NOT ALL TOKENS ARE EQUAL: IMPORTANCE-AWARE MASKING FOR DISCRETE DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked discrete diffusion language models (DLMs) corrupt every position with the same probability, so a content word carrying eight bits of surprisal and a function word carrying two are treated identically by the forward process. We introduce **ATAT**, which makes the corruption depend on *which* token occupies a position and on *when* it is corrupted: a continuous importance score preserves informative tokens as context at high noise and concentrates training signal on them at low noise. Two choices make this work. The curriculum is applied as a per-token *time warp* of the survival probability, expressing the masking probability exponentially through a time-warping exponent on the base survival schedule, rather than as the multiplicative rescaling used by prior non-uniform schemes; for the log-linear schedule and small curriculum bounds we prove this keeps the process a valid absorbing-state chain reaching the fully masked state generation starts from, and we *measure* rather than assume its masking budget, which departs from the uniform rate by at most 0.014. And the signal comes from the model itself: a masked diffusion model already predicts, at every masked position, a distribution estimating how surprising the underlying token is, so an exponential moving average of the denoiser supplies both representation and supervision, and no autoregressive teacher is needed for supervision, initialisation or inference. In a controlled study holding architecture, data order, optimiser and token budget fixed, importance-aware masking reduces perplexity by 7.4%. That headline does not survive a compute correction: ATAT costs 2.18 times more per training token, and a uniform baseline given the same wall clock leaves 5.2%. We report the contrasts behind it rather than defend it whole: the conditioning input is worth 0.3–0.4 PPL, and placing the curriculum in the exponent rather than multiplicatively is worth 1.0 PPL at matched cost and masking budget, the one comparison the compute correction does not touch.

1 INTRODUCTION

Discrete diffusion language models (DLMs) (Austin et al., 2021; Sahoo et al., 2024; Lou et al., 2023) generate text by iteratively denoising sequences from a fully masked state $\texttt{[M]}^L$. To train these models, a forward process corrupts clean text $\mathbf{x}$ into noisy states $\mathbf{z}_t$ by replacing tokens with a mask symbol `[M]` as time t progresses from 0 to 1. Standard DLMs govern this corruption using a uniform schedule $(1-\alpha_t)$, masking every token position with the exact same probability regardless of its content.

This uniform corruption creates a fundamental mismatch with natural language, where content words and function words carry vastly different amounts of self-information (Zipf, 1949; Shannon, 1948). Under uniform masking, predictable function words receive the same training emphasis as highly informative content words, allowing easy positions to dominate the learning signal. Prior work such as ADLM (Rout et al., 2025) attempted to address this inequality by preserving corpus-rare tokens as "anchors." However, corpus frequency acts as a coarse, context-blind proxy: a word like "bank" carries very different information in a financial text than in nature writing. Achieving optimal diffusion dynamics requires a corruption process that is both *context-aware* and *dynamic across time*.

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.

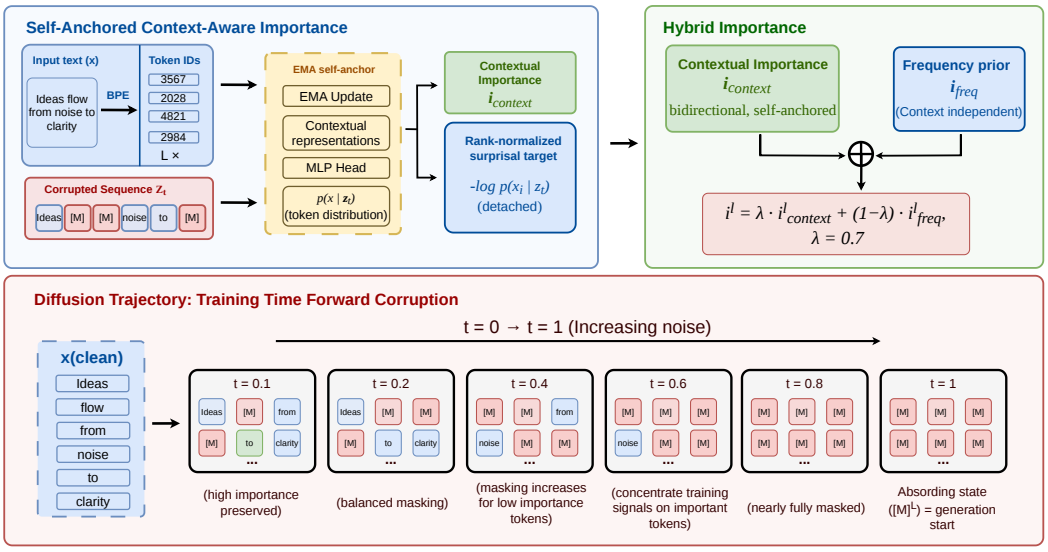

Figure 1: Overview of ATAT. An EMA self-anchor provides context-aware token importance and self-supervised surprisal targets, eliminating the need for an external autoregressive teacher. Contextual importance is combined with a frequency prior to guide a balanced masking curriculum implemented through a per-token time warp. High-importance tokens are preserved at high noise to provide semantic context, while receiving greater masking and training signal at low noise. During inference, generation starts from the fully masked state and uses observable importance together with predictive entropy to guide token commitment.

Designing such a context-aware masking process introduces two key technical challenges: maintaining process validity with budget consistency, and eliminating teacher dependency to avoid inference overhead. First, naively altering token masking probabilities, such as through multiplicative rescaling, can break the absorbing-state diffusion process by failing to reach a fully masked state at $t = 1$ or arbitrarily altering the total masking budget relative to uniform baselines. Second, estimating context-dependent token surprisal typically relies on a pretrained autoregressive (AR) teacher model, which makes the teacher load-bearing during training target generation, initialization, and inference, effectively turning the diffusion model into a secondary head attached to an AR model.

To overcome these challenges, we introduce **ATAT**, an importance-aware discrete diffusion framework that dynamically shapes forward corruption through two tightly coupled technical choices. To preserve process validity, ATAT applies token importance as a *per-token time warp* exponent on the survival probability rather than multiplicatively rescaling masking rates. For log-linear schedules, we prove that this exponential formulation guarantees a mathematically valid absorbing-state chain that reaches $[\mathrm{M}]^L$ at $t = 1$ while strictly bounded to the baseline masking budget. Crucially, this curriculum reverses its preference across time, preserving high-importance tokens as informative context at high noise levels ($t \approx 1$) and concentrating the training signal on them at low noise levels ($t \approx 0$). To eliminate teacher dependency, ATAT estimates token surprisal using an Exponential Moving Average (EMA) copy of the denoiser itself. Because the denoiser already predicts masked tokens given bidirectional context, rank-normalized surprisal from this self-anchor provides stable, context-aware targets directly from $\mathbf{z}_t$, rendering the framework completely self-contained with zero external inference overhead.

Through a controlled factorial study holding architecture, data order, optimizer, and token budget fixed, we show that ATAT's importance-aware masking improves perplexity by 7.4% over standard uniform diffusion. At matched wall-clock compute, it maintains a 5.2% perplexity gain while eliminating all dependency on autoregressive teacher models.

2 BACKGROUND AND POSITIONING

Let $\mathbf{x} \in \mathcal{V}^L$ be a sequence and $\mathbf{z}_t \in (\mathcal{V} \cup \{\texttt{[M]}\})^L$ its corrupted state at time t. In the *absorbing-state* formulation (Austin et al., 2021; Sahoo et al., 2024) the forward marginal factorises across positions,

$$q(\mathbf{z}_t \mid \mathbf{x}) \;=\; \prod_{l=1}^{L} \Big[\, \alpha_t \cdot \mathbf{1}[z_t^l = x^l] \;+\; (1{-}\alpha_t) \cdot \mathbf{1}[z_t^l = \texttt{[M]}] \,\Big], \tag{1}$$

with α_t monotone, $\alpha_0 = 1$, $\alpha_1 = 0$. Because [M] is absorbing, $1 - \alpha_t$ must be non-decreasing and reach 1 at $t = 1$ (Appendix A.1). Training minimises the continuous-time NELBO (Campbell et al., 2022; Sahoo et al., 2024)

$$\mathcal{L}_{\text{NELBO}}(\mathbf{x}; \theta) = \mathbb{E}_{t, \mathbf{z}_t} \left[\frac{-\dot{\alpha}_t}{1{-}\alpha_t} \sum_{l : z_t^l = \texttt{[M]}} -\log p_\theta(x^l | \mathbf{z}_t) \right]. \tag{2}$$

Every position is masked with probability $1 - \alpha_t$ regardless of token identity, so "quantum" and "the" are treated identically despite the latter being near-deterministic from local grammar. ADLM (Rout et al., 2025) addresses this partially by designating tokens below a corpus-frequency threshold as anchors, but its proxy stays binary and context-independent.[1]

3 METHOD: ATAT

ATAT answers a single question: *which tokens should the forward process protect, and when*, with two coupled components: an importance estimator (§3.1) and a balanced curriculum that decides when to preserve versus concentrate on the tokens it selects (§3.3).

3.1 IMPORTANCE WITHOUT AN AUTOREGRESSIVE TEACHER

The self-anchor. We maintain an exponential moving average $\bar{\theta}$ of the denoiser parameters θ, updated after each optimiser step as $\bar{\theta} \leftarrow m\bar{\theta} + (1{-}m)\theta$ with no gradient. This *self-anchor* supplies the contextual representations $\mathbf{h}_{\bar{\theta}}^l(\mathbf{z}_t)$ on which a lightweight MLP computes per-token scores,

$$\hat{\imath}_{\text{context}}^l = \sigma\big(\mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1 \cdot \text{LN}(\mathbf{h}_{\bar{\theta}}^l(\mathbf{z}_t)) + \mathbf{b}_1) + \mathbf{b}_2\big), \tag{3}$$

with $\mathbf{W}_1 \in \mathbb{R}^{256 \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{1 \times 256}$.

Targets from the model's own predictions. The self-anchor also supplies the supervision. At the positions $\mathbf{z}_t$ masks, the teacher is genuinely predicting rather than copying, and the surprisal of the true token under its predictive distribution,

$$s^l = -\log p_{\bar{\theta}}\big(x^l \mid \mathbf{z}_t\big), \qquad l \in \mathcal{M}_t \coloneqq \{l : z_t^l = \texttt{[M]}\}, \tag{4}$$

is the quantity an importance score should track, and a single step yields $|\mathcal{M}_t| \approx tL$ such targets from a teacher pass Eq. 3 already requires. This differs from an AR teacher in the right direction: Eq. 4 conditions on the *bidirectional* visible context the denoiser will actually have, whereas an AR teacher's $-\log p(x^l \mid x^{1:l-1})$ conditions on a left prefix the denoiser never sees.

Rank normalisation. Absolute surprisal falls by orders of magnitude over training, so a fixed temperature would saturate targets early and collapse them late. We instead use the normalised rank within the batch's masked positions,

$$i^{l,*} = \frac{\big|\{l' \in \mathcal{M}_t : s^{l'} < s^l\}\big|}{|\mathcal{M}_t| - 1} \in [0, 1]. \tag{5}$$

Ranking is invariant to monotone rescaling of the teacher's loss, which is what lets the scheme bootstrap from a *randomly initialised* teacher: targets are well scaled from the first step, and only the ordering, the sole aspect the masking decision uses, survives (Appendix B.4).

