CALIBRATING NOISE SCHEDULES FOR CONTINUOUS LANGUAGE DIFFUSION WITH TOKEN UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous language diffusion models use noise schedules to allocate training capacity and guide inference across varying corruption levels. Existing schedules often adapt heuristics from image diffusion, rely on costly denoiser queries, or assume uniform token distributions, overlooking how embedding geometry and token frequency affect recoverability under Gaussian noise. We present **kNN Information Scheduling (KIS)**, an approach that bases noise scheduling on token frequencies and distances between embeddings without querying a denoiser. By modeling token uncertainty under Gaussian noise, KIS tracks conditional entropy across log signal to noise ratios (log-SNRs) and expresses its decrease as a normalized entropy reduction. Instead of replacing schedules without constraints, KIS applies *constrained calibration*: it shifts the training distribution's median while keeping its standard deviation fixed and applies a positive affine correction to the interior sampling levels in log-SNR. Across various architectures (ELF, FLM) and datasets, KIS significantly accelerates inference and enhances generation quality. On FLM-B, KIS improves generative perplexity by up to 25.6% on OpenWebText and 47.1% on LM1B under limited budgets. Notably, a calibrated FLM at 256 steps outperforms its uncalibrated version at 1024 steps, using four times fewer sampling steps. Ablation studies confirm that affine calibration outperforms direct quantile placement at 8–32 steps in the reported comparisons, offering key insights into the relationship between discrete priors, continuous representations, and diffusion dynamics.

1 INTRODUCTION

Continuous language diffusion has emerged as a compelling alternative to discrete autoregressive generation, corrupting real vector representations of tokens, such as learned embeddings, continuous simplexes, and frozen encoder latents, via Gaussian perturbation processes (Li et al., 2022; Hu et al., 2026; Lee et al., 2026; Gulrajani & Hashimoto, 2023). In these architectures, the noise schedule plays a dual operational role: it dictates the allocation of gradient optimization capacity across noise intensities during training, and determines the trajectory discretization grid for the numerical ODE/SDE solver during inference. Continuous image diffusion provides scheduling principles (Kingma et al., 2021; Karras et al., 2022). In language, token recoverability depends not only on the global signal to noise ratio, but also on embedding scale, distances between token embeddings, and skewed corpus token frequencies. Consequently, a scalar noise level alone fails to capture the true semantic information retained along the corruption path.

Current text schedules attempt to address this through heuristic hyperparameter tuning (Hu et al., 2026), uniform decoding error rules (Lee et al., 2026), or measurements from denoisers that track errors in the reverse process and training losses (Sabour et al., 2024; Stancevic & Ambrogioni, 2025; Chen et al., 2026; Raya et al., 2026). Although informative, strategies that query a denoiser require additional model evaluations, and their estimates depend on the denoiser's prediction accuracy as well as token recoverability. Rules that ignore token frequencies omit this prior effect: higher prior frequency expands a token's pairwise decision region at the same noise level (Figure 1a,b). This gap highlights the need for an analytical reference that *requires no denoiser queries* and grounds token uncertainty directly in empirical corpus priors and representation geometry.

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.

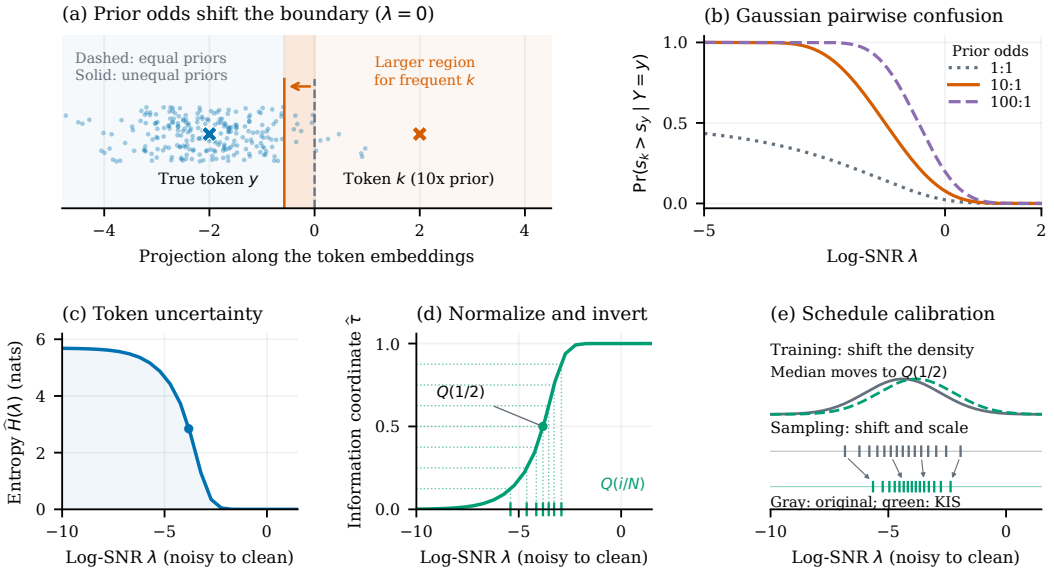


Figure 1: **Deriving diffusion noise schedules from token priors and representation geometry. (a–b)** Analytic characterization of decision boundaries affected by priors and Gaussian pairwise confusion at embedding distance $\|e_y - e_k\|_2 = 4$. **(c–e)** Estimated conditional entropy on OpenWebText using normalized T5-small lookup embeddings ($K = 8$). Conditional entropy $H(\lambda)$ induces a normalized information coordinate $\widehat{\tau}(\lambda)$ and corresponding quantiles $Q(i/N)$. For training, KIS centers the noise distribution at the informational median $Q(1/2)$. For inference, KIS applies an optimal positive affine correction to the interior log-SNR sampling levels, preserving their relative gaps. Gray and green curves denote original and calibrated schedules, respectively. The posterior is analytic, and entropy is estimated by Monte Carlo without denoiser queries.

Using corpus counts and fixed embeddings, we estimate token uncertainty by Monte Carlo under the analytic posterior. The Gaussian I–MMSE relation between mutual information and minimum mean squared error (Guo et al., 2005) connects conditional entropy decay in log signal to noise ratio (log-SNR) to Bayes denoising error. Normalizing the entropy decrease defines an information coordinate whose inverse gives noise levels at equal increments of measured uncertainty reduction. Rather than directly replacing sampling levels with the reference quantiles, we introduce **kNN Information Scheduling (KIS)** under *constrained calibration* (Figure 1c–e). KIS repositions the training density median to align optimization capacity with the information midpoint ($Q(1/2)$) while preserving loss weighting, and fits an optimal positive affine transformation to the sampling grid that directs evaluations toward informative regimes while preserving relative gaps in log-SNR.

KIS accelerates inference and enhances generation quality across diverse continuous diffusion paradigms (contextual latent flows in ELF, one hot flows in FLM, and learned embeddings in LangFlow). On OWT, KIS reduces FLM-B generative perplexity by up to 25.6%, with improvements at small sampling budgets reaching 47.1% on LM1B. Notably, a calibrated FLM at 256 steps surpasses its baseline at 1024 steps, using four times fewer sampling steps. Alignment before training similarly yields steady reductions in loss on withheld data and in perplexity across corpora. Our contributions are:

1. **An Information Metric Without Denoiser Queries:** We establish an analytical framework combining empirical token frequencies and representation geometry to compute conditional entropy under Gaussian noise, producing a normalized information coordinate without querying neural denoisers (§3.2–§3.3).

2. **A Constrained Calibration Paradigm:** We design calibration rules (a median shift for training densities and an affine interior correction for sampling) that align existing schedules with the reference derived from entropy (§3.4).

3. **Comprehensive Empirical Validation:** We evaluate KIS across multiple continuous diffusion formulations, training regimes, and evaluation budgets (§4).

2 Related Work

Continuous Language Diffusion and Representation Spaces. Continuous language diffusion formulates text generation by corrupting continuous representations with Gaussian processes and decoding intermediate states into discrete tokens via rounding or auxiliary decoders (Li et al., 2022; Gong et al., 2023). Rather than diffusing over discrete categorical states (Lou et al., 2024; Sahoo et al., 2024; Nie et al., 2025), continuous approaches operate across varied representations, including continuous probability simplexes (Chen et al., 2023; Han et al., 2023; Karimi Mahabadi et al., 2024), continuous score or flow models (Gulrajani & Hashimoto, 2023; Dieleman et al., 2022; Chen et al., 2026; Lee et al., 2026), and contextual latent representations (Hu et al., 2026; Jo & Hwang, 2025; Yang et al., 2026; Bansal & Sanghavi, 2026; Du & Ma, 2026). However, embedding discrete vocabularies into manifolds of real vectors is acutely sensitive to representation scale and metric anisotropy (Gao et al., 2022; Ye et al., 2024). Unlike image diffusion with uniform pixel ranges (Kingma et al., 2021; Karras et al., 2022), language features nonuniform geometric clustering and token distributions with long tails, implying that identical scalar noise increments do not correspond to uniform degrees of semantic confusion. KIS addresses this by explicitly evaluating the geometry of each representation alongside corpus statistics to quantify intrinsic token identifiability.

