$$

Both proposals are evaluated using the same frozen spectral geometry and the combined scalar in Eq. 18. If the half-step reaches τ_t, it is accepted. Otherwise, if the full step reaches τ_t, the full step is accepted. If neither reaches the target, the proposal with lower combined overlap is retained. The accepted state becomes the starting point for the next local iteration, after which the spectral bases are refreshed.

Uncertainty Scaling. The local optimizer determines an endpoint $\mathbf{H}^i_{l,\star}$, but the full displacement is applied only if it does not substantially increase the model's predictive uncertainty relative to the clean trajectory. Let $\mathbf{p}^0_n$ and $\mathbf{p}^\star_n$ denote the vocabulary distributions at position n under the clean state and the optimized endpoint, respectively. Using natural logarithms,

$$H^0_n = -\sum_v p^0_{n,v} \log p^0_{n,v} \qquad H^\star_n = -\sum_v p^\star_{n,v} \log p^\star_{n,v} \tag{25}$$

For each token we define

$$q_n = \exp\big(-\max\big(0,\, \mathcal{H}^\star_n - \mathcal{H}^0_n\big)\big) \tag{26}$$

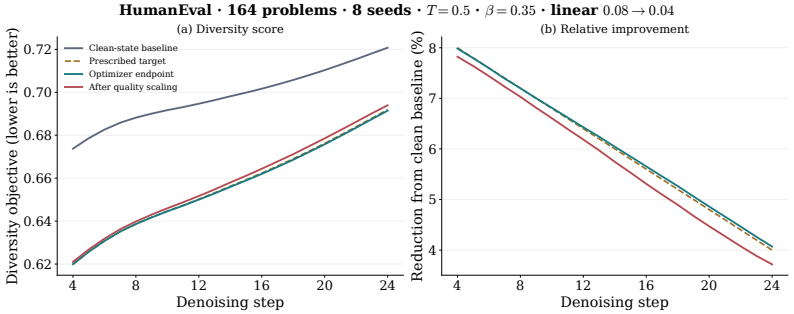

Figure 6: Threshold profile for diversity and Converged threshold of optimizer in Sonar.

Hence an unchanged or reduced entropy gives $q_n = 1$, while increases in uncertainty exponentially reduce the accepted movement.

For the revealed and masked position sets $\mathcal{R}$ and $\mathcal{M}$, respectively,

$$Q_r = \frac{1}{|\mathcal{R}|} \sum_{n \in \mathcal{R}} q_n \qquad Q_m = \frac{1}{|\mathcal{M}|} \sum_{n \in \mathcal{M}} q_n \tag{27}$$

and the per-sample uncertainty scale is

$$Q = \min(Q_r^2, Q_m) \tag{28}$$

Importantly, uncertainty does not define an additional search direction. It is applied only after the local diversity optimization has finished, scaling the total displacement from the original clean state:

$$\mathbf{H}^i_{l,\text{applied}} = \mathbf{H}^i_{l,0} + Q^i \left(\mathbf{H}^i_{l,\star} - \mathbf{H}^i_{l,0} \right) \tag{29}$$

Ordinary MDLM denoising then resumes from the resulting intervention states.

A.2 Hyperparameter

Table 2: Hyperparameters used for diversity-based generation. For Sonar, the diversity threshold is annealed from its initial to final value over the intervention window.

Method	Parameter	Low	Medium	High
Sonar	Diversity threshold	$0.040 \rightarrow 0.020$	$0.080 \rightarrow 0.040$	$0.120 \rightarrow 0.060$
ODD	Diversity weight α	8	16	64
Sonar diversity weighting β			0.35	
Intervention steps			4–24	
Intervention layer (LLaDA)			16	
Intervention layer (Dream)			13	

We consider three diversity-strength profiles, denoted Low, Medium, and High, which differ only in the diversity threshold. In each profile, the threshold decreases from its initial value at intervention step 4 to its final value at step 24. We fix $\beta = 0.35$ across all experiments. Diversification is therefore active only during steps 4–24, avoiding intervention during the earliest and latest portions of denoising. For both models, we intervene at the middle transformer layer: layer 16 for LLaDA and layer 13 for Dream. The RBF bandwidth is determined using the median heuristic over pairwise distances among the pooled candidate representations.

