UNIFYING AUTOREGRESSIVE AND DIFFUSION-BASED SEQUENCE GENERATION

Nima Fathi[*], **Torsten Scholak & Pierre-André Noël**
ServiceNow Research
nima.fathi@mila.quebec
{torsten.scholak,pierre-andre.noel}@servicenow.com

ABSTRACT

We present significant extensions to diffusion-based sequence generation models, blurring the line with autoregressive language models. We introduce *hyperschedules*, which assign distinct noise schedules to individual token positions, generalizing both autoregressive models (*e.g.*, GPT) and conventional diffusion models (*e.g.*, SEDD, MDLM) as special cases. Second, we propose two *hybrid token-wise noising processes* that interpolate between absorbing and uniform processes, enabling the model to fix past mistakes, and we introduce a *novel inference algorithm* that leverages this new feature in a simplified context inspired from MDLM. To support efficient training and inference, we design attention masks compatible with KV-caching. Our methods achieve state-of-the-art perplexity and generate diverse, high-quality sequences across standard benchmarks, suggesting a promising path for autoregressive diffusion-based sequence generation.

1 INTRODUCTION

Generative diffusion models, primarily recognized for their impressive image generation performance in continuous domains (Yang et al., 2023), are rapidly gaining traction in language modeling, a discrete domain historically dominated by autoregressive (AR) models such as the GPT family (Radford et al., 2019; Brown et al., 2020). Contrary to the perceived visceral separation between autoregressive and diffusion models, and despite their distinct historical development, this work reveals a fundamental connection: autoregressive models are, in essence, a form of diffusion.

The core principle behind diffusion models involves *prescribing* a "noising" process that gradually destroys information in training data samples, subsequently *learning* a neural network that progressively generates new samples from "pure noise" with a denoising process. The noising process acts as a form of data augmentation: together with the original training dataset, it specifies the *curriculum* on which the generator (denoiser) is trained. Part of the attraction for these models arises from their rich theoretical grounding, resulting in concrete practical techniques. In particular, a model's training and inference environments can be decoupled, allowing for a compute-budget knob at inference time, and guidance techniques adapting a model's behavior to specific situations.

Despite the common use of Gaussian noise in continuous diffusion, the underlying principles can be adapted to discrete state spaces (Austin et al., 2021; Zhou et al., 2023; Lou et al., 2024). Common practices for sequence generation have the noising process randomly and independently substituting some original tokens by completely unrelated ones, *i.e.*, *uniformly* sampled tokens or a special "mask" *absorbing* state. A noise schedule determines token replacement probabilities at different points in the curriculum. The resulting sequence at the schedule's highest noise level retains no mutual information with the original sequence. The generator is then trained on this curriculum to enable the production of novel sequences.

This work unifies AR and diffusion sequence generation by introducing *hyperschedules*, allowing different positions in the sequence to be affected by different noise schedules. We establish that autoregressive models, such as GPT, can be understood as diffusion models without data augmen-

[*]Also at Québec Artificial Intelligence Institute (Mila) and McGill University.

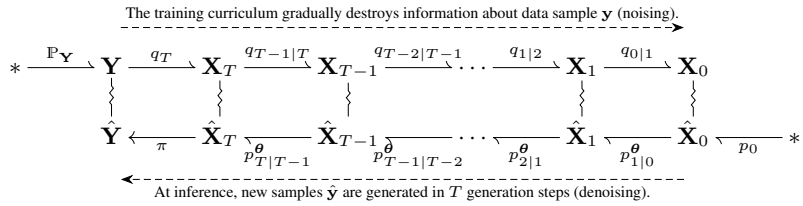


Figure 1: Generative diffusion models *prescribe* (through $q_{t|t+1}$) a curriculum process $\{\mathbf{X}_t\}$, then *learn* (through $p^{\boldsymbol{\theta}}_{t+1|t}$) a reverse process $\{\hat{\mathbf{X}}_t\}$ so that the marginal distributions match at each step t (vertical squiggly lines). $\mathbf{Y}$ is the training dataset and $\hat{\mathbf{Y}}$ is the generated output. This work focuses on discrete diffusion for sequence generation: our $\mathbf{Y}$, $\hat{\mathbf{Y}}$, $\mathbf{X}_t$ and $\hat{\mathbf{X}}_t$ are all sequences of discrete tokens, using the identity for both q_T and π. We show that standard autoregressive models (*e.g.*, GPT) are an extreme case of this framework, a unification enabling a vast continuum of new diffusion models, including autoregressive-like ones.

tation, utilizing a discrete noise schedule comprising only "full noise" and "no noise" levels. This unification expands model design space and enables a variety of generalized AR-like approaches.