Noise Scheduling: Analytical Rules vs. Denoiser Measurements. Noise schedules govern both the distribution of gradient optimization effort during training and the step allocation for inference trajectories (Kingma et al., 2021; Karras et al., 2022; Esteves & Makadia, 2026). Existing text schedules typically rely on heuristic hyperparameter tuning (Hu et al., 2026), uniform decoding error bounds (Lee et al., 2026), or diagnostics that depend on a denoiser that track errors in the reverse process, online losses, and token cross entropy (Sabour et al., 2024; Stancevic & Ambrogioni, 2025; Raya et al., 2026; Chen et al., 2026). These diagnostics require model evaluations, and their uncertainty estimates depend on the denoiser's prediction accuracy as well as token recoverability. Conversely, analytical error rules like FLM implicitly assume uniform token distributions, ignoring that frequent tokens maintain substantially larger decision regions under Gaussian noise. KIS resolves this tension by providing an analytical reference that requires no denoiser queries and integrates empirical unigram priors and metric separation into an explicit Bayesian posterior.

Information Dynamics and Calibration for Numerical Solvers. Perspectives from information theory connect diffusion objectives to progressive information acquisition (Kingma et al., 2021), supported by the I–MMSE identity relating mutual information in Gaussian channels to Bayes denoising error (Guo et al., 2005). While prior efforts estimate entropic progress via denoiser evaluations (Stancevic & Ambrogioni, 2025; Raya et al., 2026), the deployment of a schedule also depends on the numerical solver and learned model (Sabour et al., 2024; Ambrogioni et al., 2026). As detailed in Appendix C, KIS explicitly separates intrinsic uncertainty measurement from operational schedule deployment: we shift the training median and use an affine fit for sampling, then compare these choices with direct quantile placement.

3 Method

Using corpus token statistics and representation geometry, KIS derives analytical noise schedules for continuous language diffusion without querying a denoiser. It calibrates existing training densities and sampling trajectories within a constrained parameter family (Figure 1).

3.1 Problem Setup

The Token Corruption Channel. Let $\mathcal{V}$ denote the target vocabulary, $\pi_y > 0$ the empirical token prior estimated over a reference corpus ($\sum_{y\in\mathcal{V}} \pi_y = 1$), and $e_y \in \mathbb{R}^d$ a fixed embedding assigned to token y. To isolate the fundamental interaction between representation geometry and vocabulary prior from sequence context and network capacity, we formalize diffusion corruption via an isolated

channel with one fixed vector per token:

$$Y \sim \pi, \qquad Z_\lambda = \alpha_\lambda e_Y + \sigma_\lambda \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_d), \qquad \lambda = \log \frac{\alpha_\lambda^2}{\sigma_\lambda^2}, \tag{1}$$

where $\alpha_\lambda, \sigma_\lambda > 0$ parameterize the noise schedule, and $\lambda \in \mathbb{R}$ represents the log signal to noise ratio (log-SNR), with increasing λ indicating progressively cleaner states. This formulation with fixed vectors excludes sequence context, positional encodings, and denoiser parameters. We can therefore estimate the conditional entropy curve offline, using fixed embeddings before training or jointly learned embeddings from a trained checkpoint.

Calibration Objectives. We consider continuous language diffusion model schedules parameterized by a continuous density $q_0(\lambda)$ during training or an N-step discrete sampling schedule defined by interior log-SNR nodes $x_1 < \cdots < x_{N-1}$ bounded by fixed terminal endpoints. While $q_0(\lambda)$ governs the distribution of gradient optimization effort across noise intensities, the sampling schedule dictates trajectory discretization for the underlying numerical solver. Our goal is to recalibrate the allocation of these schedules within a bounded domain $[L, U]$ using the conditional entropy curve.

3.2 Quantifying Token Uncertainty

Conditional Entropy as the Bayes Risk. We measure the intrinsic token ambiguity at corruption level λ via the minimal expected categorical cross entropy (log loss). Over the space of all normalized predictors $r(\cdot \mid z)$, this optimal loss coincides with the conditional entropy of the corruption channel:

$$H(\lambda) = \inf_r \mathbb{E}_{Y, Z_\lambda}[-\log r(Y \mid Z_\lambda)] = H(Y \mid Z_\lambda). \tag{2}$$

Because the Bayes posterior uniquely attains the minimum in equation 2, $H(\lambda)$ measures irreducible token uncertainty under Gaussian corruption, accounting for the full posterior distribution rather than only whether the most probable token is correct. For the Gaussian channel equation 1, the posterior admits an explicit closed form:

$$s_k(z, \lambda) = \pi_k \exp\left(-\frac{\|z - \alpha_\lambda e_k\|_2^2}{2\sigma_\lambda^2}\right), \qquad p_\lambda(k \mid z) = \frac{s_k(z, \lambda)}{\sum_{j \in \mathcal{V}} s_j(z, \lambda)}. \tag{3}$$

Equation 3 shows how prior frequency ratios π_k / π_y and embedding distances $\|e_k - e_y\|_2$ jointly affect token recoverability. Pairwise confusion probabilities and their analytic decision boundaries are derived in Appendix A.1.

Full Posterior and Approximation with the Top K Candidates. Evaluating the exact denominator in equation 3 over the full vocabulary $\mathcal{V}$ becomes computationally intensive for large token sets. To ensure tractability across representations, KIS either computes the full posterior directly or evaluates a localized approximation over a candidate neighborhood A_K. Specifically, let $\mathcal{N}_K(z, \lambda)$ denote the top K candidates possessing the smallest negative unnormalized log posterior scores $-\log s_k(z, \lambda)$, and define the active candidate set as $A_K = \mathcal{N}_K(z, \lambda) \cup \{y\}$. The localized loss and its Monte Carlo expectation are given by:

$$\widetilde{\ell}_K(y, z; \lambda) = -\log \frac{s_y(z, \lambda)}{\sum_{j \in A_K} s_j(z, \lambda)}, \qquad \widetilde{H}_K(\lambda) = \widehat{\mathbb{E}}_{Y \sim \pi, \varepsilon}\big[\widetilde{\ell}_K(Y, Z_\lambda; \lambda)\big], \tag{4}$$

where $\widehat{\mathbb{E}}$ denotes empirical expectation over Monte Carlo draws of prior tokens and Gaussian perturbations. Truncating the normalizer to A_K underestimates the loss under the full posterior: letting r_{tail} represent the posterior mass outside A_K, the pointwise discrepancy satisfies $\ell_{\text{full}} - \widetilde{\ell}_K = -\log(1 - r_{\text{tail}}) \geq 0$. In our implementation, FLM (Lee et al., 2026) evaluates the exact posterior over the full vocabulary, whereas contextual latent proxies and learned embeddings use a finite candidate set ($K = 8$).

3.3 Normalizing and Inverting Conditional Entropy

Let $\widehat{H}(\lambda)$ denote the nonincreasing fit to the estimated entropy curve (Appendix A.2). Its rate of decrease, $-\frac{d}{d\lambda}\widehat{H}(\lambda)$, measures how quickly token uncertainty is resolved as noise recedes. Across

a calibrated log-SNR interval $[\lambda_{\mathrm{n}}, \lambda_{\mathrm{c}}]$ with $\widehat{H}(\lambda_{\mathrm{n}}) > \widehat{H}(\lambda_{\mathrm{c}})$, we define the normalized entropy reduction $\widehat{\tau}(\lambda) \in [0,1]$ (the *information coordinate*) and its inverse $Q(u)$ (the *information quantiles*):

$$\widehat{\tau}(\lambda) = \frac{\widehat{H}(\lambda_{\mathrm{n}}) - \widehat{H}(\lambda)}{\widehat{H}(\lambda_{\mathrm{n}}) - \widehat{H}(\lambda_{\mathrm{c}})}, \qquad Q(u) = \inf\{\lambda : \widehat{\tau}(\lambda) \geq u\}, \quad 0 < u < 1. \tag{5}$$

Under this mapping, $Q(u)$ identifies the noise level at which a fraction u of the measured entropy decrease has occurred. Evaluating Q across uniform intervals yields a sampling schedule with equal information increments:

$$\lambda_i^{\mathrm{info}} = Q(i/N), \qquad i = 0, \ldots, N, \tag{6}$$

with boundaries explicitly assigned as $Q(0) = \lambda_{\mathrm{n}}$ and $Q(1) = \lambda_{\mathrm{c}}$. When $\widehat{\tau}$ is continuous and strictly increasing, each subinterval in equation 6 encapsulates an identical information increment of $1/N$.