A.3 Algorithmic Novelty Evaluation

Preparing the anonymized problems. For this analysis, we chose the best setting per baseline for the two benchmarks and two MDLMs, i.e. the results correspond to the four plots in fig. 4. We created 164 subfolders, each corresponding to a HumanEval problem. In each problem folder,

Table 3: Diversity strength for best result per method in table 1. Parameters for Low, Medium, High shown in table 2.

		Diversity Strength					
		HumanEval			GSM8K		
Dream	**Method**	$\tau = 0.5$	0.75	1.0	0.5	0.75	1.0
	Base	-	-	-	-	-	-
	ODD	High	High	Low	Medium	High	High
	Sonar	High	High	Medium	High	Medium	Medium
LLaDA	**Method**	$\tau = 0.5$	1.0	1.5	0.5	1.0	1.5
	Base	-	-	-	-	-	-
	ODD	Medium	Low	Low	High	High	High
	Sonar	Medium	Medium	Low	High	High	Medium

we again created three subfolders, simpy named 'method_1', 'method_2', and 'method_3'. For each problem, we randomly assigned a baseline to a method number. For example, for problem 0, methods 1, 2, and 3 could be the base model, ODD, and Sonar respectively, while for problem 1, the order could be ODD, Sonar, and base model instead. This randomized ordering, in addition to explicit instructions to the LLM to ignore the order of the methods, is done in an attempt to (a) prevent any sort of order or sequence-based bias (e.g. it is possible the LLM always gives the first method a better score), or to prevent an approximate identification of methods.

Further, in the instruction to the LLM judge, we mentioned: *Each problem folder contains solutions from multiple methods in a randomized order. That is, for problem 0, methods (1, 2, 3) may be some (X, Y, Z), for problem 1 the methods may be some (W, Z, X), for 2 it may be some (U, V, Y) and so on. There may be more than 3 total methods, and for each problem, a random subset of 3 methods are used to solve the problem.*

By explicitly instructing that there may be an unknown number of methods, we attempt to prevent the LLM judge from discerning that there are always exactly 3 unique attackers, and using this information in its analysis.

Evaluation instructions. We use the following prompt:

What we require: for each problem i, for each method_{j} (j in 1, 2, 3), go over each of the 8 seed files. For each seed file, there are 16 solutions. Read the 16 solutions for every seed k and evaluate objectively: how many unique algorithmic approaches have been attempted in order to solve this problem?

Do this evaluation irrespective of final correctness or pass/fail. Then write down a count in the form: [problem i, method j, seed k]: X unique algorithmic approaches (X between 1 and 16). Write down a brief explanation in a markdown file, what those unique ideas/algorithmic approaches were.

For degenerate responses that are not meaningful (they actually are not solving the problem), or solutions that are just writing comments/docstrings etc. but not writing any code to solve the problem, count them as 0 (no contribution to algorithmic novelty count).

*So at the end, we should have 2 artifacts: a csv file with 164 rows and 8 * 3 = 24 columns, and a markdown file that records what unique ideas/algorithms have been explored per problem. Each group of 8 columns reports the algorithmic novelty / 16 for the 3 methods corresponding to a problem. Once you give your initial verdict, look again at all the solutions of all the methods; make sure you are not biased towards the first method you see. Make corrections to your score as needed.*

*For each problem, the markdown file should contain a rubric that identifies all unique approaches ever attempted on this problem, by any method. This allows a holistic comparison; for each method, you can then identify which labeled ideas were retrieved. After the initial analysis is done, also analyze the total number of unique algorithms attempted over the union of all random seeds, i.e. a total of 8 seeds * 16 responses (128 generation outcomes) per method per problem.*

Result analysis. In practice, we use GPT-6-Pro as our LLM judge. After the evaluation has been done on the anonymized and randomized set, we de-anonymize the results, and present them in fig. 5. This method-blind evaluation indicates that Sonar explores a broader set of algorithmic approaches than the baselines, consistent with its improved HumanEval coverage and pass@K performance.

Supplementary. We provide the anonymized and randomized humaneval samples, the de-anonymization mapping, as well as the LLM judge's blind feedback of algorithmic novelty in the supplementary material.