Recent (Sahoo et al., 2024; Ou et al., 2024; Shi et al., 2024) and concurrent (Liu et al., 2024; Kim et al., 2025; Peng et al., 2025; Nie et al., 2025; Wang et al., 2025; Arriola et al., 2025) works have focused on mask diffusion models (MDMs). By specializing on the "absorb" noising mechanism, these MDMs enable great simplifications over a general-purpose treatment: neural networks no longer need an explicit noise-level dependency, and a more standard loss function can be used. However, our work shows that the same feats and simplifications can be accomplished in a general, non-MDM case. Motivated by the same rationale that led concurrent MDMs to conceive "remasking" strategies, we introduce hybrid noising processes, interpolating between the "absorb" and "uniform" processes to combine the benefits of both and achieve state-of-the-art performances, with aspects further improved by our novel adaptive correction sampler (ACS) inference algorithm. Our hyperschedule-equipped approach also supports KV-caching and efficient training.

In summary, our main contributions are:

- we unify AR and diffusion sequence generation by introducing *hyperschedules*;
- we consider hybrid noising processes, reaping benefits from both leading noising processes and achieving state-of-the-art performances, with and without our novel ACS algorithm;
- our hyperschedule-powered hybrid processes generalize multiple concurrent developments to non-MDM setting, including efficient training and KV-caching.

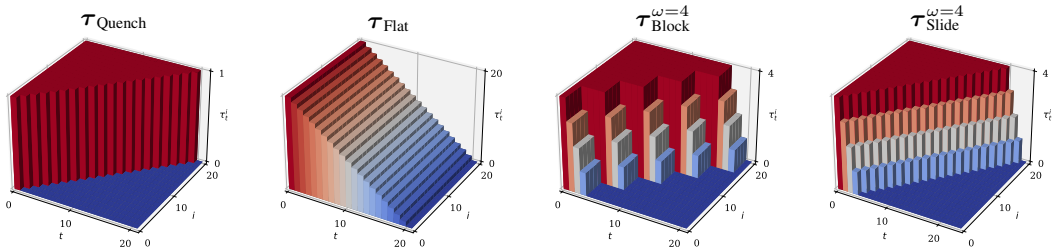
2 Abstract Sequence Generators

In this section, we reconcile autoregressive and diffusion-based sequence generation models by abstracting-out their respective implementation details, instead emphasizing their shared essence. Note that we focus on unconditional generation without loss of generality; conditional generation (*e.g.*, prompting) may be recovered as a special case.

2.1 Generative Diffusion Models

For our present purpose, a *generator* is a procedure that yields an output $\hat{\mathbf{y}}$ according to a certain probability distribution $\mathbb{P}_{\hat{\mathbf{Y}}}$. We focus on procedures composed of $T \in \mathbb{N}^*$ discrete steps (*e.g.*, "calls" to a neural network) each updating a *state* $\hat{\mathbf{x}}_t$ to a state $\hat{\mathbf{x}}_{t+1}$ using a conditional probability distribution $\mathbb{P}_{\hat{\mathbf{X}}_{t+1}|\hat{\mathbf{X}}_t}(\hat{\mathbf{x}}_{t+1}|\hat{\mathbf{x}}_t) = p^{\boldsymbol{\theta}}_{t+1|t}(\hat{\mathbf{x}}_{t+1}|\hat{\mathbf{x}}_t)$ with learned parameters $\boldsymbol{\theta}$. The stochastic process $\{\hat{\mathbf{X}}_t\}$ is seeded with an *initial state* $\hat{\mathbf{x}}_0$ that is either a constant or sampled from a provided (*i.e.*, not learned) distribution $\mathbb{P}_{\mathbf{X}_0}(\mathbf{x}_0) = p_0(\mathbf{x}_0)$. The output $\hat{y} = \pi(\hat{x}_T)$ is a deterministic function of $\hat{x}_T$. Given a dataset sampled from a data distribution $\mathbb{P}_{\mathbf{Y}}$, our goal is to train the parameters $\boldsymbol{\theta}$ so that the generator's marginal output distribution $\mathbb{P}_{\hat{\mathbf{Y}}}$ matches the data's $\mathbb{P}_{\mathbf{Y}}$ (hereafter noted $\hat{\mathbf{Y}} \text{–}\!\!\sim\!\!\text{–} \mathbf{Y}$).