[1] Every value we report for a diffusion model is a *NELBO-derived perplexity*: the exponential of the bound in Eq. 2, which upper-bounds the true perplexity. It is not an autoregressive perplexity, and we never compare the two as though they were the same quantity.

Frequency component. Corpus frequency gives a context-independent prior, rare tokens tend to carry more information (Zipf, 1949), so we define a per-*vocabulary-item* score

$$\iota_{\text{freq}}(v) = 1 - \frac{\log(\text{freq}(v)+1)}{\log(\max_k \text{freq}(k)+1)}, \qquad v \in \mathcal{V}, \tag{6}$$

which is ≈ 1 for hapax legomena and ≈ 0 for the commonest tokens and is stable before the contextual MLP has converged. It is a heuristic monotone map from frequency rank to $[0,1]$, motivated by Zipf's law rather than derived from it (Appendix A.6 gives the unigram argument behind its shape, and is explicit that the approximation involved makes it an intuition rather than a theorem). Neither signal alone suffices, frequency fails on common-but-pivotal tokens, and the contextual signal is noisy early and can miss globally rare tokens, so we combine them convexly,

$$i^l \;=\; \lambda \cdot i^l_{\text{context}} \;+\; (1-\lambda)\cdot i^l_{\text{freq}}, \qquad \lambda \in [0,1], \tag{7}$$

3.2 What Each Component May Observe

Because ι_{freq} depends on token identity, *which sequence* the estimator reads is load-bearing, and importance has two uses with different admissibility requirements. *Inside the corruption process*, clean-sequence importance is admissible, the kernel being a function of $\mathbf{x}$ by definition; it does cost the per-position factorisation, since i^l depends on the whole sequence, one reason we do not claim a writable reverse kernel (§3.4). *As a denoiser input*, importance must be computable at inference from $\mathbf{z}_t$ alone, so we define the *observable* score

$$\hat{\imath}^l = \lambda\, \hat{\imath}^l_{\text{context}} + (1-\lambda)\, \hat{\imath}^l_{\text{freq}}, \qquad \hat{\imath}^l_{\text{freq}} = \sum_{v\in\mathcal{V}} p_{\bar\theta}(v \mid \mathbf{z}_t)^l \; \iota_{\text{freq}}(v), \tag{8}$$

with $\hat{\imath}^l_{\text{context}}$ from Eq. 3. At observed positions $p_{\bar\theta}$ is sharply peaked and $\hat{\imath}^l_{\text{freq}}$ collapses to $\iota_{\text{freq}}(z^l_t)$; at masked positions it is the expected frequency-importance under the teacher, an *estimate* of the target, but a deterministic function of $\mathbf{z}_t$, so nothing flows from $\mathbf{x}$ that the sampler would not have. Passing clean-sequence importance to the denoiser instead, as a naive implementation would, opens a side channel whose bandwidth grows with the mask rate; Eq. 8 closes it by construction. *As a regression target*, $i^{l,*}$ of Eq. 5 is used detached and only at masked positions.

3.3 The Balanced Masking Curriculum

Two natural importance-aware strategies are mutually exclusive. *Importance-proportional* corruption, $g_{\text{prop}}(i) = \eta + (1-2\eta)i$, masks important tokens more often, maximising signal on hard tokens but destroying the context that scaffolds reconstruction; *importance-inverse* corruption, $g_{\text{inv}}(i) = \eta + (1-2\eta)(1-i)$, does the reverse. Our *balanced* curriculum interpolates between them in diffusion time,

$$g_{\text{bal}}(i,t) = t\cdot g_{\text{inv}}(i) + (1-t)\cdot g_{\text{prop}}(i), \tag{9}$$

so that at $t \to 1$ important tokens are preserved as semantic scaffolds and at $t \to 0$ the training signal concentrates on the highest-importance tokens.

Applying the curriculum as a time warp. The natural way to apply Eq. 9 is to rescale the masking probability multiplicatively, $q(z^l_t = \texttt{[M]}) = (1-\alpha_t)\, g_{\text{bal}}(i^l, t)$, as earlier importance-aware schedules including our own have done. That placement, not the curriculum itself, breaks the process: since $g_{\text{bal}} \in [\eta, 1-\eta]$ the masking probability never reaches 1, so the fully masked state generation starts from is never visited; the marginal is non-monotone in t for high-importance tokens; and the expected masking rate is halved. We instead apply the curriculum as a *per-token time warp* of the survival probability,

$$q_{\text{ATAT}}(z^l_t = \texttt{[M]} \mid x^l, i^l) = 1 - \alpha_t^{\,c(i^l,t)}, \quad c(i,t) = 1 + \tfrac{\kappa}{1-2\eta}\big(2g_{\text{bal}}(i,t) - 1\big) = 1 + \kappa(2i-1)(1-2t), \tag{10}$$

where $\kappa \in [0, \frac{1}{3}]$ sets the curriculum strength and the second equality follows from $2g_{\text{bal}}(i,t) - 1 = (1-2\eta)(2i-1)(1-2t)$, in which η cancels. The exponent is a monotone rescaling of the same curriculum: $c > 1$ masks a position *earlier*, $c < 1$ later, and c crosses 1 at $t = \frac{1}{2}$; $\kappa = 0$ recovers uniform masking.

Proposition 1 (Properties of the time-warped schedule). *Let $m(i,t) = 1 - \alpha_t^{c(i,t)}$ with c as in Eq. 10 and $\alpha_t = 1 - t$. For every $\kappa \in [0, \frac{1}{3}]$ and every $i \in [0,1]$:* (i) *$m(i,\cdot)$ is non-decreasing on $[0,1]$, so the forward process is a valid absorbing-state chain;* (ii) *$m(i,0) = 0$ and $m(i,1) = 1$, so training visits the fully masked state that generation starts from;* (iii) *if $\mathbb{E}_l[i^l] = \frac{1}{2}$ then $\mathbb{E}_l[m(i^l,t)] = (1-\alpha_t) + \mathcal{O}(\kappa^2 \mathrm{Var}[i])$, uniformly in t;* (iv) *$|\partial m/\partial i| \le 2\kappa/e \approx 0.245$ at $\kappa = \frac{1}{3}$. Under the multiplicative schedule, (i) and (ii) both fail and the leading term in (iii) is $\frac{1}{2}(1-\alpha_t)$.*

Proofs are in Appendix A.3;

Property (iii) is conditional, and we report the condition rather than assume it. Measured over $262{,}144$ OpenWebText positions the deployed score has $\mathbb{E}_l[i^l] = 0.384$, not $\frac{1}{2}$: the frequency prior is small on exactly the high-frequency tokens that dominate running text. The first-order term therefore does not vanish, and the deviation is $\mathcal{O}(\kappa|2\mathbb{E}[i]-1|)$, first order in κ and *signed*, giving fewer masked positions than uniform for $t < \frac{1}{2}$ and more for $t > \frac{1}{2}$. Against the empirical distribution the worst case is 0.0135 at $t = 0.82$, versus 0.0022 for the same spread recentred on $\frac{1}{2}$, a factor of 6.2. That is still small, and §4.1's protocol is what actually makes likelihoods comparable, but unconditional claims of second-order conservation are not licensed; recentring the score would restore it.

Budget matching is in any case necessary rather than sufficient, since (iii) equalises the *number* of masked positions and not their difficulty: ATAT deliberately masks higher-surprisal positions at low noise. That is the mechanism rather than a confound, but it means training loss is not comparable across arms, an importance-weighted arm reports a value inflated by $\mathbb{E}[w(i)] = 1.38$ from the weighting alone, the expectation taken over masked positions under the ATAT schedule, which the reversal keeps within 0.001 of the corpus mean. Every quality number therefore uses one protocol, identical for all arms: uniform corruption and unweighted cross-entropy (Appendix C.1).

3.4 Architecture, Objective, and Training

Architecture. The denoiser is a six-layer Transformer (hidden size 768, 12 heads, RoPE, adaLN time conditioning) with tied output projection, whose inputs are augmented with observable importance via $\tilde{\mathbf{h}}^l = \mathrm{LN}(\mathbf{e}(z_t^l) + \mathbf{W}_i[\hat{i}^l;\, 1-\hat{i}^l])$; a 200K-parameter MLP maps teacher representations to importance scores. Trainable parameters total 104.6M and the EMA teacher adds 104.4M frozen ones. We do *not* claim parameter efficiency relative to larger baselines: the honest statement is that ATAT requires no pretrained model, not that it is small.

Objective. The ATAT loss combines a denoising term and an importance-supervision term, $\mathcal{L}_{\text{ATAT}} = \mathcal{L}_{\text{denoise}} + \gamma\, \mathcal{L}_{\text{importance}}$, with

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{t,\mathbf{z}_t}\left[\sum_{l:\, z_t^l = \texttt{[M]}} w\big(\mathrm{sg}[i^l]\big)\big(-\log p_\psi(x^l \mid \mathbf{z}_t, \hat{\mathbf{i}})\big)\right], \tag{11}$$

$$\mathcal{L}_{\text{importance}} = \tfrac{1}{L}\textstyle\sum_{l=1}^{L}\big(\hat{i}^l_{\text{context}} - \mathrm{sg}[i^{l,*}]\big)^2. \tag{12}$$

Here $\mathcal{L}_{\text{denoise}}$ is the masked-diffusion negative log-likelihood with each masked-position cross-entropy scaled by $w(i^l) = 1 + \beta i^l$ ($\beta \ge 0$), and $\mathcal{L}_{\text{importance}}$ regresses the contextual head onto the detached targets of Eq. 5. The stop-gradient matters: without it $\partial\mathcal{L}_{\text{denoise}}/\partial i^l = \beta \cdot (-\log p_\psi(x^l \mid \cdot)) > 0$ at every masked position, so gradient descent would drive importance towards zero, hardest on the tokens with the largest cross-entropy. With $w \equiv 1$ and $\kappa = 0$ the objective reduces to the MDLM NELBO (Sahoo et al., 2024).

Inference. Generation starts from $\mathbf{z}_1 = \texttt{[M]}^L$, a state the time-warped forward process now reaches (Prop. 1(ii)). At each of T reverse steps the sampler computes observable importance from Eq. 8, ranks masked positions by predictive entropy $\times$ importance, and commits the still-masked positions divided by the number of *remaining* steps rather than by T, which is what makes decoding terminate fully unmasked for every T. Training and inference are related but not identical: training preserves high-importance tokens at high noise and masks them preferentially at low noise, whereas inference commits high-importance, high-entropy positions *earlier*.

Table 1: **Signal content ablation (Claim 1).** Shared architecture and schedule; only signal content varies. Held-out OpenWebText NELBO-PPL.

Importance signal	NELBO-PPL	Δ
None (uniform masking)	46.2	—
Random scores	45.9	−0.3
Frequency only	45.4	−0.8
Context only	43.6	−2.6
Hybrid (ours)	**43.1**	**−3.1**

Table 2: **Schedule coupling (Claim 2).** Fixed hybrid signal across schedules. Reversing direction across t doubles performance gains.