B Derivation

B.1 Lower-Layer Intervention as Jacobian-Induced Ellipsoidal Search

We derive the final-space interpretation used in Eq. 7. Let $\mathbf{g}_L = \nabla_{\mathbf{H}_L}\mathcal{L}$ and locally linearize the downstream transformation as

$$\delta\mathbf{H}_L \approx \mathbf{J}_{l\to L}\delta\mathbf{H}_l \tag{30}$$

A first-order intervention at layer l therefore solves

$$\min_{\delta\mathbf{H}_l} \langle \mathbf{g}_L, \mathbf{J}_{l\to L}\delta\mathbf{H}_l\rangle_F \qquad \text{s.t.} \qquad \|\delta\mathbf{H}_l\|_F^2 \le \epsilon \tag{31}$$

Using the adjoint relation,

$$\langle \mathbf{g}_L, \mathbf{J}_{l\to L}\delta\mathbf{H}_l\rangle_F = \left\langle \mathbf{J}_{l\to L}^\top \mathbf{g}_L, \delta\mathbf{H}_l\right\rangle_F \tag{32}$$

so the minimum of the linear objective over the Euclidean ball is attained at

$$\delta\mathbf{H}_l^\star = -\sqrt{\epsilon}\,\frac{\mathbf{J}_{l\to L}^\top \mathbf{g}_L}{\|\mathbf{J}_{l\to L}^\top \mathbf{g}_L\|_F} \tag{33}$$

The corresponding first-order displacement at the final representation is therefore

$$\delta\mathbf{H}_L^\star = \mathbf{J}_{l\to L}\delta\mathbf{H}_l^\star = -\sqrt{\epsilon}\,\frac{\mathbf{J}_{l\to L}\mathbf{J}_{l\to L}^\top \mathbf{g}_L}{\|\mathbf{J}_{l\to L}^\top \mathbf{g}_L\|_F} \tag{34}$$

Thus, up to its normalization

$$\delta\mathbf{H}_L^\star \propto -\mathbf{J}_{l\to L}\mathbf{J}_{l\to L}^\top \mathbf{g}_L \tag{35}$$

We can equivalently describe the same search directly in the final representation space. A displacement $\delta\mathbf{H}_L$ reachable through the linearized downstream model must satisfy

$$\delta\mathbf{H}_L \in \mathrm{Col}(\mathbf{J}_{l\to L}) \tag{36}$$

Among all lower-layer perturbations that produce a given reachable $\delta\mathbf{H}_L$, the pseudoinverse gives the minimum-norm preimage. Hence the lower-layer trust region in Eq. 31 is equivalently written as

$$\min_{\delta\mathbf{H}_L \in \mathrm{Col}(\mathbf{J}_{l\to L})} \langle \mathbf{g}_L, \delta\mathbf{H}_L\rangle_F \qquad \text{s.t.} \qquad \left\|\mathbf{J}_{l\to L}^\dagger \delta\mathbf{H}_L\right\|_F^2 \le \epsilon \tag{37}$$

For the following derivation, we identify representation perturbations with their vectorizations, which preserves Frobenius inner products and norms. Consider the reduced singular value decomposition

$$\mathbf{J}_{l\to L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \qquad \mathbf{J}_{l\to L}^\dagger = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top \tag{38}$$

where $\mathbf{\Sigma}$ contains the nonzero singular values. Every reachable final-layer displacement can therefore be written as

$$\delta\mathbf{H}_L = \mathbf{U}\mathbf{Z} \tag{39}$$

Substituting this representation into the constraint gives

$$\left\|\mathbf{J}_{l\to L}^{\dagger}\delta\mathbf{H}_L\right\|_2^2 = \left\|\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{Z}\right\|_2^2 \tag{40}$$

$$= \mathbf{Z}^\top\mathbf{\Sigma}^{-2}\mathbf{Z} \le \epsilon \tag{41}$$

Thus the Euclidean ball at layer l maps to an ellipsoid in the reachable subspace of the final representation, with principal directions $\mathbf{U}$ and axis lengths determined by the singular values of the downstream Jacobian. The final-space problem now becomes

$$\min_{\mathbf{Z}} \left\langle \mathbf{U}^\top\mathbf{g}_L, \mathbf{Z}\right\rangle \qquad \text{s.t.} \qquad \mathbf{Z}^\top\mathbf{\Sigma}^{-2}\mathbf{Z} \le \epsilon \tag{42}$$