In general, generative diffusion models Yang et al. (2023) *prescribe* the evolution $\mathbb{P}_{\mathbf{X}_t|\mathbf{X}_{t+1}}(\mathbf{x}_t|\mathbf{x}_{t+1}) = q_{t|t+1}(\mathbf{x}_t|\mathbf{x}_{t+1})$ of a *curriculum* $\{\mathbf{X}_t\}$ conditional on a data sample $\mathbf{y}$ ac-



(a) Quenched (AR). (b) Flat annealing. (c) Block annealing. (d) Slide annealing.

Figure 2: We introduce $\boldsymbol{\tau}$*-hyperschedules*, subjecting different token positions i with different noise levels (red high; blue low) at different generation step t. (a) Standard AR models (*e.g.*, GPT) determine tokens one by one, "quenching" each of them to full determination in a single step – they are an extreme case of a diffusion model. (b) Standard diffusion models (*e.g.*, SEDD) gradually anneal all tokens independently of their position. (c) Block-wise application of flat annealing, here for blocks of width $\omega = 4$. (d) Annealing with a sliding window ("smoothed" AR), here using window width $\omega = 4$. These last two examples share important features of both AR and diffusion models.

cording to $\mathbb{P}_{\mathbf{X}_T|\mathbf{Y}}(\mathbf{x}_T|\mathbf{y}) = q_T(\mathbf{x}_T|\mathbf{y})$.[1] Different training strategies have been developed to align the marginal distributions $\hat{\mathbf{X}}_t \rightsquigarrow \mathbf{X}_t$ at every step t. In a sense, the curriculum $\{\mathbf{X}_t\}$ breaks the goal $\hat{\mathbf{Y}} \rightsquigarrow \mathbf{Y}$ into T simpler steps. Figure 1 provides a diagrammatic summary.

2.2 Sequence Generators

This work focuses on datasets of sequences $\mathbf{y} = (y^0, y^1, \cdots, y^{d-1}) \in \mathcal{Y}^d$ composed of $d \in \mathbb{N}^*$ tokens from a finite set $\mathcal{Y}$. Although the curriculum $\mathbf{x}_t = (x_t^0, \cdots, x_t^{d-1}) \in \mathcal{X}^d$ may in general rely on a continuous $\mathcal{X}$, all our explicit examples use $\mathcal{X} = \mathcal{Y}$, $q_T(\mathbf{y}|\mathbf{y}) = 1$ and $\pi(\hat{\mathbf{x}}_T) = \hat{\mathbf{x}}_T$. In any case, the output sequence $\hat{\mathbf{y}} \in \mathcal{Y}^d$ and states $\mathbf{x}_t \in \mathcal{X}^d$ match these domains. We further focus on curriculum-prescribing processes that factorize in terms of per-token transitions $q^i_{t|t+1}(x'|x)$, *i.e.*, $q_{t|t+1}(\mathbf{x}_t|\mathbf{x}_{t+1}) = \prod_i q^i_{t|t+1}(x^i_t|x^i_{t+1})$. We now consider two important examples.

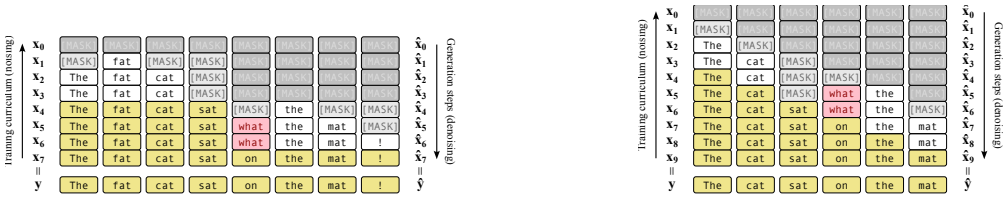
Standard AR. Most modern language models predict tokens *autoregressively*: one token at a time, each conditional on the tokens that precede it. This is an extreme case of a diffusion model where the conditional probability $q_{t|t+1}$ is actually a deterministic function that masks-out the t-th token, ultimately resulting in $\mathbf{x}_0$ composed solely of MASK tokens. At generation, $\hat{\mathbf{x}}_0$ starts as all masks, and each $p^{\boldsymbol{\theta}}_{t+1|t}(\hat{\mathbf{x}}_{t+1}|\hat{\mathbf{x}}_t)$ predicts the sole entry $\hat{x}^t_{t+1}$ that differs between $\hat{\mathbf{x}}_{t+1}$ and $\hat{\mathbf{x}}_t$. Implementations need not actually track MASK tokens because they always end the sequence. Training on a sample $\mathbf{y}$ uses cross-entropy loss on $\mathbb{P}_{\hat{X}^i_{i+1}|\hat{\mathbf{X}}^{:i}_i}(y^i|\mathbf{y}^{:i})$ for each position i.[2]