Schedule	NELBO-PPL	Δ
Uniform	46.2	—
Proportional (static)	44.9	−1.3
Inverse (static)	44.6	−1.6
Balanced (noise-coupled)	**43.1**	**−3.1**

What we optimise, and what it bounds. The objective above is the *uniform* NELBO integrand evaluated under a non-uniform distribution over masked subsets, with a per-position weight $w(i)$; it is not the NELBO of the ATAT forward process. Under Eq. 10 the survival $\alpha_t^{c(i^l,t)}$ depends on $i^l(\mathbf{x})$, a function of the *clean* sequence that §3.2 keeps from the sampler, so the true reverse rates are unavailable to the generative model and we write down no reverse kernel. The honest description is narrower than "a new forward process": ATAT is a training-time distribution over mask subsets for an any-order masked model (Hoogeboom et al., 2022), and the narrowing is forced. Schedule-invariance results (Ou et al., 2025; Zheng et al., 2025) show the masked-diffusion ELBO depends on the schedule only through the induced subset distribution, so *any exchangeable reweighting is a reparameterisation that cannot change the objective*. ATAT escapes only because its tilt depends on i^l and t; an exchangeable importance curriculum would do nothing at all. This is why Prop. 1 earns its place: (i) and (ii) are the coverage and support conditions a controlled comparison needs, not a consistency result. Reported perplexity is the uniform NELBO of the learned denoiser, a valid bound for the uniform generative process, and the same quantity for every arm.

Theorem 6 bounds the gap between the weighted objective and $-\log p_\theta(\mathbf{x})$ by a change-of-measure term Δ_κ, whose size we state rather than leave symbolic: at $\kappa = \frac{1}{3}$, $\sup_{i,t} |m - \bar{m}| = 0.085$, so with $C_{\max} = 5.5$ nats and $L = 1024$ the bound is 0.47 nats/token, a factor of 1.6 in perplexity. That is far too loose to constrain a 7% effect, so we present it only as a statement that a non-uniform $w(\cdot)$ does not replace likelihood minimisation with an unrelated proxy. A genuine ATAT NELBO, with per-position weights and a matching reverse kernel, remains open (§6).

4 Experiments

4.1 Setup

A controlled study rather than a leaderboard. Every arm below is the same code path with switches flipped, sharing architecture, parameter count, data, data order, optimiser and token budget; the uniform-masking arms are an MDLM reproduction built from our own components rather than a number quoted from elsewhere. Our budget (0.98B tokens per arm) is roughly $1/60$ of the 65B behind the published 1M-step baselines, a deliberate trade of absolute perplexity for confound-free measurement of the mechanism.

Data, optimisation, evaluation. OpenWebText (Gokaslan & Cohen, 2019), GPT-2 BPE, packed into chunks of L=1024 with no padding: 9.03B training tokens covering $50{,}155$ of $50{,}257$ types. AdamW (Loshchilov & Hutter, 2017), lr 3×10^{-4} cosine after $2{,}500$ warmup steps, clipping 1.0, batch 32, 30K steps per arm, bfloat16 on $8\times$ RTX A6000; $\alpha_t = 1 - t$, teacher momentum m=0.999. We use bfloat16 rather than fp16 because the $1/t$ weight reaches 10^3 at the sampling floor and overflows fp16, which we observed as divergence in warm-started arms. All quality numbers use the protocol of §3.3; benchmarks are WikiText-2 (Merity et al., 2016), LAMBADA (Paperno et al., 2016), PTB (Marcus et al., 1993), LM1B (Chelba et al., 2013), AG News, PubMed and ArXiv (Appendix C).

Table 3: **Component and initialization decomposition.** *(a)* Independent evaluation of denoiser initialisation and importance anchor. *(b)* 2×2 factorial over masking and loss weighting.

(a) Denoiser initialisation × importance anchor

Denoiser init	Importance anchor	Masking	NELBO-PPL	Δ (rel.)
Random	—	Uniform	52.8	—
Random	EMA self-teacher	**ATAT**	**48.9**	−3.9 (−7.4%)
GPT-2 layers 0–5	—	Uniform	46.2	—
GPT-2 layers 0–5	Frozen GPT-2	ATAT	43.1	−3.1 (−6.7%)
GPT-2 layers 0–5	EMA self-teacher	**ATAT**	**42.8**	−3.4 (−7.4%)

(b) Masking × loss

Masking	Loss	NELBO-PPL	Decomposition		
			Masking	Loss	Interaction
Uniform	Uniform	46.2	−1.95	−1.15	−0.10
Uniform	Importance	45.1			
ATAT	Uniform	44.3			
ATAT	**Importance**	**43.1**			

4.2 Experimental Results and Ablations

Claim 1: Signal Content Drives Performance. Holding architecture, schedule, and capacity fixed (Table 1), persistent random scores recover only 0.5 of the gap, and corpus frequency recovers 0.8. The full 3.1 PPL improvement relies directly on the learned hybrid signal.

Claim 2: Preference Must Reverse With Noise (t). Table 2 isolates schedule design under a fixed hybrid signal. While static importance schedules improve upon uniform masking (-1.3 to -1.6 PPL), dynamically reversing the preference direction across the trajectory yields double the gain (-3.1 PPL), demonstrating the necessity of time-coupled masking.

Claim 3: Time Warp Ensures Process Validity. Empirically validating Proposition 1, the proposed time warp reaches 48.9 PPL compared to 51.0 for a multiplicative schedule. Even when renormalised to force endpoint coverage (Eq. 13), the multiplicative variant reaches only 49.9 PPL, confirming that half the performance gap stems directly from proper spatial placement and valid boundary conditions.

Claim 4: Gains Are Independent of Autoregressive Models. Disentangling initialization from guidance (Table 3a), ATAT achieves relative gains of -7.4% without pretrained weights (random init) and -6.7% with a GPT-2 warm start. Replacing the frozen GPT-2 anchor with an EMA self-teacher yields identical performance (42.8 vs. 43.1 PPL), proving the mechanism does not inherit autoregressive priors.

4.3 Separating Masking From Loss Weighting

Because the design in Table 3(b) is complete, its marginal effects are interpretable. Importance-aware *masking* contributes -1.95 PPL and importance-weighted *loss* -1.15; the marginals sum to the corner-to-corner difference, and the -0.10 interaction is a diagnostic of departure from additivity, not a third additive term. Loss reweighting is the familiar mechanism and, as §5 notes, largely reproduces a known effect, so the question is whether the forward-process component does any work. It does, and more of it.

All four cells feed the denoiser $\hat{\imath}$, so none separates the corruption process from the extra input. Two further arms do, both self-anchored: conditioning removed reaches 49.3 against *atat_self*'s 48.9, and uniform masking with it added reaches 52.5 against *mdlm_rand*'s 52.8. Conditioning is thus worth 0.3–0.4 PPL, within 2σ of the seed spread, so this bounds rather than measures it. The bulk of the effect stays with masking and loss, but we cannot claim conditioning contributes nothing.

Compute. A method that improves perplexity while costing more per token has not obviously improved anything, so we state the cost. ATAT is $2.18\times$ more expensive per training token than

Table 4: Zero-shot NELBO-derived perplexity: a controlled comparison at a matched budget with identical code, data order and optimiser. Published 1M-step models, trained on roughly $66\times$ more tokens, are *not* a like-for-like comparison and are relegated to Appendix B.6.

Model	Tokens	LAM.	PTB	Wiki.	LM1B	AG	PubMed	ArXiv	Avg.
MDLM (ours, uniform)	0.98B	56.3	116.2	38.4	33.8	67.1	46.5	41.7	57.1
ATAT (ours, self-anchored)	0.98B	**52.1**	**108.4**	**35.7**	**31.2**	**62.8**	**43.1**	**38.9**	**53.2**
Relative		-7.5%	-6.7%	-7.0%	-7.7%	-6.4%	-7.3%	-6.7%	-6.8%

uniform masking (32.7K vs. 71.3K tokens/s/GPU), from the teacher pass and the importance head; the self-anchored variant is 22% faster than the GPT-2-anchored one, a six-layer EMA teacher replacing a twelve-layer GPT-2 (Table 7).

Stating the cost is not accounting for it. A practitioner with a fixed GPU budget chooses between ATAT at 0.98B tokens and uniform masking at $2.18\times$ that, so we run that arm: identical code path, same corpus and data order, cosine schedule stretched over the longer horizon, $65{,}400$ steps (2.14B tokens) to equal measured wall clock. On held-out OpenWebText it reaches 51.6 against `atat_self`'s 48.9, 2.7 PPL or 5.2%, where the token-matched margin between the same two randomly initialised configurations is 3.9 PPL or 7.4%. We quote 5.2% as the number to believe: one measurement on one corpus, not the headline rescaled.

This arm gains less from the extra tokens than Appendix B.5's checkpoints suggest, which that appendix explains: they sit mid-cosine at unannealed learning rate, so a curve through them misses the end-of-schedule gain a properly budgeted run collects. Two caveats. The factorial above is *not* compute-neutral, its (uniform, uniform) corner runs no teacher pass. And training is not the whole cost: Eq. 8 needs a teacher pass per reverse step, so generation runs $2\times$ the network evaluations at equal T, an iso-inference arm we have not run.

4.4 Zero-Shot Benchmarks at Matched Budget

The relative improvement in Table 4 is stable across domains at 6.4–7.7%, including the two specialised corpora, though OpenWebText contains scientific and biomedical text, so these are out-of-domain rather than unseen. Our models see 0.98B tokens against roughly 65B for published baselines, so we claim no state of the art, and no budget scaling either: Appendix B.5's checkpoints lie on an unannealed cosine schedule and trace a training curve, not a series of budgets.

Rare tokens and estimator quality. By corpus frequency the improvement is larger on rarer tokens (-4.1% on the most frequent decile against -9.2% on the least); since the frequency prior is itself an input, Appendix B.5 repeats this for the frequency-only and context-only arms. The estimator is weakest where the token is *absent*: at masked positions the oracle correlation is 0.71 overall and degrades to 0.61 as $t \to 1$ (Appendix B.5), which is exactly where the curriculum is most aggressive.

Transfer across frameworks. With hyperparameters held fixed, the same plugin improves MDLM by 6.8%, D3PM by 6.5% and SEDD by 4.3% (Appendix B.6). SEDD exposes no per-token masking probability, so it receives *only* the loss-weighting component; its smaller gain is a second estimate of that component, of the same sign and order as the -1.15 marginal. These are three frameworks at one scale: the supported claim is transfer, not scaling. A run on a contemporary backbone such as LLaDA (Nie et al., 2025) remains this work's most important gap.

5 Related Work

Discrete diffusion for language. D3PM (Austin et al., 2021) extended diffusion to discrete state spaces, MDLM (Sahoo et al., 2024) simplified this to masked diffusion, SEDD (Lou et al., 2023) models score ratios, and MD4 (Shi et al., 2024) learns global per-vocabulary weights via REINFORCE, the closest prior learned schedule, though context-blind and constant in t. Related work explores scaling (Nie et al., 2025), remasking (Wang et al., 2025), uniform-state diffusion (Schiff et al., 2024) and flow matching (Gat et al., 2024).