The minimizing point on this ellipsoid is

$$\mathbf{Z}^\star = -\sqrt{\epsilon}\,\frac{\mathbf{\Sigma}^2\mathbf{U}^\top\mathbf{g}_L}{\|\mathbf{\Sigma}\mathbf{U}^\top\mathbf{g}_L\|_2} \tag{43}$$

Mapping back to the final representation

$$\delta\mathbf{H}_L^\star = \mathbf{U}\mathbf{Z}^\star \tag{44}$$

$$= -\sqrt{\epsilon}\,\frac{\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top\mathbf{g}_L}{\|\mathbf{\Sigma}\mathbf{U}^\top\mathbf{g}_L\|_2} \tag{45}$$

$$= -\sqrt{\epsilon}\,\frac{\mathbf{J}_{l\to L}\mathbf{J}_{l\to L}^\top\mathbf{g}_L}{\|\mathbf{J}_{l\to L}^\top\mathbf{g}_L\|_F} \tag{46}$$

This is exactly the displacement obtained by directly solving the lower-layer problem in Eq. 31. Therefore, intervening through an earlier layer is locally equivalent to optimizing the final-layer objective over a Jacobian-induced ellipsoidal neighborhood rather than the isotropic neighborhood obtained by direct final-layer intervention.

B.2 Kernelized Evaluation of the RKHS Spectral Projection

For candidate i, recall the unnormalized second-moment operator

$$\mathbf{C}^i = \sum_{n=1}^{c} \phi(\widehat{\mathbf{h}}_{L,n}^i) \otimes \phi(\widehat{\mathbf{h}}_{L,n}^i) \quad \text{where} \quad (a\otimes b)f = a\,\langle b, f\rangle_{\mathcal{H}} \tag{47}$$

We note that $\mathbf{C}^i$ is self-adjoint and is a finite sum of rank-one operators. Therefore, by the spectral theorem for compact self-adjoint operators, its nonzero spectrum admits the finite decomposition

$$\mathbf{C}^i = \sum_{k=1}^{r_i} \lambda_k^i\, u_k^i \otimes u_k^i, \qquad \lambda_k^i > 0, \qquad r_i \le c \tag{48}$$

where $\{u_k^i\}_{k=1}^{r_i}$ is orthonormal in $\mathcal{H}$. The nonzero spectral functions of $\mathbf{C}^i$ admit a finite coordinate representation over the observed RKHS features. Using the definition of the tensor product, the eigenvalue equation

$$\mathbf{C}^i u_k^i = \lambda_k^i u_k^i \tag{49}$$

can be written as

$$u_k^i = \frac{1}{\lambda_k^i}\sum_{n=1}^{c} \left\langle \phi(\widehat{\mathbf{h}}_{L,n}^i), u_k^i\right\rangle_{\mathcal{H}} \phi(\widehat{\mathbf{h}}_{L,n}^i) \tag{50}$$

Define the corresponding token-coordinate vector $\mathbf{v}_k^i \in \mathbb{R}^c$ by

$$v_{n,k}^i = \frac{\left\langle \phi(\widehat{\mathbf{h}}_{L,n}^i), u_k^i\right\rangle_{\mathcal{H}}}{\sqrt{\lambda_k^i}} \tag{51}$$

Then

$$u_k^i = \frac{1}{\sqrt{\lambda_k^i}} \sum_{n=1}^{c} v_{n,k}^i \, \phi(\widehat{\mathbf{h}}_{L,n}^i) \tag{52}$$

The components of the Gram matrix are given by

$$[\mathbf{G}^i]_{mn} = \left\langle \phi(\widehat{\mathbf{h}}_{L,m}^i), \phi(\widehat{\mathbf{h}}_{L,n}^i) \right\rangle_{\mathcal{H}} \tag{53}$$