From Lookup Embeddings to Contextual Latents. Under Gaussian noise, uniformly scaling fixed token embeddings by $s > 0$ shifts the quantiles to $Q_s(u) = Q(u) - 2\log s$, with the prior fixed and measurement endpoints shifted accordingly (Proposition 2, Appendix A.2). We use this identity to motivate a scale correction from lookup embeddings to contextual latents. It does not imply that contextualization preserves distances or entropy curves. For Embedded Language Flows (ELF) (Hu et al., 2026), which corrupt contextual representations according to $z_t = tx + (1-t)c\varepsilon$ ($c = 2$), the model log-SNR corresponds to $\lambda_{\mathrm{model}} = 2\,\mathrm{logit}(t) - 2\log c$. We compute the reference quantiles from frozen T5-small lookup embeddings, normalized to unit RMS per coordinate using corpus frequency weights, and empirical token frequencies. Let r be the median RMS across sequences of contextual latents after ELF normalization. We use $Q_{\mathrm{model}} = Q_{\mathrm{profile}} - 2\log r$ as a reference adjusted for scale, rather than a measurement of contextual latent entropy. For a logit-normal training density $\mathrm{logit}(t) \sim \mathcal{N}(P_{\mathrm{mean}}, P_{\mathrm{std}}^2)$, the center implied by the information median is:

$$P_{\mathrm{mean}}^* = \frac{Q_{\mathrm{model}}(1/2)}{2} + \log c = \frac{Q_{\mathrm{profile}}(1/2)}{2} + \log \frac{c}{r}, \tag{7}$$

while keeping the standard deviation P_{std} unchanged.

3.4 Constrained Schedule Calibration

We use information quantiles (equation 6) to calibrate existing schedules: shifting the training median and fitting a positive affine transformation to the interior sampling levels in log-SNR (Algorithms 1 and 2). We call these restricted adjustments *constrained calibration*.

Algorithm 1 KIS for training

Require: Data $\mathcal{D}$, model θ, density q_0, quantiles Q, loss ℓ_θ

$\delta \leftarrow Q(1/2) - \mathrm{median}(q_0)$
$q_\delta(\lambda) \leftarrow q_0(\lambda - \delta)$
for each training update **do**
  Draw $x \sim \mathcal{D}$, $\lambda \sim q_\delta$, noise ε
  Corrupt x along the original path
  Update θ with ℓ_θ and original optimizer
end for
return θ, q_δ

Algorithm 2 KIS for sampling

Require: Interior nodes x_i, quantiles Q, time map g, original endpoints

$y_i \leftarrow Q(i/N)$, $i = 1, \ldots, N-1$
Fit (a^*, b^*) by Eq. 9
$\lambda_i' \leftarrow a^* + b^* x_i$
Map λ_i' to model time with g
Restore the original endpoints
return calibrated sampling schedule

Training Calibration: Centering. Rather than distorting the training noise schedule density, we shift $q_0(\lambda)$ such that its median aligns with the informational midpoint $Q(1/2)$, where half of the measured entropy decrease has occurred:

$$q_\delta(\lambda) = q_0(\lambda - \delta), \qquad \delta = Q(1/2) - \mathrm{median}(q_0). \tag{8}$$

This shift preserves the shape and standard deviation of q_0 in log-SNR and retains the original loss weight w, giving the expected objective $\int q_\delta(\lambda) w(\lambda) R_\theta(\lambda)\, d\lambda$ without changing the architecture or optimizer. For ELF models, this calibration directly instantiates the analytic update P_{mean}^* specified in equation 7.

Table 1: **KIS computes a noise schedule shift before training that improves ELF-B on both corpora.** KIS reduces GenPPL from 121.2 to 90.1 on LM1B (25.7% relative reduction) and from 27.9 to 25.8 on OWT (7.5%). Baselines are listed at their original budgets (Hu et al., 2026).

	LM1B, $L = 128$		OWT, $L = 1024$	
Model	Steps	GenPPL ↓	Steps	GenPPL ↓
Published references				
AR Transformer	AR	66.7	AR	35.9
MDLM (Sahoo et al., 2024)	128	103.9	1024	104.9
Duo (Sahoo et al., 2025)	128	97.6	1024	77.6
FLM (Lee et al., 2026)	1024	96.9	1024	62.2
LangFlow (Chen et al., 2026)	128	92.2	1024	36.5
ELF-B	128	121.2	128	27.9
+ KIS training	128	**90.1**	128	**25.8**

Sampling Calibration: Affine Correction. For inference with frozen checkpoints, we project the original interior sampling levels $x_1 < \cdots < x_{N-1}$ toward the information quantiles $y_i = Q(i/N)$ via an optimal positive affine transformation:

$$(a^*, b^*) = \arg\min_{a,\, b \geq b_{\min}} \sum_{i=1}^{N-1} (a + b x_i - y_i)^2, \qquad b_{\min} > 0, \tag{9}$$

where $b_{\min} = 10^{-6}$ ensures strict monotonicity. Equation 9 is equivalent to minimizing the squared 2-Wasserstein distance in one dimension between the empirical distributions of the transformed nodes and the target quantiles (Appendix A.3). A positive affine transformation preserves the interior log-SNR gap ratios:

$$\frac{\lambda'_{i+1} - \lambda'_i}{\lambda'_i - \lambda'_{i-1}} = \frac{x_{i+1} - x_i}{x_i - x_{i-1}},$$

This is an algebraic property, not evidence of improved solver stability. The calibrated levels $\lambda'_i = a^* + b^* x_i$ are mapped to model time via $g(\lambda)$ and combined with the original endpoints (Algorithm 2). These operations need not preserve ratios between time steps.

4 Experiments

We empirically validate the proposed method across two deployment regimes: training distribution alignment (§4.1) and trajectory recalibration during inference (§4.2). We perform targeted ablations to study calibration constraints, nonuniform corpus priors, and neighborhood approximations (§4.3). Following established text diffusion benchmarks (Hu et al., 2026; Lee et al., 2026), we evaluate text generation quality using generative perplexity (GenPPL ↓) evaluated by GPT-2 Large, alongside mean unigram token entropy per sample (H_{tok}, in nats) to monitor sample diversity. For sampling calibration, we hold the underlying model checkpoint, tokenizer, numerical ODE/SDE solver, classifier-free guidance configuration, and evaluation scoring pipeline strictly invariant between paired baseline and calibrated runs.

4.1 Before Training: Aligning the Noise Schedule

For ELF (Hu et al., 2026), equation 7 analytically maps the channel information median $Q(1/2)$ (computed from corpus token statistics and frozen T5-small lookup embeddings, then adjusted for contextual RMS) into the central parameter P_{mean} of the native logit-normal noise density. This shifts P_{mean} from its heuristic default of -1.5 to -1.1743 on LM1B (Chelba et al., 2013) and -1.2162 on OWT (Gokaslan et al., 2019). In log-SNR space, the native density median is situated 0.65 (LM1B) and 0.57 (OWT) units toward the regime with high noise relative to $Q(1/2)$. This discrepancy reveals that stock ELF allocates more training steps to severe corruption regimes.

This alignment before training yields substantial performance gains, reducing GenPPL from 121.2 to 90.1 on LM1B (25.7% relative reduction) and from 27.9 to 25.8 on OWT (7.5%) (Table 1). Notably,

Table 2: **Without retraining, KIS improves the majority of frozen models across three representations.** Each KIS row changes only the sampling schedule of the model above it. We report GenPPL together with unigram token entropy per sample H_{tok}.

	LM1B, $L = 128$			OWT, $L = 1024$		
Model	Steps	GenPPL ↓	H_{tok}	Steps	GenPPL ↓	H_{tok}
ELF-B (105M)	128	121.20	3.400	32	23.55	5.154
+ KIS	128	**114.40**	3.390	32	**21.59**	5.072
ELF-M (342M)	128	112.08	3.486	64	21.70	5.180
+ KIS	128	**110.05**	3.477	64	**21.50**	5.151
ELF-L (652M)	128	**108.88**	3.235	64	**23.30**	5.280
+ KIS	128	109.62	3.515	64	24.15	5.241
FLM-B	1024	97.95	4.293	1024	62.23	5.330
+ KIS	1024	**93.95**	4.278	1024	**46.28**	5.148
LangFlow	128	**92.24**	4.310	1024	**36.53**	5.250
+ KIS	128	94.08	4.208	1024	37.26	5.225

Table 3: **Calibrated schedules matter most when steps are scarce.** Entries report measured Original → KIS GenPPL at different budgets for each model and corpus pair.