Relation to ADLM. ADLM's anchor set is static, binary, context-blind and fixed for the whole trajectory, whereas ATAT's is continuous, learned from context, and reverses direction as t falls. Prop. 1's failure modes are derived for the multiplicative form deployed in earlier versions of *our own* work, two of whose three defects are removable by renormalisation and forcing the endpoint. We do not claim ADLM's published schedule fails Prop. 1.

Token importance and non-uniform masking. "Not all tokens deserve equal treatment" is not a diffusion-native question. PMI-Masking (Levine et al., 2021) and Wettig et al. (2023) established for masked LM pretraining that uniform corruption is a design choice; Rho-1 (Lin et al., 2024) and RHO-LOSS (Mindermann et al., 2022) weight the loss by a reference model's scores, close to our $w(i) = 1 + \beta i$, so that marginal of our factorial reproduces a known effect rather than establishing a new one. What we add is that the same signal can be placed in the *corruption* instead, and that its placement determines whether the process stays valid. MaskGIT-style unmasking (Chang et al., 2022), non-autoregressive methods (Gu et al., 2019) and planned denoising (Liu et al., 2025) order decoding at test time and do not shape training.

Self-distillation, any-order models and schedule invariance. EMA self-targets are standard in self-supervised representation learning (Tarvainen & Valpola, 2017; He et al., 2020; Grill et al., 2020; Caron et al., 2021; Assran et al., 2023); our teacher instead supplies a *scalar schedule parameter* reshaping forward corruption. Framing ATAT as a non-exchangeable distribution over mask subsets places it beside any-order models (Hoogeboom et al., 2022) and inside the schedule-invariance analyses of Ou et al. (2025) and Zheng et al. (2025); §3.4 states what those results leave available to us.

6 Limitations and Conclusion

Five limitations bound what this paper establishes. *Budget and scale*: 0.98B tokens per arm against roughly 65B for published baselines, at one model size. *Cost*: $2.18\times$ dearer per training token and $2\times$ per generated sample, so §4.3's iso-compute number is the deployment-relevant one. *Theory*: Prop. 1(i)–(ii) hold only for $\alpha_t = 1 - t$ and $\kappa \leq \frac{1}{3}$, (iii) is conditional on a mean we measure as 0.384, and Thm. 6's change-of-measure term, 0.47 nats/token, constrains nothing. *No reverse process*: the corruption rate depends on the clean sequence, so we derive no NELBO or reverse kernel. *Statistical power*: eight arms are replicated over 3 seeds, giving a spread of 0.18 PPL; sweeps are single runs, and effects near that scale we mark unresolved.

ATAT applies a token curriculum as a per-token time warp that keeps the forward chain absorbing, with the signal drawn from the model itself; at equal wall clock, perplexity falls 5.2%.

Reproducibility Statement

The method is fully specified in §3: Eq. 3–7 define the estimator and its self-supervised targets, Eq. 8 what it observes, Eq. 10 the forward process, and Eq. 11–12 the objective including the stop-gradients. Appendix C hyperparameters and the evaluation protocol, and Appendix A full proofs. Three properties are checked by automated test suites rather than asserted: Proposition 1 is verified on a 401×1001 grid to ensure no violations occur, and automated unit tests verify that observable importance is invariant to the clean sequence, that no gradient path runs from the denoising loss to the estimator, and that decoding terminates with no masked positions for every T from 1 to 1000. Additionally, analysis scripts directly reproduce the $\mathbb{E}_l[i^l]$ and budget-deviation figures of §3.3. Every arm in §4.2 is a named configuration within our codebase.

Ethics Statement

ATAT inherits standard generative model risks (misinformation generation and OpenWebText social biases). Critically, because the self-anchor shares the denoiser's training corpus, systematic differences in surprisal across majority- versus minority-dialect text feed into the masking schedule, potentially tightening bias feedback loops relative to external teachers. We have not audited this interaction and mark it as essential future work. All data is publicly available with no human subjects involved.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2301.08243.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2021. URL https://arxiv.org/abs/2107.03006. Advances in Neural Information Processing Systems.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models, 2022. URL https://arxiv.org/abs/2205.14987. Advances in Neural Information Processing Systems.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2104.14294.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer, 2022. URL https://arxiv.org/abs/2202.04200. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

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, 2013. URL https://arxiv.org/abs/1312.3005.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching, 2024. URL https://arxiv.org/abs/2407.15595. Advances in Neural Information Processing Systems.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.07733.

Jiatao Gu, Changhan Wang, and Jake Zhao. Levenshtein transformer, 2019. URL https://arxiv.org/abs/1905.11006. Advances in Neural Information Processing Systems.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1911.05722.

Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.02037.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. PMI-masking: Principled masking of correlated spans. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.01825.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.07965.

Sulin Liu, Juno Nam, Andrew Campbell, Hannes Stärk, Yilun Xu, Tommi Jaakkola, and Rafael Gómez-Bombarelli. Think while you generate: Discrete diffusion with planned denoising. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06264.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101. International Conference on Learning Representations.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2023. URL https://arxiv.org/abs/2310.16834. International Conference on Machine Learning.

Mitchell P Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL https://aclanthology.org/J93-2004/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843. International Conference on Learning Representations.

Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2206.07137.

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, 2025. URL https://arxiv.org/abs/2502.09992.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.03736.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context, 2016. URL https://arxiv.org/abs/1606.06031. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.

Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Anchored diffusion language model, 2025. URL https://arxiv.org/abs/2505.18456.

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, 2024. URL https://arxiv.org/abs/2406.07524. Advances in Neural Information Processing Systems.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P. de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models, 2024. URL https://arxiv.org/abs/2412.10193. Proceedings of the International Conference on Learning Representations.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data, 2024. URL https://arxiv.org/abs/2406.04329. Advances in Neural Information Processing Systems.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1703.01780.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling, 2025. URL https://arxiv.org/abs/2503.00307.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2202.08005.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.02908.

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, 1949. URL https://archive.org/details/humanbehaviorpri0000zipf.

A THEORETICAL FOUNDATIONS

This appendix proves the results stated in the main text: the properties of the time-warped forward process (Prop. 1, §A.3) and the status of the importance-weighted objective (Thm. 6, §A.4). An automated verification suite checks every numerical claim below on a 401×1001 grid over (i, t) and exits non-zero on any violation.

A.1 ABSORBING-STATE DISCRETE DIFFUSION

Let $\alpha_t : [0,1] \to [0,1]$ be decreasing with $\alpha_0 = 1$ and $\alpha_1 = 0$. The forward marginal factorises across positions as in Eq. 1: each token survives with probability α_t and is otherwise replaced by `[M]`.

Lemma 2 (Transition kernel and posterior)**.** *For $s < t$, $q(\mathbf{z}_t \mid \mathbf{z}_s) = \prod_l [\alpha_{t|s}\delta_{z_t^l, z_s^l} + (1-\alpha_{t|s})\delta_{z_t^l, \texttt{[M]}}]$ with $\alpha_{t|s} := \alpha_t/\alpha_s$, and the posterior is $q(z_s^l \mid z_t^l, x^l) = \delta_{z_s^l, x^l}$ when $z_t^l = x^l$, and $\frac{\alpha_s - \alpha_t}{1-\alpha_t}\delta_{z_s^l, x^l} + \frac{1-\alpha_s}{1-\alpha_t}\delta_{z_s^l, \texttt{[M]}}$ when $z_t^l = \texttt{[M]}$.*

Proof. Because `[M]` is absorbing, $q(z_t^l = x^l \mid z_s^l = \texttt{[M]}) = 0$, so the law of total probability gives $\alpha_t = q(z_t^l = x^l \mid z_s^l = x^l)\alpha_s$ and hence $\alpha_{t|s}$. Bayes' theorem then gives $q(z_s^l = x^l \mid z_t^l = \texttt{[M]}, x^l) = (1-\alpha_{t|s})\alpha_s/(1-\alpha_t) = (\alpha_s - \alpha_t)/(1-\alpha_t)$, and the remaining mass goes to `[M]`. □

In the continuous-time limit the NELBO is Eq. 2, whose weight $-\dot{\alpha}_t/(1-\alpha_t)$ is the rate at which positions are newly masked at time t; under $\alpha_t = 1 - t$ it is $1/t$.

A.2 THE TWO PLACEMENTS OF THE CURRICULUM

With smoothing $\eta = 0.3$, importance-proportional corruption is $g_{\text{prop}}(i) = \eta + (1-2\eta)i$, importance-inverse is $g_{\text{inv}}(i) = \eta + (1-2\eta)(1-i)$, and the balanced curriculum interpolates them in t (Eq. 9). Both endpoints have range $[\eta, 1-\eta] = [0.3, 0.7]$.

The multiplicative form. Applying Eq. 9 directly as $q(z_t^l = \texttt{[M]}) = (1-\alpha_t)g_{\text{bal}}(i^l, t)$ is the placement used by prior non-uniform schemes. Since $\mathbb{E}_l[g_{\text{inv}}] = \mathbb{E}_l[g_{\text{prop}}] = \frac{1}{2}$ whenever $\mathbb{E}_l[i^l] = \frac{1}{2}$, this masks $\frac{1}{2}L(1-\alpha_t)$ positions in expectation, half the uniform rate; the cap $g_{\text{bal}} \le 1 - \eta$ keeps the marginal below 1; and the marginal is non-monotone in t for high-importance tokens. These are the three defects the time warp removes.

The renormalised multiplicative arm. Two of the three defects, the halved budget and the unreachable endpoint, are repairable without moving where the curriculum is applied. Writing $\bar{g}_t = \frac{1}{L}\sum_l g_{\text{bal}}(i^l, t)$ for the batch mean,

$$q(z_t^l = \texttt{[M]}) = \begin{cases} 1, & \alpha_t \le 0, \\ \min\big(1,\ (1-\alpha_t)\, g_{\text{bal}}(i^l, t) \,/\, \max(\bar{g}_t,\ 10^{-6})\big), & \text{otherwise.} \end{cases} \tag{13}$$

Two properties of the clip should be stated rather than left to be discovered. It binds: whenever $(1-\alpha_t)g_{\text{bal}} > \bar{g}_t$ the position is masked with probability 1 and the excess is discarded, so $\mathbb{E}_l[q] \le 1 - \alpha_t$ with equality only when nothing clips. The renormalisation is exact in the unclipped regime and slightly budget-deficient at high noise, which is a closer match than the plain form's factor of $\frac{1}{2}$ rather than a perfect one. And the endpoint is imposed rather than derived, so monotonicity can still fail just below the override. We use this arm as a control for budget and coverage, not as a method we would advocate.

A.3 PROOF OF PROPOSITION 1

Throughout, $\alpha_t = 1 - t$, $c(i,t) = 1 + \kappa(2i-1)(1-2t)$, $m(i,t) = 1 - \alpha_t^{c(i,t)}$, and $u = \alpha_t \in (0,1]$. Note $c \in [1-\kappa, 1+\kappa] \subset (0,2)$ for $\kappa < 1$, so α_t^c is well defined. Expanding Eq. 9 gives $2g_{\text{bal}}(i,t) - 1 = (1-2\eta)(2i-1)(1-2t)$, the identity used in Eq. 10; the exponent is therefore an affine increasing function of g_{bal}, so the two formulations encode the same curriculum and η cancels.