Using Eq. 52,

$$[\mathbf{G}^i \mathbf{v}_k^i]_m = \sum_{n=1}^{c} \left\langle \phi(\widehat{\mathbf{h}}_{L,m}^i), \phi(\widehat{\mathbf{h}}_{L,n}^i) \right\rangle_{\mathcal{H}} v_{n,k}^i \tag{54}$$

$$= \left\langle \phi(\widehat{\mathbf{h}}_{L,m}^i), \sum_{n=1}^{c} \phi(\widehat{\mathbf{h}}_{L,n}^i) v_{n,k}^i \right\rangle_{\mathcal{H}} \tag{55}$$

$$= \sqrt{\lambda_k^i} \left\langle \phi(\widehat{\mathbf{h}}_{L,m}^i), u_k^i \right\rangle_{\mathcal{H}} \tag{56}$$

$$= \lambda_k^i v_{m,k}^i \tag{57}$$

Hence

$$\mathbf{G}^i \mathbf{v}_k^i = \lambda_k^i \mathbf{v}_k^i \tag{58}$$

so the RKHS spectral function u_k^i and token-coordinate vector $\mathbf{v}_k^i$ are two representations of the same nonzero spectral mode. For the retained rank-r projector

$$\mathbf{P}_U^i = \sum_{k=1}^{r} u_k^i \otimes u_k^i \tag{59}$$

consider token n of candidate j. Using Eq. 52,

$$\left\langle u_k^i, \phi(\widehat{\mathbf{h}}_{L,n}^j) \right\rangle_{\mathcal{H}} = \frac{(\mathbf{v}_k^i)^\top \mathbf{G}_{:,n}^{ij}}{\sqrt{\lambda_k^i}}, \tag{60}$$

where

$$[\mathbf{G}^{ij}]_{mn} = \left\langle \phi(\hat{\mathbf{h}}_{L,m}^i), \phi(\hat{\mathbf{h}}_{L,n}^j) \right\rangle_{\mathcal{H}} \tag{61}$$

Since the retained u_k^i are orthonormal,

$$\left\| \mathbf{P}_U^i \phi(\widehat{\mathbf{h}}_{L,n}^j) \right\|_{\mathcal{H}}^2 = \sum_{k=1}^{r} \left(\left\langle u_k^i, \phi(\widehat{\mathbf{h}}_{L,n}^j) \right\rangle_{\mathcal{H}} \right)^2 \tag{62}$$

$$= (\mathbf{G}_{:,n}^{ij})^\top \mathbf{V}_r^i (\mathbf{\Lambda}_r^i)^{-1} (\mathbf{V}_r^i)^\top \mathbf{G}_{:,n}^{ij} \tag{63}$$

Therefore the spectral overlap in Eq. 9 is evaluated exactly as

$$\mathcal{L}_U^{i \leftarrow j} = \frac{1}{c} \sum_{n=1}^{c} \sqrt{(\mathbf{G}_{:,n}^{ij})^\top \mathbf{V}_r^i (\mathbf{\Lambda}_r^i)^{-1} (\mathbf{V}_r^i)^\top \mathbf{G}_{:,n}^{ij}} \tag{64}$$

C Ablations Studies

C.1 Effect of the Spectral Mixing Weight

The two components of our diversity objective capture complementary aspects of representation geometry: $\mathcal{L}_U$ measures overlap in the RKHS spectral functions, while $\mathcal{L}_V$ measures overlap in their token-position structure. We control their relative contribution using β. To examine the sensitivity to this choice, we sweep $\beta \in \{0, 0.1, 0.2, 0.3, 0.35\}$ on 56 selected HumanEval problems using four seeds.

Figure 7 reports both the average number of problems solved per seed and the number of distinct problems solved across all four seeds. The dependence on β is monotonic. Using only the $\mathcal{L}_U$