Model	Corpus	Small N	Original → KIS	Large N	Original → KIS
ELF-B	OWT	4	139.20 → 116.48	32	23.55 → 21.59
ELF-M	OWT	4	299.37 → 232.45	32	25.62 → 25.38
ELF-L	OWT	4	194.73 → 217.55	32	30.51 → 31.62
FLM-B	LM1B	8	252.89 → 133.67	1024	97.95 → 93.95
LangFlow	LM1B	4	437.74 → 610.20	32	113.45 → 125.22

on LM1B, this correction enables ELF-B to surpass prior diffusion models using 128 steps (which span 92.2–103.9), even though those methods rely on independently tuned training systems. These results demonstrate that embeddings and vocabulary statistics provide sufficient signal to identify superior optimization regimes *a priori*, bypassing empirical schedule grid searches.

4.2 After Training: Trajectory Recalibration for Frozen Checkpoints

The same information coordinate can be deployed on a frozen denoiser to reallocate the finite evaluation budget. Holding the trained weights, numerical solver, step budget, and decoding pipeline fixed, we recalibrate only the interior log-SNR grid via the positive affine transformation formulated in Section 3.4 (Table 2).

Efficacy Across Diverse Embedding Geometries. The impact of calibration using token frequencies is most pronounced in FLM-B (Lee et al., 2026), whose default sampling schedule uses decoding error with uniform token frequencies. Reevaluating this exact one hot geometry under the posterior weighted by empirical token frequencies reduces GenPPL by 25.6% on OWT and 4.1% on LM1B. Similarly, for continuous latent flows (ELF-B and ELF-M), our T5-small lookup proxy with the scale correction consistently improves generation quality across both corpora. These perplexity reductions are achieved with negligible degradation in sample diversity, with unigram token entropy H_{tok} barely shifting.

Improved Gains Under Compute Constraints. Table 3 and Figure 2 compare original and calibrated schedules across paired sampling budgets. At minimal budgets, KIS achieves dramatic perplexity reductions: 16.3% for ELF-B/OWT ($N = 4$), 22.4% for ELF-M/OWT ($N = 4$), and 47.1% for FLM-B/LM1B ($N = 8$). As budgets expand, these margins moderate to 8.3%, 0.9%, and 4.1%, respectively, confirming that scheduling guided by information provides the greatest marginal return when solver evaluations are severely constrained. Remarkably, calibrated FLM-B at only 256 steps achieves a GenPPL of 94.73, outperforming the uncalibrated baseline at 1024 steps (97.95), using $4\times$ fewer sampling steps.

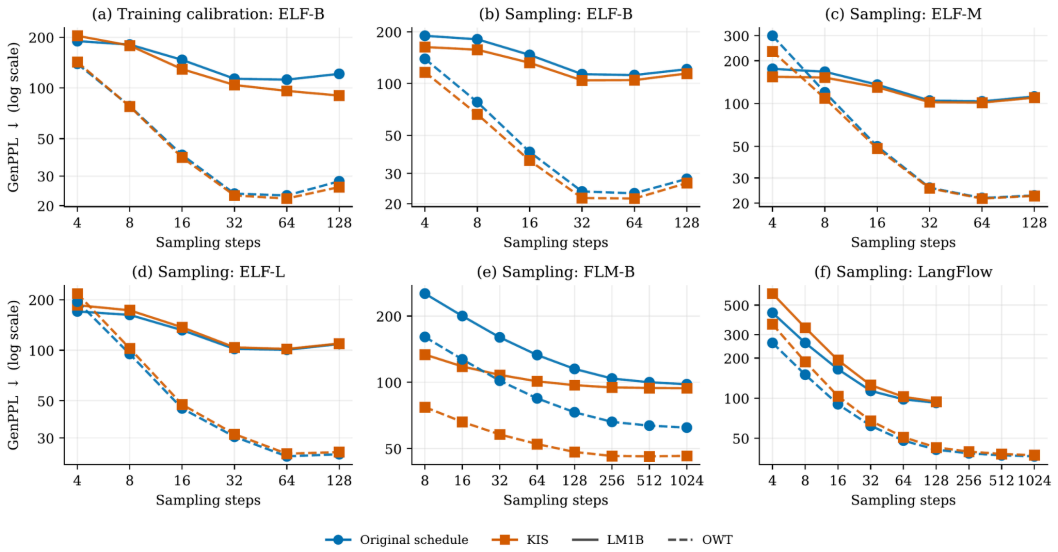


Figure 2: **KIS across sampling budgets.** Panel (a) compares original and calibrated ELF-B training. Panels (b–f) compare original and calibrated sampling grids for frozen models.

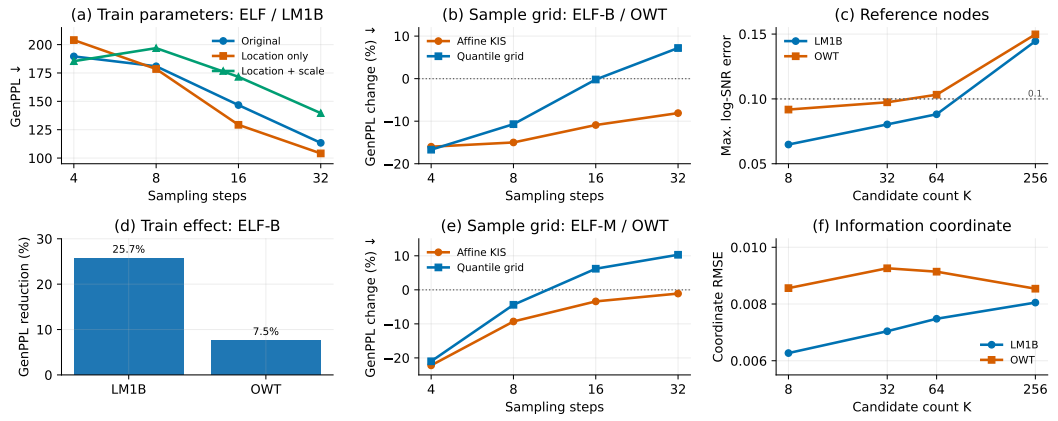


Figure 3: **Comparing training and sampling calibration rules.** Left (a, d): shifting only the training median gives lower GenPPL at 8–32 steps than also narrowing the distribution. Middle (b, e): affine calibration gives lower GenPPL than direct placement at equal information increments across 8–32 steps. Right (c, f): the $K = 8$ reference closely tracks the information coordinate from the full posterior.

4.3 Ablation Studies

Rather than replacing default schedules with unconstrained information quantiles, KIS calibrates within restricted parametric families. We isolate the necessity of this constraint, the role of empirical token distributions, and the estimator's sampling fidelity (Figure 3).

Training Allocation: Median and Standard Deviation. Table 4 compares a shift of the median alone with a joint adjustment that, in this experiment, reduces P_{std} from 0.800 to 0.555 at the same $P_{\text{mean}} = -1.174$. The narrower distribution gives lower GenPPL at 4 steps (185.4 vs. 204.0), but higher GenPPL at 8–32 steps (e.g., 139.6 vs. 104.1 at $N = 32$). This comparison favors the shift of the median alone in equation 8 at 8–32 steps. It does not establish a general disadvantage of adjusting the standard deviation.

Table 4: **Training control: shifting the median alone or also narrowing the distribution.** Generation uses the original sampling schedule. Lower GenPPL is better.

Training distribution	P_{mean}	P_{std}	$N = 4$	$N = 8$	$N = 16$	$N = 32$
Original	-1.500	0.800	189.6	180.9	146.7	113.4
Median only	-1.174	0.800	204.0	**178.5**	**129.3**	**104.1**
Median + std.	-1.174	0.555	**185.4**	196.9	171.6	139.6

Table 5: **Both affine calibration and a reference that uses token frequencies matter.** Entries are median paired GenPPL changes (%) relative to the original schedule. Lower is better.

Model	Schedule	4 steps	8 steps	16 steps	32 steps
Grid parameterization					
ELF-B	Affine KIS	-16.0	-15.0	-10.9	-8.1
ELF-B	Quantile grid $Q(i/N)$	-16.7	-10.7	-0.2	$+7.2$
ELF-M	Affine KIS	-22.2	-9.3	-3.4	-1.1
ELF-M	Quantile grid $Q(i/N)$	-21.0	-4.4	$+6.2$	$+10.3$
Uncertainty reference					
ELF-B	Affine KIS	-16.2	-14.4	-10.5	-7.6
ELF-B	FLM rule, direct	-34.2	-10.0	$+6.6$	$+12.0$
ELF-B	FLM rule, affine	-33.9	-11.3	$+5.8$	$+9.7$
ELF-M	Affine KIS	-22.1	-9.6	-4.5	-0.7
ELF-M	FLM rule, direct	$+27.0$	$+20.1$	$+26.5$	$+30.4$
ELF-M	FLM rule, affine	$+26.9$	$+20.0$	$+26.5$	$+27.6$

Inference Discretization: Relative Gap vs. Direct Quantile. During inference, directly placing evaluation nodes at uniform information quantiles $Q(i/N)$ is competitive only under very sparse discretization ($N = 4$, Table 5). Across all intermediate regimes ($N \in [8, 32]$), affine calibration systematically outperforms the raw quantile grid. Unconstrained quantile placement degrades quality at 32 steps and inflates GenPPL by $+7.2\%$ on ELF-B and $+10.3\%$ on ELF-M (Figure 3b,e).