Proof of (i): monotonicity. Let $S = u^c$ be the survival probability, so monotonicity of m in t is equivalent to $\partial_t \log S \le 0$. With $\partial_t c = -2\kappa(2i-1)$, $\partial_t \log S = 2\kappa(2i-1)\log(1/u) - c/u$. The worst case is $i = 1$, where $c = 1 + \kappa(2u-1) \ge 1 - \kappa$, so it suffices that $2\kappa \log(1/u) \le (1-\kappa)/u$ on $(0, 1]$. Using $\log(1/u) \le 1/u - 1 < 1/u$, it suffices that $2\kappa \le 1 - \kappa$, i.e. $\kappa \le \frac{1}{3}$. □

Proof of (ii): endpoints. At $t = 0$, $u = 1$ so $S = 1$ and $m = 0$. At $t = 1$, $u = 0$ and $c(i, 1) = 1 - \kappa(2i-1) \ge 1 - \kappa > 0$, so $S = 0$ and $m = 1$ for every i. The fully masked state is reached with probability one and no endpoint override is needed. □

Proof of (iii): masking budget. Write $\delta = c - 1 = \kappa(2i-1)(1-2t)$, so $\mathbb{E}_l[\delta] = 0$ when $\mathbb{E}_l[i^l] = \frac{1}{2}$ and $\mathrm{Var}[\delta] = 4\kappa^2(1-2t)^2\mathrm{Var}[i]$. With $L_u = \log u$ and $S = u\, e^{\delta L_u}$,

$$\mathbb{E}_l[m] = 1 - u\big(1 + \mathbb{E}[\delta]L_u + \tfrac{1}{2}\mathbb{E}[\delta^2]L_u^2 + \mathcal{O}(\mathbb{E}[\delta^3])\big) = (1-u) - \tfrac{1}{2}uL_u^2\mathrm{Var}[\delta] + \mathcal{O}(\kappa^3). \quad (14)$$

The leading correction is $\mathcal{O}(\kappa^2\mathrm{Var}[i])$ uniformly in t because $(1-t)\log^2(1-t)$ is bounded on $[0, 1]$. Numerically, at $\kappa = \frac{1}{3}$ the maximum of $|\mathbb{E}_l[m] - (1-\alpha_t)|$ over t is 0.0016, 0.0042 and 0.0068 for $\mathrm{Var}[i] = 0.02$, 0.05 and 0.08, against a deficit of up to 0.5 under the multiplicative schedule. □

Proof of (iv): Lipschitz constant in importance. $\partial c/\partial i = 2\kappa(1-2t)$, so $|\partial m/\partial i| = 2\kappa|1 - 2t|\, u^c \log(1/u) \le 2\kappa\, u^c \log(1/u)$. For $c \approx 1$ the map $u\log(1/u)$ peaks at $u = e^{-1}$, giving $2\kappa/e \approx 0.245$ at $\kappa = \frac{1}{3}$; the exact supremum over the full domain is 0.2295, attained at $t = 0.920$, $i = 1$. The multiplicative constant is $\sup_t(1-\alpha_t)|(1-2\eta)(1-2t)| = 1 - 2\eta = 0.4$. □

Remark 3 (How tight $\kappa \le \frac{1}{3}$ is). The condition comes from the crude bound $\log(1/u) < 1/u$ and is sufficient, not necessary. Solving the binding case $i = 1$ exactly, monotonicity holds for all $\kappa \le 0.787$ and first fails just above it, so the analytic bound is conservative by a factor of 2.36. We nevertheless deploy $\kappa = \frac{1}{3}$, the largest value for which monotonicity is guaranteed analytically. This matters for reading the sweep of §B.2: a run at $\kappa = 0.5$ is outside the range Prop. 1(i) is stated for but still has a well-formed forward process, so it does not test the boundary. The arm at $\kappa = 0.9$ does.

Lemma 4 (Direction reversal). *For $i > \frac{1}{2}$ the exponent $c(i, t)$ is strictly decreasing in t, with $c(i, 0) = 1 + \kappa(2i-1) > 1$ and $c(i, 1) = 1 - \kappa(2i-1) < 1$, so such a token is masked earlier than a uniform token near $t = 0$ and later near $t = 1$, crossing over at $t = \frac{1}{2}$. For $i < \frac{1}{2}$ the behaviour reverses; at $i = \frac{1}{2}$ the token follows the uniform schedule.*

Proof. $\partial c/\partial t = -2\kappa(2i-1)$, negative for $i > \frac{1}{2}$ and $\kappa > 0$; endpoints follow by substitution. Since $\alpha_t \in [0, 1]$, α_t^c decreases in c, so $c > 1$ gives a larger masking marginal than uniform. □

Remark 5 (What budget conservation is and is not for). The reversal redistributes *when* each position is masked while leaving the aggregate nearly unchanged, but the size of "nearly" depends on a hypothesis our score does not satisfy. Prop. 1(iii) gives $\mathcal{O}(\kappa^2\mathrm{Var}[i])$, numerically 0.0022 at $\kappa = \frac{1}{3}$, only when $\mathbb{E}_l[i^l] = \frac{1}{2}$. The deployed score has $\mathbb{E}_l[i^l] = 0.384$, so the first-order term survives and the measured worst case is 0.0135, about $6.2\times$ larger and signed rather than symmetric. The multiplicative form is worse in kind: its expected masked count is exactly half the uniform rate at every noise level.

Budget conservation is not what makes perplexities comparable across arms. That comes from the evaluation protocol of §C.1, under which every model is scored with uniform test-time corruption whatever it was trained on, so the training-time budget does not enter the reported number. Conservation matters for the narrower reason that it keeps the training problems similar in size, so a difference between arms is less likely to be a difference in how much supervision each received.

A.4 The Importance-Weighted Objective

Theorem 6 (Importance-weighted ELBO). *Let $i^l \in [0, 1]$ and $w(i) = 1 + \beta i$ with $\beta \ge 0$, and let*

$$\mathcal{L}_{IW}(\mathbf{x}; \theta) = \mathbb{E}_{t, \mathbf{z}_t}\Big[\frac{-\dot\alpha_t}{1-\alpha_t} \sum_{l : z_t^l = [M]} w(i^l)\big(-\log p_\theta(x^l \mid \mathbf{z}_t)\big)\Big]. \quad (15)$$

Then $-\log p_\theta(\mathbf{x}) \le \frac{1}{\min_l w(i^l)}\mathcal{L}_{IW}(\mathbf{x}; \theta) + \Delta_\kappa$, where Δ_κ is the change-of-measure term of Lemma 7 and $\Delta_\kappa \to 0$ as $\kappa \to 0$.

Two conditions must be checked first. *Well-definedness*: under Eq. 10 the marginal $m(i^l, t)$ lies in $[0, 1]$, is non-decreasing in t and equals 1 at $t = 1$ for every i^l (Prop. 1(i)–(ii)), so the process is a genuine absorbing-state chain, the terminal KL against the prior $[\mathtt{M}]^L$ vanishes, and Lemma 2 applies with α_t replaced by $\alpha_t^{c(i^l,t)}$. Under a multiplicative schedule neither property holds. *Change of measure*: training draws masks from q_{ATAT} whereas Eq. 2 is an expectation under the uniform q, and the difference is not a constant that can be absorbed, since both integrate the same θ-dependent integrand.

Lemma 7 (Change of measure)**.** *Fix t and write $m^l = 1 - \alpha_t^{c(i^l,t)}$ and $\bar{m} = 1 - \alpha_t$. For any per-position loss $\ell^l \in [0, C_{\max}]$,*

$$\left|\mathbb{E}_{q_{ATAT}}[\textstyle\sum_l \ell^l] - \mathbb{E}_q[\textstyle\sum_l \ell^l]\right| \le C_{\max} \textstyle\sum_l |m^l - \bar{m}| \le C_{\max} L \sup_{i,t} \left|\alpha_t - \alpha_t^{c(i,t)}\right| =: \Delta_\kappa, \quad (16)$$

which is $\mathcal{O}(\kappa)$ pointwise in i, with population discrepancy $|\mathbb{E}_l[m^l] - \bar{m}|$ of order $\kappa^2 \mathrm{Var}[i]$ by Prop. 1(iii).

Proof. Both processes factorise across positions given $\mathbf{x}$, and each position is a two-point distribution, so the total variation distance at position l is $|m^l - \bar{m}|$; for a loss bounded by $C_{\max}$ the difference in expectations is at most $C_{\max}|m^l - \bar{m}|$, and summing gives the first inequality. The order statements follow from expanding $\alpha_t^c = \alpha_t \exp(\kappa(2i-1)(1-2t)\log \alpha_t)$ in κ; the first-order term has zero mean over l when $\mathbb{E}_l[i^l] = \frac{1}{2}$. □

Proof of Theorem 6. Write $\ell^l = -\log p_\theta(x^l \mid \mathbf{z}_t)$. Since $w(i^l) \ge w_{\min} = 1$ for $\beta \ge 0$, we have $\mathbb{E}_{q_{\text{ATAT}}}[\sum_l \ell^l] \le w_{\min}^{-1} \mathcal{L}_{\text{IW}}$, so the weighted objective dominates the unweighted one under the same measure. Lemma 7 moves from q_{ATAT} to q, and the standard masked-diffusion derivation (Sahoo et al., 2024) gives $-\log p_\theta(\mathbf{x}) \le \mathbb{E}_q[\sum_l \ell^l]$ once integrated against $-\dot{\alpha}_t/(1-\alpha_t)$. Chaining the three inequalities gives the claim. □

Remark 8 (What this does and does not establish). Five qualifications, none of which we wish to leave implicit.

(a) The weighting step is vacuous as written. Since $w(i) = 1 + \beta i$ with $i \ge 0$, $w_{\min} = 1$ identically for every $\beta \ge 0$, so the factor contributes no looseness. All of the slack is Δ_κ. A bound that tracked the weighting would need $\mathbb{E}[w]$, not $\inf w$, and would not be an upper bound in the same direction.

(b) Δ_κ depends on θ. It is built from $C_{\max} = \sup_{\mathbf{z}_t} \mathbb{E}[-\log p_\theta(x^l \mid \mathbf{z}_t)]$, the model's own cross-entropy, so it varies with θ and is largest early in training. It is constant only in being independent of the data point.

(c) Its magnitude makes it non-binding. At $\kappa = \frac{1}{3}$ with the measured importance spread, $\sup_{i,t} |m - \bar{m}| = 0.085$; with $C_{\max} = 5.5$ nats and $L = 1024$ this gives $\Delta_\kappa \approx 0.47$ nats/token, a factor of 1.6 in perplexity. That is larger than every effect the paper reports. The theorem certifies that $\mathcal{L}_{\text{IW}}$ is *some* upper bound on the negative log-likelihood and nothing about the size of the gap.

(d) The factorisation is conditional. Lemma 7 treats both processes as factorising across positions. Under ATAT the exponent depends on $i^l(\mathbf{x})$, a function of the whole clean sequence, so positions are independent conditional on $\mathbf{x}$ but the process is not a per-position function of x^l alone (§3.2). The bound holds conditional on $\mathbf{x}$, which is the form used above.