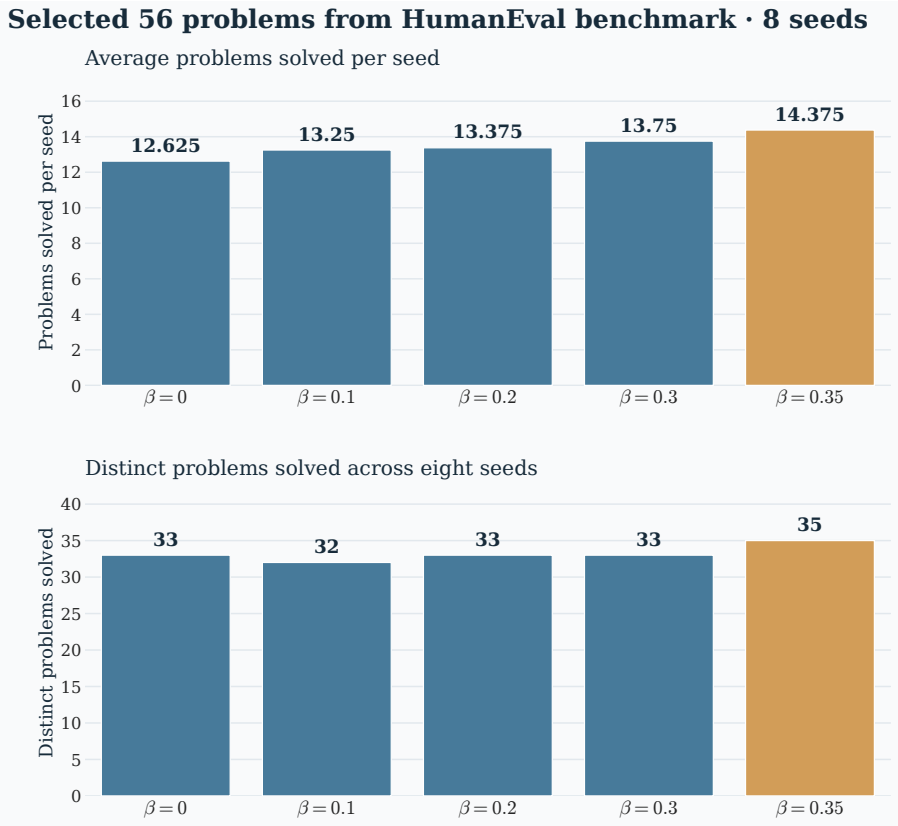

Figure 7: Ablation of the spectral mixing weight β on 56 selected HumanEval problems over 8 seeds. Top: average number of problems solved per seed. Bottom: number of distinct problems solved across all four seeds. The selected setting $\beta = 0.35$ achieves the strongest result under both measures.

component, corresponding to $\beta = 0$, solves 12.625 problems per seed on average and $33/56$ distinct problems. Introducing $\mathcal{L}_V$ as complementary to diversity objective improves these values. Among the tested values, $\beta = 0.35$ gives the strongest result, solving 14.375 problems per seed on average and $35/56$ distinct problems. We therefore use $\beta = 0.35$ in our main experiments. These results indicate that the two spectral views are most useful when combined, while placing excessive weight on either component can reduce the effectiveness of the search.

C.2 Behavioral Effect of the Intervention Layer

Our geometric analysis in Sec. 3.2 motivates measuring diversity at the output-facing representation while applying the intervention at an earlier layer. We additionally test whether this choice translates into a behavioral difference by comparing an intermediate-layer intervention at layer 15 against direct intervention at layer 32 while retaining the final-layer diversity objective.

As shown in Fig. 8, the earlier intervention solves 15.4 problems per seed on average, compared with 12.6 for direct layer-32 intervention. Across the five seeds, it also solves a larger union of problems, $34/56$ compared with $29/56$. Thus, the advantage of the earlier steering site is not limited to requiring a smaller representation edit in the mechanistic analysis; under this ablation it also leads to more successful solutions. This result is consistent with the view that the remaining transformer blocks provide useful downstream dynamics to apply diversity in the last layer rather than this achieved directly by diversity objective gradient.