Corpus Statistics: Necessity of Empirical Frequency Weighting. The lower tier of Table 5 substitutes parameters in KIS with FLM's decoding rule with uniform token frequencies. While the uniform rule performs strongly on ELF-B at 4 steps, KIS consistently surpasses it across practical budgets ($N \geq 8$). More critically, on ELF-M, transferring this uniform rule severely degrades performance across all step counts ($+20.0\%$ to $+30.4\%$ GenPPL increase), whereas KIS maintains steady gains. Because geometric separation interacts nonlinearly with skewed unigram frequencies, schedules derived under uniform assumptions fail to generalize across representation spaces.

5 Conclusion

We introduced KIS to calibrate continuous language diffusion schedules using information theory, corpus token priors, and representation geometry. Without denoiser queries, KIS maps corruption levels to a normalized semantic information coordinate through conditional entropy under Gaussian noise. Constrained calibration shifts the training median and applies a positive affine transformation to interior log-SNR sampling levels, significantly improving sample quality. KIS enables models to match or surpass baseline generation quality with up to $4\times$ fewer evaluation steps.

Limitations and Future Scope. Our approach uses static token vectors shared across positions, without modeling the dynamic effects of attention with multiple heads or deep sequence context. For models that diffuse encoder latents, the reference uses an approximate latent proxy rather than the fully coupled sequence latent space. Future work includes adding n-gram or local contextual priors to the channel posterior, extending constrained calibration to discrete diffusion, and updating representation geometry as the training schedule evolves.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing, literature discovery and brainstorming, and software development, including code generation, debugging, and documentation. The authors reviewed and revised text produced with AI assistance, checked cited literature against the original sources, and inspected and tested code produced with AI assistance. Outputs generated with AI were not used as experimental measurements. The authors take responsibility for the final manuscript, including its claims, results, references, and associated code and artifacts.

ETHICS STATEMENT

This work studies noise schedule calibration for language diffusion models using existing text benchmarks. Such models may reproduce biases, harmful content, or private information present in their training data. Schedule calibration does not address these underlying risks. Our evaluation measures generative perplexity and token entropy, which do not establish fairness, factual reliability, privacy protection, or safety. Any deployment of models using this method would require additional evaluation and safeguards appropriate to the intended application.

REPRODUCIBILITY STATEMENT

Section 3 defines the information coordinate and the training and sampling calibration rules, with implementation procedures in Algorithms 1 and 2. Section 4 describes the evaluated models and datasets, evaluation metrics, sampling budgets, and ablation comparisons. Appendix A provides the channel derivations, estimation with finite candidate sets and normalization conventions, and details for each sampler. Appendix B states the assumptions and derivations underlying the relations from information theory and describes the analysis of the fitted curves. Appendix D gives the measurement settings, controlled training protocols, and inference details for fixed checkpoints.

REFERENCES

Luca Ambrogioni, Giulio Franzese, Alberto Foresti, Gabriel Raya, Bac Nguyen, Georgios Batzolis, Yuhta Takida, Naoki Murata, Chieh-Hsin Lai, and Yuki Mitsufuji. From atoms to entropy: Optimal noise allocation for diffusion training in the convex regime. *arXiv preprint arXiv:2607.20540*, 2026.

Parikshit Bansal and Sujay Sanghavi. Token time continuous diffusion for language modeling. *arXiv preprint arXiv:2607.14106*, 2026.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *International Conference on Learning Representations*, 2023.

Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. LangFlow: Continuous diffusion rivals discrete in language modeling. *arXiv preprint arXiv:2604.11748*, 2026.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Zhicheng Du and Lan Ma. Continuous language diffusion as a decoder-interface problem. *arXiv preprint arXiv:2606.08810*, 2026.

Carlos Esteves and Ameesh Makadia. Spectrally-guided diffusion noise schedules. *arXiv preprint arXiv:2603.19222*, 2026.

Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. Difformer: Empowering diffusion models on the embedding space for text generation. *arXiv preprint arXiv:2212.09412*, 2022.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus, 2019.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations*, 2023.

Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-based diffusion language models. In *Advances in Neural Information Processing Systems*, 2023.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, April 2005. ISSN 0018-9448. doi: 10.1109/tit.2005.844072. URL https://doi.org/10.1109/TIT.2005.844072.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. ELF: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026.

Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. *arXiv preprint arXiv:2502.11564*, 2025.

Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E. Peters, and Arman Cohan. TESS: Text-to-text self-conditioned simplex diffusion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M. Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2024.

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. *arXiv preprint arXiv:2502.09992*, 2025.

Gabriel Raya, Bac Nguyen, Georgios Batzolis, Yuhta Takida, Dejan Stancevic, Naoki Murata, Chieh-Hsin Lai, Yuki Mitsufuji, and Luca Ambrogioni. Noise scheduling as information-guided allocation in diffusion training. *arXiv preprint arXiv:2602.18647*, 2026.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. In *International Conference on Machine Learning*, 2024.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T. Chiu, and Volodymyr Kuleshov. The diffusion duality. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52584–52619. PMLR, 2025. URL https://proceedings.mlr.press/v267/sahoo25a.html.

Dejan Stancevic and Luca Ambrogioni. Entropic time schedulers for generative diffusion models. *arXiv preprint arXiv:2504.13612*, 2025.

Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language. *arXiv preprint arXiv:2605.18530*, 2026.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. DINOISER: Diffused conditional sequence learning by manipulating noises. *Transactions of the Association for Computational Linguistics*, 2024.

A Details of information measurement and calibration

A.1 Bayes log loss and token competition

For any categorical predictor r, conditioning on the noisy observation gives

$$\mathbb{E}[-\log r(Y \mid Z_\lambda)] = H(Y \mid Z_\lambda) + \mathbb{E}_{Z_\lambda}[\mathrm{KL}(p_\lambda(\cdot \mid Z_\lambda) \,\|\, r(\cdot \mid Z_\lambda))]\,. \tag{10}$$

Nonnegativity of Kullback–Leibler (KL) divergence yields equation 2. This is the standard Bayes risk characterization of conditional entropy, used here to define the measurement target.

In the token corruption model, each token y has one fixed embedding e_y. The Gaussian likelihood gives an exact pairwise explanation of the interaction between embedding geometry and token frequency. For distinct token embeddings, define $r_{yk}(\lambda) = e^{\lambda/2}\|e_y - e_k\|_2$. Under $Y = y$, the log score ratio $\log(s_k/s_y)$ is Gaussian with mean $\log(\pi_k/\pi_y) - r_{yk}^2/2$ and variance r_{yk}^2. Therefore

$$\Pr[s_k(Z_\lambda, \lambda) > s_y(Z_\lambda, \lambda) \mid Y = y] = \Phi\left(\frac{\log(\pi_k/\pi_y)}{r_{yk}(\lambda)} - \frac{r_{yk}(\lambda)}{2}\right), \tag{11}$$

where Φ is the standard Gaussian cumulative distribution function. This identity explains pairwise ambiguity. The entropy calculation uses the joint categorical posterior and does not assume independent pairwise errors.

Figure 1(a–b) illustrates two token embeddings. Panel (a) shows projected Z_λ/α_λ for the embedding of the true token at -2, the embedding of a competing token at $+2$, $\lambda = 0$, and prior odds $\pi_k/\pi_y = 10$. The boundary moves from 0 to $-\log(10)/4 \simeq -0.576$, enlarging the competitor's region. Dots are 240 Gaussian draws with vertical jitter used only for display. Panel (b) shows the exact pairwise posterior comparison probability from Equation 11 at three prior odds and fixed embedding distance. These are analytic channel quantities, not corpus measurements.