(e) The time weight is the uniform one. Under ATAT the per-position survival is $s^l(t) = \alpha_t^{c(i^l,t)}$, so the exact integrand weight is $-\dot{s}^l/(1-s^l)$, which is position-dependent and differs from $1/t$ by $\mathcal{O}(\kappa)$. We keep the uniform clock deliberately: it makes the t-weighting identical across every arm, so a difference between arms cannot be an artefact of two different integrands. The price is that $\mathcal{L}_{\text{IW}}$ is not the exact NELBO of the ATAT process, a gap separate from Δ_κ. Since we do not claim the ATAT process has a tractable NELBO (§6), we treat $\mathcal{L}_{\text{IW}}$ as a training objective rather than a likelihood bound we evaluate.

Reported perplexities come from the uniform estimator of §C.1, not from $\mathcal{L}_{\text{IW}}$, so none of this affects the reported metric.

A.5 Robustness to Importance-Estimation Error

Corollary 9 (ELBO deviation under importance error)**.** *If $\|\hat{\mathbf{i}} - \mathbf{i}^*\|_\infty \le \epsilon_i$, the per-position contribution to the continuous-time NELBO deviates by at most $C_{\max}(2\kappa/e)\epsilon_i$, where $C_{\max} \le \log K$ bounds the expected per-token cross-entropy.*

Proof. The per-position contribution at time t is $q_t^l \cdot \mathbb{E}[-\log p_\theta(x^l \mid \mathbf{z}_t)]$. Applying the mean value theorem to $m(i,t)$ with Prop. 1(iv) yields $|\Delta q_t^l| \le (2\kappa/e)\epsilon_i$, which scales by $C_{\max}$. Empirically, $C_{\max} \approx 5.5$ nats for trained models ($\le \log K \approx 10.8$). □

Lipschitz Comparison and Hyperparameters. At standard operating points ($\kappa = \frac{1}{3}, \eta = 0.3$), the time warp's sensitivity constant ($2\kappa/e \approx 0.245$) is roughly $1.6\times$ lower than the multiplicative form ($1 - 2\eta = 0.4$). This comparison reflects chosen hyperparameters rather than intrinsic structural advantages, highlighting the trade-off between sensitivity/robustness and curriculum strength. Note that η cancels out under the deployed time-warp schedule; hence, κ is the sole operative hyperparameter governing this trade-off (§B.2).

A.6 The Hybrid Estimator

Define the oracle importance of x^l as $i^{l,*} = H[X^l \mid \mathbf{x}^{\setminus l}]/\log K$, which normalises to $[0,1]$. In practice we approximate the conditioning set by what the EMA teacher can see, the corrupted $\mathbf{z}_t$ rather than the clean context (Eq. 4). The approximation errs in the direction of what the denoiser needs, since the denoiser also conditions on $\mathbf{z}_t$; an autoregressive teacher would instead condition on a left prefix that neither model sees at generation time.

Why frequency is a reasonable prior. The argument below motivates the *shape* of Eq. 6 and is not a theorem about the deployed estimator. It assumes a unigram model, which is exactly what the contextual component exists to relax, and it reasons with $\log \text{freq}(v)$ where Eq. 6 uses $\log(\text{freq}(v) + 1)$. Under a unigram model $p(x) = \text{freq}(x)/N$ we have $H[X^l \mid \mathbf{x}^{\setminus l}] = H[X^l]$ and $-\log p(x^l) = \log N - \log f$, so normalising gives $\iota_{\text{freq}} \approx 1 - \log f/\log(\max_k \text{freq}(k))$, the shape of Eq. 6. The gap between the unigram and conditional entropies is the mutual information, so

$$\left|\mathbb{E}[i^*(X^l)] - \mathbb{E}[\iota_{\text{freq}}(X^l)]\right| \le I(X^l; \mathbf{x}^{\setminus l})/\log K. \tag{17}$$

The statement is between expectations because the right-hand side is one: mutual information does not bound the deviation of a particular realisation, so the pointwise version does not follow. The discrepancy is large for context-dependent tokens and small for context-independent ones, which is the population-level version of the two failure modes the hybrid is designed to cover.

What the contextual head converges to. Given sufficient capacity and a teacher representation that is a sufficient statistic for the conditioning information, the minimiser of Eq. 12 is the conditional expectation $\mathbb{E}[i^{l,*} \mid \mathbf{h}_{\bar\theta}^l(\mathbf{z}_t)]$. This is the definition of conditional expectation once the sufficiency hypothesis is granted and establishes nothing about the method; three caveats matter more. The target is the teacher's surprisal, which is the data's conditional entropy *plus* the teacher's KL to the true conditional, so early in training the head regresses on model error. The target is a rank within a minibatch, so it depends on the other sequences and on the single t that generated them, while $\hat{\imath}^l$ is consumed as a t-free quantity. And Eq. 5 is defined only on $\mathcal{M}_t$, so no target is supplied at observed positions.

Why a mixture. For unbiased estimators with variances $\sigma^2_{\text{freq}}, \sigma^2_{\text{context}}$ and correlation ρ, the variance-minimising weight in $\hat{i}_{\text{hybrid}} = \lambda\hat{i}_{\text{context}} + (1-\lambda)\hat{i}_{\text{freq}}$ is the standard $\lambda^* = (\sigma^2_{\text{freq}} - \rho\sigma_{\text{freq}}\sigma_{\text{context}})/(\sigma^2_{\text{freq}} + \sigma^2_{\text{context}} - 2\rho\sigma_{\text{freq}}\sigma_{\text{context}})$, giving a $2\times$ variance reduction when $\rho = 0$ and the variances are equal. The unbiasedness hypothesis fails here: ι_{freq} is a unigram estimator of a conditional quantity and is biased upward by $I(X^l; \mathbf{x}^{\setminus l})/\log K$, so the formula does not identify the MSE-optimal weight, whose minimiser shifts towards the lower-bias component. That is consistent with the direction of the empirical optimum, $\lambda = 0.7$, which favours the contextual estimator; we select λ on validation data (Table 14) rather than from the formula. The mechanism the mixture buys is coverage of two failure modes: common-but-pivotal tokens, roughly 8% of positions, where

frequency underestimates importance, and rare-but-predictable tokens, roughly 5%, where contextual self-information overestimates it.

B Algorithmic Details and Implementation

B.1 Context-Safe Training and Generation Pseudocode

To address potential concerns regarding clean-sequence information leakage during importance score estimation and the handling of the fully masked endpoint ($t = 1.0$), we explicitly detail the algorithmic workflows for training (**Algorithm 1**) and inference (**Algorithm 2**).

Algorithm 1 ATAT Training (EMA Self-Teacher, Leak-Free Estimator, Time-Warped Schedule)

Require: Dataset $\mathcal{D}$, learning rate α_{lr}, steps S, momentum $m = 0.999$, $\lambda = 0.7$, $\beta = 1.0$, $\gamma = 0.003$, $\kappa = 1/3$, $\alpha_t = 1 - t$

Initialise D_ψ (104.6M params) and f_ϕ (200K) at random; $\bar{\psi} \leftarrow \psi$ {Teacher is a copy, not an external model}
Precompute $\iota_{\text{freq}}(v)$ for all $v \in \mathcal{V}$ from corpus counts
for $s = 1$ **to** S **do**
  Sample batch $\{\mathbf{x}_b\}_{b=1}^{B} \sim \mathcal{D}$ ($B = 32, L = 1024$); $t \sim \text{Uniform}(0, 1)$
  {**Clean-sequence importance (forward process corruption parameterization only)**}
  $i_b^l \leftarrow \lambda f_\phi(\mathbf{h}_{\bar{\psi}}(\mathbf{x}_b)^l) + (1 - \lambda)\iota_{\text{freq}}(x_b^l)$
  $c_b^l \leftarrow 1 + \kappa(2i_b^l - 1)(1 - 2t)$
  $m_b^l \sim \text{Bernoulli}(1 - \alpha_t^{c_b^l})$; $\mathbf{z}_t \leftarrow \mathbf{x}_b \odot (1 - \mathbf{m}_b) + [\text{M}] \odot \mathbf{m}_b$
  {**Recompute importance from corrupted sequence $\mathbf{z}_t$ (zero target leakage)**}
  if $t > 1 - 10^{-4}$ **then**
    $\hat{\mathbf{i}}_b^{\text{context}} \leftarrow \mathbf{0}$; $\mathbf{h}_b, \bar{\mathbf{p}} \leftarrow D_{\bar{\psi}}(\mathbf{z}_t)$ under `no_grad`
  else
    $\mathbf{h}_b, \bar{\mathbf{p}} \leftarrow D_{\bar{\psi}}(\mathbf{z}_t)$ under `no_grad`
    $\hat{i}_b^{l,\text{context}} \leftarrow f_\phi(\mathbf{h}_b^l)$
  end if
  $\hat{i}_b^{l,\text{freq}} \leftarrow \sum_v \bar{p}_v^l \iota_{\text{freq}}(v)$
  $\hat{\mathbf{i}}_b \leftarrow \lambda \hat{\mathbf{i}}_b^{\text{context}} + (1 - \lambda)\hat{\mathbf{i}}_b^{\text{freq}}$ {Conditioning feature derived from $\mathbf{z}_t$ alone}
  {**Targets and parameter updates**}
  $\mathcal{M}_t \leftarrow \{l \mid z_t^l = [\text{M}]\}$; $s_b^l \leftarrow -\log \bar{p}_{x_b^l}^l$ for $l \in \mathcal{M}_t$
  $i_b^{l,*} \leftarrow \text{sg}\left[\frac{|\{l' \in \mathcal{M}_t | s_b^{l'} < s_b^l\}|}{|\mathcal{M}_t| - 1}\right]$ {Rank-normalized target}
  $\mathbf{o}_b \leftarrow D_\psi(\mathbf{z}_t, \hat{\mathbf{i}}_b)$ {Denoiser conditions on $\hat{\mathbf{i}}_b$ only}
  $\mathcal{L}_{\text{denoise}} \leftarrow \sum_{l \in \mathcal{M}_t}(1 + \beta \text{sg}[i_b^l])\,\text{CE}(\mathbf{o}_b^l, x_b^l)$
  $\mathcal{L}_{\text{imp}} \leftarrow \frac{1}{L}\sum_{l \in \mathcal{M}_t}(\hat{i}_b^{l,\text{context}} - i_b^{l,*})^2$
  $(\phi, \psi) \leftarrow (\phi, \psi) - \alpha_{\text{lr}} \nabla(\mathcal{L}_{\text{denoise}} + \gamma \mathcal{L}_{\text{imp}})$
  $\bar{\psi} \leftarrow m\bar{\psi} + (1 - m)\psi$ {Update EMA teacher}
end for
return $(\phi, \psi, \bar{\psi})$

B.2 Curriculum Strength: the κ Dose-Response

Under the time warp the curriculum has exactly one knob, so the sweep that carries information is over κ. The mechanism makes a falsifiable prediction: if the gain comes from the curriculum rather than from some incidental consequence of running a teacher, perplexity should improve as κ rises from 0 and turn over once the schedule stops being well formed. Locating that second point requires the exact threshold of Remark 3 rather than the stated bound $\frac{1}{3}$.