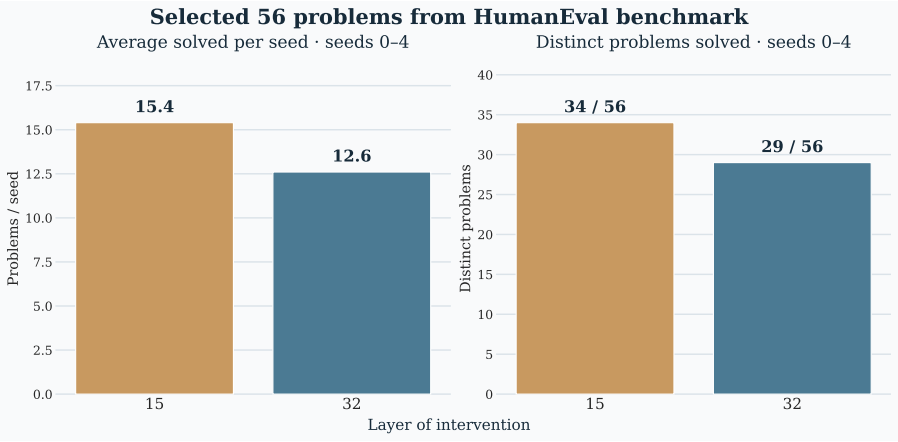

Figure 8: Behavioral comparison of intervention at layer 15 and direct intervention at layer 32 on 56 selected HumanEval problems over seeds 0–4. Left: average number of problems solved per seed. Right: number of distinct problems solved across all five seeds.

C.3 Per-Problem Layerwise Analysis

Figure 9 expands the aggregate layerwise analysis from Fig. 2 and shows the corresponding measurements separately for HumanEval problems 9, 14, and 17 at denoising steps $t \in \{5, 10, 15\}$. The individual trajectories reveal that the aggregate behavior is not produced by a single problem or denoising stage.

The first column compares a diversity objective defined locally at layer 16 with the final-layer objective optimized through the same intervention layer. Across all three problems, the overlap reduction created by the local objective is largest near layer 16 and progressively decreases as the representation propagates toward the output. In contrast, the reduction produced by the final-layer objective is initially small at the intervention site and grows through the downstream blocks, reaching its largest values near the output-facing representation. This directly illustrates that creating diversity at one layer does not imply preservation of the same geometry across depth.

The middle column measures representation movement under the same two objectives. Optimizing the local objective produces substantial movement at the intervention layer but comparatively limited additional separation through the later blocks. The final-layer objective exhibits the opposite pattern: although both methods intervene at the same layer, the downstream transformation increasingly amplifies the movement induced by final-layer feedback. The effect is visible across each of the three denoising stages and all three problems.

Finally, the right column fixes the diversity objective at the final post-RMSNorm representation and varies the intervention layer. Earlier intervention sites consistently require much smaller changes at the edited representation, while the resulting movement at the final representation can be substantially larger. As the intervention approaches layer 32, the required local edit increases and the distinction between edited-layer and final-layer movement narrows. Together, the individual plots support the same conclusion as the averaged analysis in the main paper: measurement and intervention play different roles, and earlier steering allows the model's downstream transformation to realize final-layer diversity from a smaller local edit.