Frequency odds and a Zipf specialization. Let $d_{yk} = \|e_k - e_y\|_2 > 0$ and $U = \varepsilon^\top(e_k - e_y)/d_{yk} \sim \mathcal{N}(0,1)$. Expanding the Gaussian scores under $Y = y$ gives

$$\log\frac{s_k}{s_y} = \log\frac{\pi_k}{\pi_y} + r_{yk}(\lambda)U - \frac{r_{yk}^2(\lambda)}{2}. \tag{12}$$

For a more frequent competitor $\pi_k > \pi_y$, its probability of outranking y crosses one half at

$$\lambda^*_{y\to k} = \log\left(\frac{2\log(\pi_k/\pi_y)}{d_{yk}^2}\right). \tag{13}$$

A larger prior advantage moves this crossing to cleaner log-SNR. Greater separation moves it to noisier log-SNR. If $\pi_k \le \pi_y$, there is no such crossing at positive SNR. For a rarer competitor, the pairwise probability can be nonmonotonic but stays below one half, even though the full channel's conditional entropy is monotone.

For interpretation, suppose an idealized law relating rank and frequency, $\pi_{(r)} = Cr^{-\zeta}$, with $\zeta > 0$ and smaller rank indicating greater frequency. Then, for $r_k < r_y$,

$$\log\frac{\pi_k}{\pi_y} = \zeta\log\frac{r_y}{r_k}, \qquad \lambda^*_{y\to k} = \log\left(2\zeta\log\frac{r_y}{r_k}\right) - 2\log d_{yk}. \tag{14}$$

The rank ratio controls prior asymmetry and the distance controls geometric separation. This is a pairwise crossing, not an entropy midpoint or an information quantile. It does not determine the aggregate information coordinate without neighborhood geometry and joint competition. KIS uses empirical counts and does not assume or fit a Zipf law.

A.2 Candidate subset estimation and coordinate conventions

Let $S = \sum_{k\in\mathcal{V}} s_k$ and $S_A = \sum_{k\in A_K} s_k$. The omitted posterior mass is $r_{\mathrm{tail}} = 1 - S_A/S$, so

$$\ell_{\mathrm{full}} - \widetilde{\ell}_K = \log(S/S_A) = -\log(1 - r_{\mathrm{tail}}) \ge 0. \tag{15}$$

If $r_{\text{tail}} \leq \epsilon < 1$, the pointwise gap is at most $-\log(1-\epsilon)$. Evaluating that mass requires the full normalizer or an independently justified bound. Because A_K includes the true token, the losses computed with finite K do not constitute a single normalized predictor shared across all possible labels. Isotonic regression enforces monotonicity of the measured curve but does not remove its truncation bias.

The main FLM profile computes the normalizer over its full vocabulary of 30,522 tokens and uses Monte Carlo only for the expectation over Y and Gaussian noise. The reported profiles for T5 and learned embeddings use $K = 8$ with a geometric shortlist over the full vocabulary, so the top K set after adjusting for the prior is exact. An optional smaller geometric shortlist introduces another approximation before ranking by the posterior score. The implementation forms a distance matrix over the batch and vocabulary, so a small K alone does not imply lower cost for distance computation. The default Monte Carlo measurement uses 2,048 token draws from π, 16 Gaussian replicates, and 49 noise levels on $[-18, 6]$.

The exact channel becomes more informative with increasing log-SNR. After dividing the observation by α_λ, an observation at lower log-SNR can be obtained by adding independent Gaussian noise to one at higher log-SNR. Data processing therefore makes $H(Y \mid Z_\lambda)$ nonincreasing. The fitted $\widehat{H}$ enforces this order on the Monte Carlo measurements. We interpolate the fitted curve with linear segments and use the generalized inverse in equation 5, taking the leftmost crossing at a plateau value. The endpoint assignments $Q(0) = \lambda_{\text{n}}$ and $Q(1) = \lambda_{\text{c}}$ override this convention at the two boundaries.

Proposition 1 (Equal information quantiles) *If $\widehat{\tau}$ is continuous and strictly increasing on the measurement interval, the nodes $Q(i/N)$ form the unique grid, including its endpoints, with adjacent information increments $1/N$.*

The result follows from $\widehat{\tau}(Q(u)) = u$ and uniqueness of the inverse. With plateaus, a generalized inverse still specifies a convention, but the uniqueness statement no longer applies. These standard properties clarify the coordinate and do not guarantee sampling accuracy.

Proposition 2 (Scale equivariance) *For the exact token corruption model, scaling every embedding by $s > 0$ while fixing the prior gives $H_s(\lambda) = H(\lambda + 2\log s)$. If the measurement endpoints are shifted by $-2\log s$, the information quantiles obey $Q_s(u) = Q(u) - 2\log s$.*

To see this, divide Z_λ by σ_λ to obtain the equivalent observation $e^{\lambda/2} e_Y + \varepsilon$. Scaling e_Y by s is identical to shifting λ by $2\log s$. Shifting the endpoints by the same amount preserves the numerator and denominator of the normalized coordinate. A scale conversion does not identify the uncertainty curves of arbitrary embedding geometries.

ELF normalization convention. The T5 lookup table is divided by its RMS per coordinate weighted by corpus frequency, $[\sum_y \pi_y \|e_y\|_2^2/d]^{1/2}$, before measuring the profile. The source geometry label `raw_t5` therefore does not mean that the measurement uses unnormalized rows. ELF latents are first passed through the model's configured latent normalizer. For each sequence we compute RMS across valid tokens and embedding coordinates. The median across sequences is r. The lookup normalization and contextual RMS use different averaging conventions. Equation 7 uses this statistic as a scale proxy. The scale equivariance proposition does not identify the two geometries or their entropy curves.

For the ELF profiles used here, the $K = 8$ measurements give $Q_{\text{profile}}(1/2) = -4.1756$ and $r = 0.802249$ on LM1B, and $Q_{\text{profile}}(1/2) = -4.0927$ and $r = 0.8719364$ on OWT. These values map to $Q_{\text{model}}(1/2) = -3.7349$ and -3.8186, respectively, yielding $P^*_{\text{mean}} = -1.1743$ and -1.2162.

A.3 Calibration objectives and schedule disagreement diagnostics

Extending $\widehat{\tau}$ by zero below λ_{n} and one above λ_{c} makes it a cumulative distribution function. Its median $Q(1/2)$ minimizes $\mathbb{E}|\Lambda - m|$ for Λ with CDF $\widehat{\tau}$. The rule in equation 8 uses this location criterion to set the shift. It does not solve a problem of minimizing training risk. The training variant with importance correction uses a separate proposal on the original Gaussian training distribution's

$\mu_0 \pm 6\sigma_0$ interval. With the tabulated information density $\rho(\lambda) = d\widehat{\tau}(\lambda)/d\lambda$ (defined almost everywhere by the interpolation with linear segments) and original density q_0, it sets

$$q_{\mathrm{KIS}}(\lambda) = \frac{\max\{\rho(\lambda), 0.1q_0(\lambda)\}}{M}, \qquad M = \int_{\mu_0-6\sigma_0}^{\mu_0+6\sigma_0} \max\{\rho(u), 0.1q_0(u)\}\, du, \tag{16}$$

and multiplies each sampled loss by the tabulated factor q_0/q_{KIS}. The pointwise floor preserves proposal support on plateaus of ρ. It is not a convex mixture. Up to truncation to a finite window, numerical tabulation, interpolation of the inverse CDF, and weight interpolation, this approximately restores $\int q_0(\lambda)w(\lambda)R_\theta(\lambda)\, d\lambda$ over the tabulated interval. The recorded maximum weight is 10.032. We do not impose a hard clip at 10. The omitted Gaussian probability mass is below 2×10^{-9}. This is a mass statement, not a bound on omitted weighted loss. This diagnostic is distinct from the training rule that changes only location.

For sampling, let $m = N - 1$ be the number of interior nodes. Since both $a + bx_i$ and y_i are ordered for $b > 0$, the squared Wasserstein distance in one dimension between their empirical distributions with equal weights is

$$W_2^2\left(\frac{1}{m}\sum_i \delta_{a+bx_i}, \frac{1}{m}\sum_i \delta_{y_i}\right) = \frac{1}{m}\sum_i (a + bx_i - y_i)^2. \tag{17}$$

Thus equation 9 and the squared Wasserstein objective are proportional by the positive factor $1/m$ and have the same minimizer. The constraint $b \geq b_{\min}$ only enforces positive ordering. The adapter maps the fitted interior log-SNR nodes to model time and restores the original endpoints. For ELF, the exact endpoints are $t = 0, 1$, corresponding to infinite log-SNR. The finite numerical bounds $[L, U] = [-30, 15]$ are used only for the interior construction and diagnostics and are discarded before restoring the original endpoints. Preservation of gap ratios applies to interior log-SNR nodes. Nonlinear conversion to model time and restoration of endpoints need not preserve ratios between time steps. This geometric fit does not provide a bound on solver error.