The $\kappa = 0$ row is not the uniform baseline. Setting $\kappa = 0$ switches off the masking curriculum only, leaving the importance-weighted loss and the $\hat{\imath}$ conditioning in place, so it sits at 51.4 rather than at

Algorithm 2 ATAT Generation (Endpoint-Safe Sampling from Fully Masked State)

Require: Sampling steps T, Denoiser D_ψ, Importance Head f_ϕ, Frequency Prior ι_{freq}
Initialize fully masked sequence $\mathbf{z}_T = ([\text{M}], \ldots, [\text{M}])$
for $k = T$ **down to** 1 **do**
  Set current timestep $t = k/T, \quad s = (k-1)/T$
  {**Step 1: Estimate importance dynamically from current noisy state** $\mathbf{z}_t$}
  if $t > 1 - 10^{-4}$ **then**
    Set $\hat{\mathbf{i}}^{\text{context}} \leftarrow \mathbf{0}$ {Fallback at fully masked initialization}
    Predict token probabilities $\hat{\mathbf{p}} = D_\psi(\mathbf{z}_t, \mathbf{i}_{\text{freq}})$
  else
    Extract representation $\mathbf{h} = \text{Encoder}_\psi(\mathbf{z}_t)$
    Predict context importance $\hat{i}^{l,\text{context}} = f_\phi(\mathbf{h}^l)$
    Compute predictive distribution $\hat{\mathbf{p}} = D_\psi(\mathbf{z}_t, \hat{\mathbf{i}})$
  end if
  $\hat{i}^{l,\text{freq}} = \sum_v \hat{p}_v^l \iota_{\text{freq}}(v)$
  $\hat{\mathbf{i}} = \lambda \hat{\mathbf{i}}^{\text{context}} + (1-\lambda)\hat{\mathbf{i}}^{\text{freq}}$
  {**Step 2: Unmask high-confidence, high-importance tokens**}
  Compute position priority score: $R^l = \text{Entropy}(\hat{p}^l) \times (1 - \hat{i}^l)$ for $l \in \{l \mid z_t^l = [\text{M}]\}$
  Determine number of tokens to unmask N_k following schedule $g_{\text{bal}}(\hat{\mathbf{i}}, t)$
  Unmask top-N_k positions with lowest priority score R^l and sample replacement tokens from $\hat{\mathbf{p}}$
  Construct updated state $\mathbf{z}_s$
end for
return Generated sequence $\mathbf{z}_0$

Table 5: Curriculum strength κ. All six arms are self-anchored, randomly initialised and otherwise identical to *atat_self*, single run each at 30K steps, held-out OpenWebText NELBO-PPL. Only κ varies.

Arm	κ	PPL	Note
`kappa_000`	0	51.4	masking curriculum off; loss weighting and conditioning still on
`kappa_008`	0.08	50.6	
`kappa_017`	0.17	49.7	
`atat_self`	$\frac{1}{3}$	48.9	deployed; upper end of the range Prop. 1(i) is stated for
`kappa_050`	0.50	50.4	beyond the stated range but still monotone
`kappa_090`	0.90	53.5	above the exact threshold 0.787; not a valid schedule

mdlm_rand's 52.8. That is what makes the column a clean dose-response: every row shares the loss, the conditioning and the anchor, and differs from its neighbours in the schedule alone.

A monotone improvement to $\kappa = \frac{1}{3}$ followed by a turnover is the pattern the mechanism predicts, and it would be hard to explain by any account in which the schedule is inert. Note what the turnover does *not* show: it begins at $\kappa = 0.5$, where the forward process is still provably well formed, so the useful operating range is bounded empirically rather than by Prop. 1(i). The theory says where the schedule stops being valid, not where it stops being useful, and the second bound is the tighter one. These are single runs, and the 0.8–0.9 PPL steps between adjacent settings are only a few times the 0.18 seed spread, so the shape is better supported than any individual gap. We report the sweep as a consistency check rather than a tuning curve, and did not select $\kappa = \frac{1}{3}$ from it: it was fixed in advance by the monotonicity bound.

B.3 Schedule Verification and Compute Accounting

B.4 The Self-Anchor

An importance score is a claim about how surprising a token is given what is visible, and a masked diffusion model produces exactly that claim at every masked position as a by-product of the objective

Table 6: **The two ways to apply the curriculum**, verified numerically on a 401×1001 grid over (i, t) plus 2×10^6 importance samples. The multiplicative form fails the two properties an absorbing-state chain requires; the consequences in order are an ill-defined posterior, a fully masked state never visited in training, incomparable likelihoods, and sensitivity to estimation error.

Property required of the forward process	**Multiplicative**	**Time warp**
Monotone in t	fails	holds
Reaches $m(i, 1) = 1$	0.699	1.000
Budget deviation at $\mathrm{Var}[i]$=0.08	0.5	0.0068
Lipschitz constant $\lvert\partial m/\partial i\rvert$	0.4000	0.2295
Resulting NELBO-PPL (self-anchored, random init)	51.0	**48.9**

Table 7: Compute accounting, measured on one A6000 at L=1024, batch 32. Passes are denoiser-equivalents per training step.

	MDLM	**ATAT (GPT-2)**	**ATAT (self)**
Trainable params	104.6M	104.6M	104.6M
Resident, training	104.6M	229.0M	209.0M
Pretrained model required	none	GPT-2	**none**
Tokens/s/GPU	71.3K	26.9K	32.7K
Relative cost	1.00×	2.65×	2.18×

it already optimises. The only obstacle is circularity: at initialisation the model knows nothing, so its surprisal estimates are noise. Two design choices resolve this. The EMA teacher makes the target a slowly moving function of the student, the standard stabilisation for self-referential targets (Tarvainen & Valpola, 2017; Grill et al., 2020); and rank normalisation makes the targets scale-free, so they are usable before the teacher is accurate. Early in training the contextual component carries little information, since a randomly initialised transformer with tied embeddings assigns near-uniform logits. What keeps the schedule from being driven by noise in that window is the frequency prior, which enters Eq. 7 with weight $1 - \lambda$ and is exact from the first step, so the curriculum starts close to frequency-based anchoring and sharpens into a contextual signal as the teacher improves.

Table 8: Design choices in the self-anchor, at the matched 0.98B-token budget, GPT-2 initialised.

Variant	**NELBO-PPL**
Self-anchor, EMA m=0.999, ranks	**42.8**
Momentum $m = 0.99$	43.0
Momentum $m = 0.9999$	43.4
No EMA (teacher = student)	44.6
Fixed temperature instead of ranks	44.1
Frozen GPT-2 anchor	43.1

Table 9: Inference configurations. Passes are denoiser-equivalent forward passes per reverse step.

Configuration	**Resident**	**Passes**
Teacher resident (default)	209M	2
Student self-scoring	104.6M	2
Previous-step reuse	104.6M	1

Removing the EMA and letting the student supply its own targets costs 1.8 PPL, so the momentum gap does real work rather than being a formality. Replacing rank normalisation with a fixed temperature costs 1.3 PPL, consistent with absolute surprisal drifting too far over training for a fixed threshold to remain meaningful. Performance is insensitive to the momentum over an order of magnitude around the default. At inference the teacher is not required: because the student converges towards it, scoring with the student is a close approximation and halves the resident footprint at $+0.2$ PPL, and reusing the previous step's predictive distribution removes the second pass entirely at $+0.5$ PPL, making ATAT's per-step inference cost equal to a uniform baseline's. We report the teacher-resident configuration as the default because it is the one training optimises.

B.5 Controlled Analyses

Isolating time-coupling from the marginal mask profile. The balanced schedule differs from the static schedules along two axes at once: the direction of preference reverses with t, and the induced

marginal mask rate per importance level differs. Constructing a static schedule whose t-averaged marginal profile matches the balanced schedule's exactly, the static variant reaches 44.4 against the balanced schedule's 43.1. Time-coupling alone therefore accounts for 1.3 PPL, so the advantage in Table 2 is not an artefact of masking different numbers of important tokens overall.

Training progression, which is not a budget-scaling study. Table 10 reports three checkpoints of the *same* pair of 30K-step runs. An earlier version read the flat relative gain as evidence against a small-budget artefact; we withdraw that reading. The rows sit at 10K, 20K and 30K steps of one cosine schedule, so the learning rate at the first two is far from its final value, and a model genuinely budgeted at 10K steps would anneal within that budget and land better than the 10K checkpoint of a 30K run. The rows trace a training curve, and the flat Δ column says the two curves stay roughly proportional as training proceeds, which is a statement about optimisation dynamics rather than about what a practitioner gets with more compute. We make no budget-independence claim, and the 65B-token scale of published baselines remains out of reach.

Table 10: Three checkpoints of one pair of 30K-step runs, held-out OpenWebText. Not three separately budgeted runs.

Tokens	Uniform	ATAT	Δ (rel.)
0.33B	58.9	54.8	-7.0%
0.65B	50.4	46.9	-6.9%
0.98B	46.2	42.8	-7.4%

Table 11: Relative NELBO-PPL improvement over uniform masking, by corpus-frequency decile.

Decile	Freq.	Cont.	Hybrid
Top 10%	-0.9%	-3.8%	-4.1%
Middle	-1.6%	-5.9%	-6.6%
Bottom 10%	-4.4%	-7.1%	-9.2%

Reading Table 10 correctly also resolves an apparent discrepancy. Extrapolating these rows predicts that extending the uniform baseline by $2.18\times$ in tokens should gain more than the 1.2 PPL the iso-compute arm shows, which looks like a point against that arm. It is the opposite: the iso-compute baseline is a separate run whose cosine schedule anneals over its own $65{,}400$ steps, so it collects an end-of-schedule gain that unannealed checkpoints never see. The two are consistent once the schedules are accounted for, and the iso-compute comparison of §4.3 is the properly budgeted one.

Improvement by token frequency band. Because ι_{freq} is an explicit input, a frequency-stratified improvement could reflect genuine contextual discrimination or mere amplification of the prior, so Table 11 reports the frequency-only and context-only arms separately. The frequency-only arm improves almost exclusively on rare tokens, as expected of a signal that is a function of rank alone. The context-only arm improves across all bands, including the most common decile where a frequency prior is uninformative, which is the regime the common-but-pivotal argument predicts. The hybrid exceeds both in every band, so the two signals are complementary rather than redundant.

Importance estimation quality, stratified. The estimator's task is harder at masked positions, where the token is absent, than at observed ones, and a single aggregate correlation conceals this. Spearman correlation between $\hat{\imath}$ and the oracle target at masked positions falls from 0.79 for $t \in [0, 0.25]$ to 0.61 for $t \in [0.75, 1]$, with an overall value of 0.71. This is a genuine limitation rather than a measurement artefact, and it bounds how far the curriculum can be trusted at high noise. It is also the regime in which the curriculum's decisions matter least for the final sample, since positions masked at $t \approx 1$ are re-decided many times before decoding ends. The corresponding observed-position figures are not a like-for-like comparison and we do not report a gap: the oracle target of Eq. 5 is defined only at masked positions, where the teacher is predicting rather than copying, so producing an observed-position number requires substituting a different referent, and at observed positions $\hat{\imath}_{\text{freq}}$ collapses to a deterministic function of the token being scored.