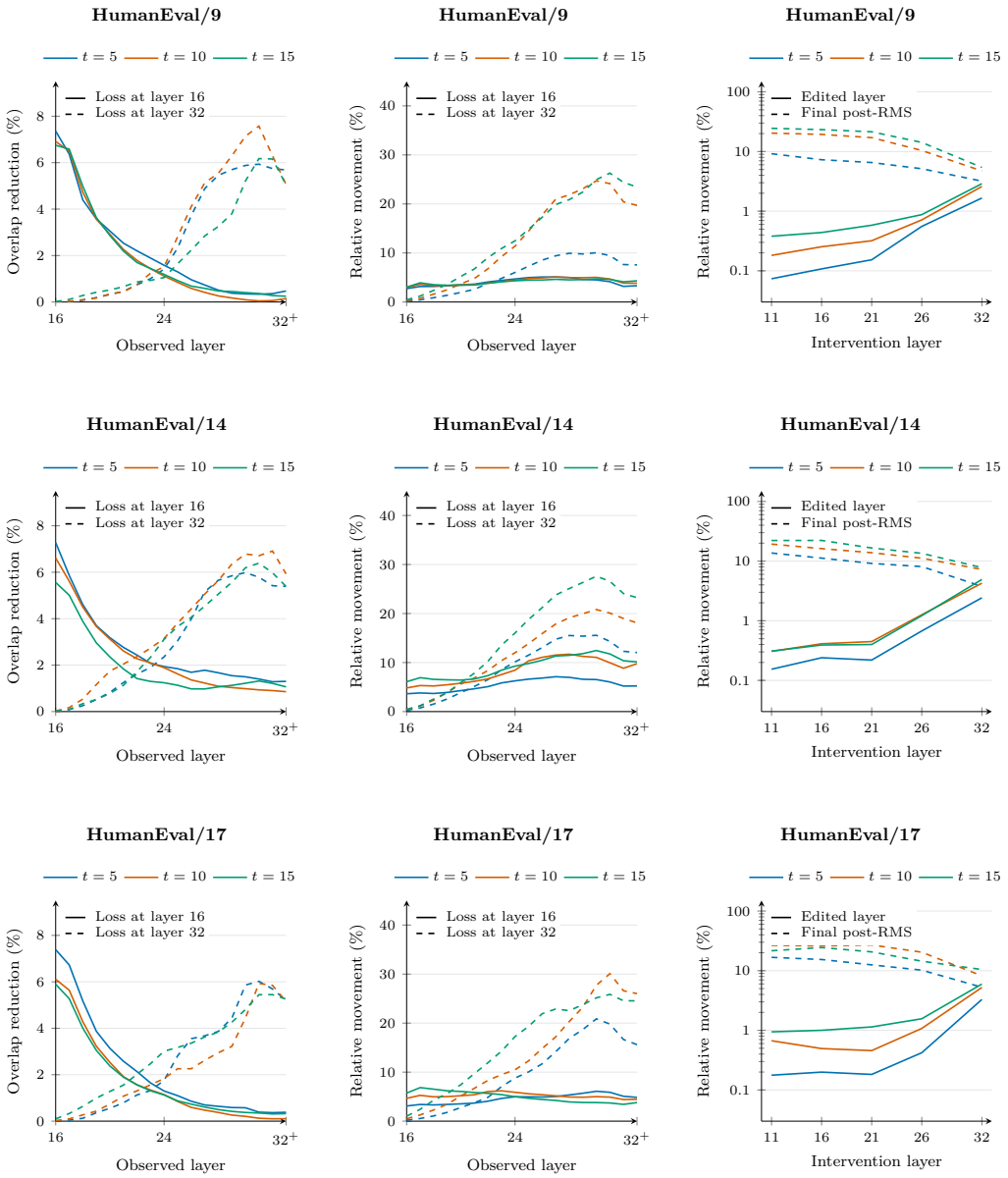

Figure 9: Per-problem version of the layerwise analysis in Fig. 2 for HumanEval problems 9, 14, and 17 at denoising steps $t = 5, 10, 15$. Left: overlap reduction measured across depth when diversity is defined locally at layer 16 (solid) or at the final representation and optimized through layer 16 (dashed). Middle: corresponding relative representation movement. Right: relative movement at the edited layer (solid) and final post-RMSNorm representation (dashed) as the intervention layer is varied while the diversity objective remains fixed at the final representation.

D Limitations & Future Directions

Sonar intentionally performs target-free exploration using only the model's own representation geometry. This allows us to study what useful behaviors can be exposed without task-specific supervision, but also makes the search dependent on how well those behaviors are organized in the learned representation space. A natural extension is to incorporate an external verifier, reward model, or reasoner to provide task-aware guidance.

Our method also introduces additional inference-time cost because representation-space exploration requires backward passes during generation. We view this primarily as an exploration and capability-probing procedure, but improving its efficiency is an important direction for broader use. This cost also limited our experiments to LLaDA and Dream at the 8B scale; evaluating larger diffusion language models such as Gemma Diffusion remains future work.

Another promising direction is to combine representation-space diversification with remasking. While Sonar steers the trajectory toward less explored regions, remasking could allow the model to revise earlier token commitments as that trajectory changes. Finally, our evaluation focuses on mathematical and code reasoning, where multiple valid solution paths are especially meaningful; whether the same geometric exploration benefits broader open-ended generation remains to be studied.

E Details of AI Usage

We used generative AI tools solely to assist with writing and presentation, including improving grammar, clarity, conciseness, and organization of the manuscript. All technical ideas, experimental design, analyses, results, and conclusions were developed and verified by the authors, who take full responsibility for the final content.