The mapped grid is an input to the unchanged sampler for each model, not necessarily the denoiser's literal query grid. In ELF's SDE sampler, each nonterminal interval determines a t_{back} query specific to the solver and the amount of noise reinjected. The terminal interval uses the original ODE update. Consequently, the calibrated and original arms are a controlled comparison within the same ELF solver and budget, but their difference cannot be attributed solely to relocating deterministic integration nodes.

The diagnostic uses a feasible translation of the original interior nodes, computed separately from the affine candidate. Let $\Delta_N^{\mathrm{raw}} = (N-1)^{-1}\sum_i (y_i - x_i)$ and $\eta = 10^{-9}\max(U - L, 1)$. For a nonempty feasible translation interval, the implementation records

$$\Delta_N^\star = \mathrm{clip}\Big(\Delta_N^{\mathrm{raw}},\, L - \min_i x_i + \eta,\, U - \max_i x_i - \eta\Big). \tag{18}$$

For budget N, write $b_N = b^*$. This clips only the translation diagnostic. Across the sampling budgets listed for each system, define $d = \mathrm{median}_N\, |\Delta_N^\star|$ and $\bar{b} = \mathrm{median}_N\, b_N$. We use the following retrospective labels for the magnitude and type of schedule disagreement. The thresholds are descriptive and are not estimates of generation benefit:

1. If $|\bar{b} - 1| > 0.20$, label the case as large shape or scale disagreement.
2. Otherwise, if $d \geq 1.0$, label the case as large location disagreement.
3. Otherwise, if $d < 0.5$ and $|\bar{b} - 1| \leq 0.10$, label the schedules as closely aligned.
4. In all remaining cases, label the diagnostic as indeterminate.

These categories summarize a descriptive diagnostic separately from the candidate grid. When the slope constraint is inactive and the node variances are nonzero, $b^* = \mathrm{Corr}(x, y)\ \mathrm{sd}(y)/\mathrm{sd}(x)$. A fitted slope therefore reflects shape mismatch as well as scale, and cannot alone certify that calibration improves generation.

B EXTENSIONS OF THE INFORMATION COORDINATE

B.1 RELATION TO LANGFLOW'S GUMBEL SCHEDULER

LangFlow uses the log noise to signal ratio (log-NSR), denoted by $\gamma = \log \mathrm{NSR} = -\lambda$, and models its entropy curve with the maximum Gumbel cumulative distribution function (CDF) (Chen et al., 2026)

$$H_\gamma = H_\infty F_{\mathrm{G}}(\gamma; \mu, \beta), \qquad F_{\mathrm{G}}(x; \mu, \beta) = \exp\left\{-\exp\left[-\frac{x-\mu}{\beta}\right]\right\}, \quad \beta > 0. \tag{19}$$

Here H_∞ is the entropy scale at the noisy limit and $f_{\mathrm{G}} = \partial F_{\mathrm{G}}/\partial x$ is the Gumbel density. LangFlow uses this density in γ for training, truncated to its finite quantile bounds. LangFlow learns this curve from a sequence denoiser's cross entropy, used as a surrogate for posterior entropy. That learned quantity need not equal the conditional entropy measured by the KIS token corruption model. The algebra below relates the two scheduling methods under the Gumbel assumption, rather than asserting equality of their measured curves. The sign change is material when this construction is expressed in our log-SNR coordinate from noise to clean. Let $\lambda_{\mathrm{n}} < \lambda_{\mathrm{c}}$ be the finite noisy and clean endpoints, respectively, and define

$$A = F_{\mathrm{G}}(-\lambda_{\mathrm{n}}; \mu, \beta), \qquad B = F_{\mathrm{G}}(-\lambda_{\mathrm{c}}; \mu, \beta), \qquad D = A - B > 0. \tag{20}$$

Then the normalized information coordinate, its density, and its inverse are

$$\tau_{\mathrm{LF}}(\lambda) = \frac{A - F_{\mathrm{G}}(-\lambda; \mu, \beta)}{D}, \tag{21}$$

$$\rho_{\mathrm{LF}}(\lambda) = \tau'_{\mathrm{LF}}(\lambda) = \frac{f_{\mathrm{G}}(-\lambda; \mu, \beta)}{D}, \tag{22}$$

$$Q_{\mathrm{LF}}(u) = -\mu + \beta \log[-\log(A - uD)], \qquad 0 \le u \le 1. \tag{23}$$

Thus, a maximum Gumbel curve in γ becomes a reflected, or minimum Gumbel, information coordinate in λ. With idealized infinite endpoints, $A = 1$ and $B = 0$, so

$$\tau_{\mathrm{LF}}(\lambda) = 1 - \exp\left\{-\exp\left[\frac{\lambda+\mu}{\beta}\right]\right\}, \qquad Q_{\mathrm{LF}}(u) = -\mu + \beta \log[-\log(1-u)]. \tag{24}$$

Gumbel special case and information rate identity. The KIS information coordinate reduces exactly to Equations 21–23 only if the entropy of the token corruption model belongs to that family with two parameters on the same coordinate and endpoints. To see why this is an additional assumption, set $s = \exp(\lambda)$, let $X = e_Y$ be the clean token's fixed embedding, and consider

$$R_s = \sqrt{s}\, X + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{25}$$

Because X is a deterministic function of Y and the channel depends on Y only through X, $I(Y; R_s) = I(X; R_s)$. Under the standard assumptions for the I–MMSE identity (additive standard Gaussian noise, finite second moment of X, natural logarithms, and differentiation with respect to the scalar SNR s), the identity of Guo et al. (2005) gives

$$-\frac{\mathrm{d}}{\mathrm{d}\lambda} H(Y \mid R_{e^\lambda}) = \frac{\mathrm{d}}{\mathrm{d}\lambda} I(X; R_{e^\lambda}) = \frac{e^\lambda}{2}\, \mathrm{mmse}_X(e^\lambda). \tag{26}$$

Here $\mathrm{mmse}_X(s) = \mathbb{E}\|X - \mathbb{E}[X \mid R_s]\|_2^2$ is the total vector Bayes MSE, summed over embedding coordinates. If MSE is reported per coordinate, it must be multiplied by d on the right side of Equation 26. It is not a trained sequence denoiser's residual. Consequently, on the finite interval used by KIS,

$$\rho_{\mathrm{KIS}}(\lambda) = \frac{e^\lambda\, \mathrm{mmse}_X(e^\lambda)}{2\{H(\lambda_{\mathrm{n}}) - H(\lambda_{\mathrm{c}})\}}. \tag{27}$$

Equations 26–27 apply to the exact entropy under Gaussian noise. A posterior approximation with finite K and subsequent isotonic projection need not satisfy this derivative identity exactly. The prior and token embedding geometry determine the MMSE curve. Several distance scales that are well separated can therefore produce several transition bands, while tokens with identical embeddings can leave irreducible conditional entropy. Neither behavior is represented exactly by a single Gumbel CDF. For distinct finite token embeddings, the information rate at low log-SNR is proportional to e^λ, while confusions at high log-SNR decay exponentially in e^λ. These tails can support a reflected Gumbel approximation, but do not fix its interior shape.

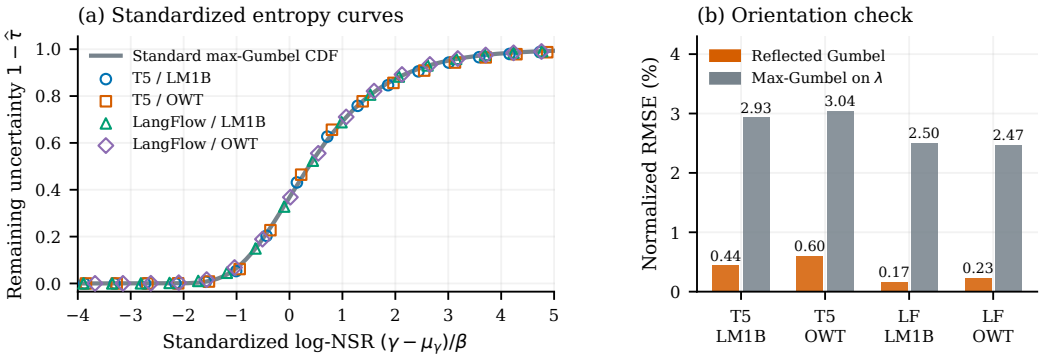


Figure 4: **Gumbel fits to normalized token uncertainty.** Left: KIS remaining uncertainty after isotonic normalization and a separate fit of location and scale to each curve, plotted against standardized log-NSR $\gamma = -\lambda$. Right: normalized root mean squared error (NRMSE) on the fitted samples for the information coordinate from the reflected family and the control with the opposite orientation. Both use two parameters and the same normalization over a finite interval. LF denotes LangFlow.