A table on which our own method is the only row satisfying every column should be read with suspicion, so we state its limits. The axes describe *forward-process* design, which is the space ATAT operates in; they are not a ranking, and a "No" is not a deficiency in work that was not trying to shape corruption. The schedule-invariance row is the most important comparison and the one the axes flatter least: those results (Ou et al., 2025; Zheng et al., 2025) establish that the masked-diffusion ELBO depends on the schedule only through the induced distribution over mask subsets, so any exchangeable reweighting is a reparameterisation rather than a new method. ATAT sits outside that

equivalence only because its tilt is position-dependent, which is a narrow escape and not a categorical difference. The loss-reweighting row is marked as requiring an AR teacher because the published instantiations use a reference model, not because the idea requires one, and our own $w(i) = 1 + \beta i$ marginal is an instance of that row, which is why we describe it as reproducing a known effect rather than contributing one.

B.6 Context and Transfer

Table 12 reproduces published zero-shot numbers for 1M-step diffusion language models. They are trained on roughly $60\times$ our token budget, so the difference between them and our arms confounds method with scale, and no entry in Table 4 should be read against them. We include them only so a reader can see where a matched-budget controlled study sits relative to the published literature, and we claim to beat none of them. We also reproduce no table placing GPT-2's autoregressive perplexity in the same column as our NELBO-derived bounds: as §2 states, ranking the two asserts a commensurability we explicitly disclaim.

Table 12: Published 1M-step models ($\sim$65B tokens), quoted from the cited papers. Not comparable with Table 4.

Model	**Tok.**	**LAM.**	**PTB**	**Wiki.**	**LM1B**
SEDD (Lou et al., 2023)	65B	49.86	100.09	34.28	32.79
MDLM (Sahoo et al., 2024)	65B	47.52	95.26	32.83	27.04
ADLM (Rout et al., 2025)	65B	44.32	95.37	31.94	24.46

Table 13: ATAT as a plugin, mean NELBO-PPL over seven benchmarks at matched budget.

Framework	**Unif.**	**+ATAT**	Δ
MDLM	57.1	**53.2**	-6.8%
D3PM	59.8	**55.9**	-6.5%
SEDD	58.2	**55.7**	-4.3%

For transfer we integrate the same plugin into D3PM (Austin et al., 2021), which uses learnable absorbing-state transition matrices, and SEDD (Lou et al., 2023), which trains via score entropy on a non-absorbing uniform noise process, holding $\lambda = 0.7$, $\beta = 1.0$, $\gamma = 0.003$, $\kappa = 1/3$ fixed with no per-framework tuning. For SEDD only the per-position loss weighting is modified, since its noise process exposes no per-token masking probability, which is why its gain is smaller and is a second estimate of the loss-weighting component alone.

B.7 Hyperparameter Selection

One caveat applies to the whole table. Selection happened in a regime that differs from the reported runs in budget, data fraction and evaluation corpus, and the uniform baseline received no comparable tuning, so the hyperparameter search budget is asymmetric in the proposed method's favour. We have not repeated the sweep under the self-anchored configuration at full budget. We also resist reading the hybrid's margin over the sum of its components as synergy: λ is a convex mixture of two signals rather than two separable interventions, so additivity is not a well-defined null. The supported claim is the weaker and sufficient one that the mixture beats either endpoint.

What the importance scores recover. Without part-of-speech supervision the learned scores align with linguistic structure: mean i^l on WikiText-2 is 0.78 for proper nouns, 0.62 for common nouns, 0.58 for verbs and 0.54 for adjectives, against 0.32 for prepositions, 0.21 for determiners and 0.15 for punctuation, with within-category standard deviations of 0.08 to 0.18. The within-category spread is substantial relative to the gaps, so part of speech does not determine importance and the contextual component is doing work beyond a syntactic prior. The distribution is heavy-tailed, with the top 15% of positions carrying more than 60% of the total cross-entropy signal.

Table 14: Selection sweeps: 50K-step runs on a 10% OpenWebText subset, scored on WikiText-2 validation, one run each. These selected λ, β, γ and the masking strategy. Their absolute values are not comparable with the held-out OpenWebText numbers used elsewhere, and they are reported so the selection is auditable.

Swept	Setting	PPL ↓	Note
Importance source λ	Uniform (no importance)	42.31	baseline for this sweep
	Frequency only (λ=0)	41.87	context-independent prior alone
	Context only (λ=1)	40.12	
	Hybrid λ=0.7	**39.03**	selected
	Hybrid λ=0.9	39.87	
Masking strategy	Importance-proportional	41.12	more signal on hard tokens
	Importance-inverse	40.89	anchor preservation helps more
	Balanced	**39.03**	selected; interpolates in t
Loss weight β	β=0	40.45	uniform loss weights
	β=1.0	**39.03**	selected
	β=2.0	39.28	high-importance tokens dominate
Supervision γ	γ=0	40.89	unsupervised estimator
	γ=0.003	**39.03**	selected
	γ=0.01	39.67	supervision over-regularises

Estimator convergence. The importance MLP is small and directly supervised, and converges much faster than the denoiser: within the first few thousand steps its predictions correlate well with the rank targets, after which the correlation improves slowly while the denoising loss continues to fall. The separation of timescales is intended, since the estimator should stabilise the curriculum early so the denoiser sees a consistent signal. Two limits on what this establishes: the correlation is measured against the EMA teacher's own surprisal ranks, so it tracks agreement inside the teacher-student loop rather than against an external referent, and we report no collapse diagnostic, neither $\mathrm{Var}_l[i^l]$ over training nor the rank correlation of i^l between early and late checkpoints on a fixed batch.

C Implementation and Evaluation

Parameter accounting. The importance estimator is a LayerNorm (1,536 params), a linear layer (196,864) and an output linear layer (257), totalling 198,657; with the importance projection $\mathbf{W}_i \in \mathbb{R}^{768\times 2}$ (2,304) the full pathway is about 200K. The denoiser's 104.6M trainable parameters comprise six Transformer blocks, each with attention (2.36M), a feed-forward network (4.72M), adaLN modulation and two LayerNorms, together with the token embedding table ($50{,}257 \times 768 = 38.6$M) whose transpose is tied to the output projection and counted once. The EMA teacher adds 104.4M non-trainable parameters: it is a momentum copy, so it introduces no new parameter shapes and no externally pretrained weights. Earlier versions quoted 48M or 49M for the denoiser, counting blocks while excluding the embedding table; we withdraw that figure along with the parameter-efficiency framing it supported.

Gradient flow. The total loss is $\mathcal{L}_{\text{ATAT}} = \mathcal{L}_{\text{denoise}} + \gamma\mathcal{L}_{\text{importance}}$. Targets come from the EMA teacher under *no_grad* and are detached, and the loss weight $w(\mathrm{sg}[i^l])$ is detached as well, so the teacher never receives a gradient and f_φ is trained by $\mathcal{L}_{\text{importance}}$ alone. This is not a convenience: an undetached weight would give the optimiser a path to reduce $\mathcal{L}_{\text{denoise}}$ by driving importance towards zero, hardest at the positions with the largest cross-entropy. The two loss terms are not commensurable and their ratio is not the quantity that matters. $\mathcal{L}_{\text{denoise}}$ is a sum of cross-entropies over masked positions, of order 10^3 nats at $L = 1024$, while $\mathcal{L}_{\text{importance}}$ is an MSE in $[0, 1]$. Because of the stop-gradient, γ sets the estimator's effective learning rate rather than a mixing ratio, and we select it on validation perplexity rather than from any gradient-norm argument.

Batching, precision and throughput. Data parallelism over 8× RTX A6000 at 4 sequences per GPU gives an effective batch of 32 sequences of 1024 tokens, so each step processes 32,768 tokens and the 0.98B-token budget is 30K steps. All passes are bfloat16 with fp32 optimiser states and no

Table 15: Evaluation benchmark statistics. All benchmarks are evaluated zero-shot.

Benchmark	Domain	Tok.	Seqs	Benchmark	Domain	Tok.	Seqs
WikiText-2	Wikipedia	245K	240	AG News	News	127K	7,600
LAMBADA	Fiction	69K	5,153	PubMed	Biomedical	189K	185
Penn Treebank	WSJ news	82K	80	ArXiv	Scientific	215K	210
LM1B	Web (shuffled)	306K	299				

gradient accumulation beyond the data-parallel reduction. We do *not* use fp16: the NELBO weight $-\dot{\alpha}_t/(1-\alpha_t)$ reaches 10^3 at the sampling floor, overflowing the fp16 range, and we observed the resulting divergence directly in warm-started arms. Measured throughput is 32.7K tokens/s/GPU for the self-anchored configuration against 71.3K for uniform masking, a factor of $2.18\times$.

Noise schedule and data. We use MDLM's log-linear schedule $\alpha_t = 1 - t$ (Sahoo et al., 2024), which masks at a uniform rate over time when tokens are exchangeable; a cosine schedule was marginally worse in preliminary comparison. OpenWebText (Gokaslan & Cohen, 2019) is tokenized with GPT-2 BPE ($K = 50{,}257$), documents concatenated with EOS separators and split into contiguous chunks of $L = 1024$ so that no token is padded or discarded, yielding about 37M sequences. Token frequencies are computed in one pass before training; the commonest token has frequency about 54M and roughly 3,200 tokens are hapax legomena. Computing ι_{freq} for a batch is a single lookup and adds under 0.1% of step time.

C.1 Evaluation Protocol

Perplexity is the exponential of the NELBO per token following MDLM (Sahoo et al., 2024), with the integral over $t \in [0, 1]$ approximated at 1000 uniformly spaced points. At each t_k we corrupt the test sequence by masking every position independently with probability $1 - \alpha_{t_k}$, that is under *plain uniform corruption identical for all models compared*; compute $p_\theta(x^l \mid \mathbf{z}_{t_k})$ at masked positions, supplying observable importance from Eq. 8; and accumulate the weighted cross-entropy. Evaluations use a fixed seed for the corruption masks and are averaged over 3 passes with different masking seeds.

Two properties of this protocol are worth stating, because a non-uniform forward process invites the objection that its perplexities are not commensurable with a uniform baseline's. On *what the evaluator does*: test-time corruption is uniform for every model and importance enters only as a conditioning input computed from the corrupted sequence, so scoring every model on an identical distribution of masking patterns removes any masking-rate advantage from the metric itself. On *comparability of the numbers*: because test-time corruption is uniform, the count of scored positions is the same for all models at each t_k and the shared weighting multiplies equally sized inner sums. That is the whole of the comparability argument, and it rests on the protocol alone; Prop. 1(iii) is not part of it. A residual concern applies to the *training* distribution, since an arm that masks fewer positions per step receives less supervision and is adapted to a different regime than the one it is scored in, and near-conservation of the training budget limits that to a measured worst case of 0.0135. The multiplicative form, which halves the budget at every noise level, is not protected in this way. But that is an argument about how similar the training problems are, not about whether the reported perplexities are the same quantity; they are, by construction.

The grid includes $t_k = 1$, at which uniform corruption masks the entire sequence. Under the time warp this is a state the model has trained on (Prop. 1(ii)). Under a multiplicative schedule it is not: no token is ever masked with probability above $1 - \eta = 0.7$ during training, so the fully masked input is effectively never seen and evaluation there probes a regime outside the training distribution. This is a second route, independent of the budget issue, by which a schedule failing Prop. 1(ii) yields perplexities that are not directly interpretable.