Orientation check after fitting. Figure 4 compares fits to four KIS token uncertainty profiles: T5-small with LM1B and OWT priors, and LangFlow's released embeddings with the corresponding tokenizers and priors. Each profile uses $K = 8$, 16 noise replicates, and 49 log-SNR levels. The two families are fitted independently and renormalized on $[-18, 6]$. With fitted log-NSR location μ_γ and scale β, remaining uncertainty $1-\widehat{\tau}$ at $z = (-\lambda-\mu_\gamma)/\beta$ is compared with $[F_{\mathrm{G}}(z;0,1)-B]/(A-B)$. The reflected family fits all four profiles more closely than the opposite orientation. Its maximum absolute errors at the seven interior $N = 8$ quantiles are 0.111/0.136 for T5 on LM1B/OWT and 0.034/0.055 for LangFlow on LM1B/OWT, in log-SNR. These fits after measurement describe the token corruption model. They neither identify its entropy with a contextual denoiser's loss nor establish a universal profile.

C Relation to other schedule measurements

Table 6 distinguishes the measured signal from the schedule it calibrates. It places the common KIS measurement and its two calibration rules alongside related approaches. The matched comparisons are in Section 4.3.

Table 6: Measured quantities and schedule adjustments. The KIS row describes the deployed rules in Section 3.4.

Method	Measurement	Schedule adjustment
FLM	Argmax error with a uniform prior in the one hot channel	Schedule based on decoding error for training and sampling.
LangFlow	Token cross entropy of a trained denoiser	Learned Gumbel time distribution and its quantile grid.
InfoNoise	Online denoising loss/information rate	Training noise distribution adapted with the objective's loss weight.
Entropic Time	Entropy change estimated by a denoiser	Sampling grid with equal estimated entropy increments.
Align Your Steps	Approximation error of the reverse process	Offline optimization of finite solver nodes.
KIS	Empirical priors and fixed token embedding geometry	Training shift that changes only location and positive affine sampling calibration with original endpoints.

Error, entropy, and the token prior. FLM's schedule based on argmax error (Lee et al., 2026) differs from KIS's posterior log loss. The matched ELF comparison in Table 5 changes both the prior and the uncertainty statistic, so it does not isolate either contribution. KIS on frozen FLM is a schedule intervention, not a retrained FLM comparison.

Measured information and learned loss. LangFlow (Chen et al., 2026), Entropic Time (Stancevic & Ambrogioni, 2025), and InfoNoise (Raya et al., 2026) obtain their signals from denoisers. KIS measures conditional entropy under a token corruption model with one fixed embedding per token. Appendix B gives a conditional Gumbel relation without identifying this conditional entropy with learned sequence cross entropy.

Allocation and solver constraints. KIS separates its uncertainty reference from deployment. We test training that changes only location and affine sampling against direct information allocation. Optimization that accounts for the solver (Sabour et al., 2024) and learning interactions across noise levels (Ambrogioni et al., 2026) motivate this distinction. The ablations test the constraints of each calibration rule.

D Training and inference details

D.1 Corruption coordinate and profile measurement

ELF coordinate. ELF uses the corruption path $z_t = tx_0 + (1-t)c\varepsilon$ with $c = 2$, where $t = 0$ is pure noise and $t = 1$ is clean text. Thus

$$\lambda(t) = 2\ln t - 2\ln\big(c(1-t)\big), \qquad t(\lambda) = \operatorname{sigmoid}\left(\frac{\lambda + 2\ln c}{2}\right). \tag{28}$$

The native logit-normal training density has $P_{\mathrm{mean}} = -1.5$ and $P_{\mathrm{std}} = 0.8$, corresponding to $\lambda \sim \mathcal{N}(-4.3863, 1.6^2)$. The denoiser predicts x_0. The constant $t_\epsilon = 0.05$ stabilizes the velocity conversion $v = (\widehat{x}_0 - z_t)/\max(1-t, t_\epsilon)$. It does not truncate the sampling grid. Sampling retains the exact endpoints $t = 0, 1$.

Measurement inputs and cost. The ELF calibration profile uses the frozen T5-small lookup table, with one vector of dimension 512 per token type, as a proxy for the contextual encoder latents. We normalize the table to unit power per coordinate, weighted by corpus frequency, before measurement. This normalization differs from the normalization of contextual latents used during training. Appendix A.2 specifies both conventions and the scale bridge $\lambda_{\mathrm{model}} = \lambda_{\mathrm{profile}} - 2\ln r$. The T5 profiles use $K = 8$, 2,048 token draws from the fixed corpus prior, 16 Gaussian replicates per draw, and 49 levels on $[-18, 6]$. Appendix A.2 lists the measured medians and bridged centers of the training density.

D.2 Controlled training diagnostics

The following protocols describe the controlled ELF-B training diagnostics. Generation seeds index repeated sampling from a fixed checkpoint.

Shared model and objective settings. The frozen T5-small encoder produces contextual latents, normalized to a configured latent mean of 0 and standard deviation of 0.2. ELF-B uses a bottleneck of dimension 128. Both corpora use Muon, zero weight decay, EMA decay 0.9999, and self conditioning probability 0.5. We apply the auxiliary decoder objective with probability 0.2, noise scale 5, and logit-normal parameters $(0.8, 0.8)$. The comparison that changes only location adjusts the denoiser's P_{mean} while retaining $P_{\mathrm{std}} = 0.8$, the architecture, and the other objective settings.

LM1B training. The LM1B runs use the training split padded to length 128, batch size 256, and 20,000 optimizer steps. The configured peak learning rate is 0.002 (base rate 0.001), with 1,000 warmup steps. The scale tokens for self conditioning guidance range from 0.5 to 5.0. The native arm and the arm that changes only location use $P_{\mathrm{mean}} = -1.5$ and -1.1743, respectively. The latter value follows from the bridged profile median through Equation 7.

OWT training. The OWT pair uses sequence length 1,024, batch size 32 per device, and eight steps of gradient accumulation. The base learning rate is 0.001, with 16,000 warmup microsteps. The native arm and the arm that changes only location use $P_{\text{mean}} = -1.5$ and -1.2162. Both arms are evaluated at 320,000 microsteps (40,000 optimizer updates) and later extended to 472,144 microsteps (59,018 updates).

Denoising loss evaluation. EMA weights are evaluated at the same 19 nominal log-SNR levels $\{-13, -12, \ldots, 5\}$, using the inverse map $t(\lambda)$ above and evaluation noise seed 0. At each level, the squared error of the x_0 prediction is averaged across embedding coordinates and tokens that are not padding. The primary diagnostic integrates these values over log-SNR using the trapezoidal rule, independently of either arm's training density. LM1B evaluation uses 512 sequences from its test split in batches of 32. OWT endpoint evaluation uses 256 sequences from the fixed OWT evaluation probe in batches of 16.

Generation checks after training. Both training arms are evaluated with the same native sampling grid constructed from $P_{\text{mean}} = -1.5$ and $P_{\text{std}} = 0.8$. ELF's stock SDE sampler uses churn $\gamma = 1.5$, classifier-free guidance scale 1, and self conditioning guidance scale 3. Each cell records GenPPL under GPT-2-large and mean token entropy per sample, as defined in Section 4.

D.3 Inference at a fixed checkpoint

ELF evaluation protocol. The replicated ELF-B and ELF-M OWT comparisons use released checkpoints with fixed weights, sequence length 1,024, batch size 32, and 512 samples per generation seed. Initial noise and subsequent sampler randomness are paired across the native, translation, and affine arms.

Native and projected grids. For an ELF run with N steps, the stock interior grid uses the i/N quantiles of the native logit-normal density in log-SNR. We construct its numerical CDF by trapezoidal integration on 9,001 points over $[-30, 15]$, then restore the exact endpoints in model time. Projection uses the bridged reference quantiles $Q_{\text{model}}(i/N)$ and transforms only the interior nodes. Translation uses Equation 18. Affine calibration uses Equation 9, with $b \geq 10^{-6}$. We clip the translation to its feasible interval. Both projections preserve $t = 0, 1$.

E Additional Ablations

Estimation Scalability: Fidelity of Truncated Neighborhoods. Under matched Monte Carlo draws, a compact candidate set of $K = 8$ closely approximates the posterior over the full vocabulary, yielding largest node errors in log-SNR of 0.0648 (LM1B) and 0.0918 (OWT), and coordinate RMSE of 0.00627 and 0.00856 (Figure 3c,f). The nonmonotonic error progression observed for larger K arises from empirical endpoint boundary normalization. Accordingly, we use $K = 8$ for T5-small profiles and exact full posteriors for FLM profiles with one hot encodings. These diagnostics establish the mathematical precision of the estimated information coordinates without requiring prohibitive log sum exp reductions over the full vocabulary and massive state spaces (Appendix A.2).