Standard diffusion. Discrete diffusion models prescribes the curriculum using $q^i_{t|t+1}(x'|x)$ given by the x-th column of a matrix $Q_{t|t+1}$ that is independent of the position i. For example, SEDD (Lou et al., 2024) uses $Q_{t|t+1} = \exp\big((\bar{\sigma}_{T-t} - \bar{\sigma}_{T-t-1})Q_{\text{tok}}\big)$, where the transition matrix Q_{tok} is one of Q_{Uniform} or Q_{Absorb} (see Appendix A), and $\bar{\sigma}_0 \leq \bar{\sigma}_1 \leq \cdots \leq \bar{\sigma}_T$ are cumulative noise schedules such that $\bar{\sigma}_0 \approx 0$ and $\bar{\sigma}_T \approx \infty$. In words, using Q_{Uniform} gradually replaces tokens by random ones, while Q_{Absorb} gradually replaces them by MASK. Generation starts from $\hat{\mathbf{x}}_0$ sampled from the stationary distribution p_0 (*i.e.*, random non-mask tokens for Q_{Uniform} and all-masks for Q_{Absorb}), and $p^{\boldsymbol{\theta}}_{t+1|t}$ is learned in terms of diffusion weighted denoising score entropy (DWDSE) (Lou et al., 2024).

Although MDMs technically correspond to the above for the case Q_{Absorb}, the literature has converged on a much simpler formulation in terms of $1 \approx \alpha_0 \geq \alpha_1 \geq \cdots \geq \alpha_T \approx 0$ such that x^i_t has probability α_{T-t} to be the original token y^i and probability $1 - \alpha_{T-t}$ to be MASK. The resulting transition matrix $Q_{t|t+1} = \mathbb{1} + (1 - \frac{\alpha_{T-t}}{\alpha_{T-t-1}})Q_{\text{Absorb}}$ may be further simplified, allowing training $p^{\boldsymbol{\theta}}_{t+1|t}$ using a weighted cross-entropy loss (Sahoo et al., 2024).

[1] Note that t here goes *down* from T to 0 as we align our notation with generation steps. Existing diffusion work often take the opposite perspective, tracking a "noise level" $T - t$.

[2] Indices $^{a:b}$ go from a (inclusive) to b (exclusive); omitted a or b are implicit 0 or d, respectively.



(a) $d = 8$ tokens using $\boldsymbol{\tau}^{\omega=4}_{\text{Block}}$.

(b) $d = 6$ tokens using $\boldsymbol{\tau}^{\omega=4}_{\text{Slide}}$.

Figure 3: Examples of our hyperschedule-powered hybrid model in action. Gold tokens are settled; dark grey mask tokens are worthless; and the remaining tokens are active. Active mask (light gray) are "known unknowns". Here the "uniform" component of our hybrid process introduces one error in the curriculum (red), teaching the model to fix such errors at inference. Although both hyperschedules generate at the limit rate of $\rho = 1$ token per step in the long-sequence limit, $\boldsymbol{\tau}^{\omega=4}_{\text{Slide}}$ experiences an initial overhead of $\omega - 1$ steps.

2.3 Transformers

Like most modern language models, all sequence generators considered in this work are implemented as transformers (Vaswani et al., 2017), more specifically the backbone from Peebles & Xie (2023). We call ALIGNED the transformer configuration used in most masked language models and diffusion models: each transformer cell predicts (output) the same token position as the one it receives (input). Conversely, we call SHIFTED the configuration used in most autoregressive models: each cell predicts the *next* token in the sequence. To simplify discussions, we fully commit to our sequence generator abstraction when indexing positions, irrespective of these input/output configuration (see details in Appendix B).

3 Autoregressive Sequence Diffusion

3.1 Hyperschedules

We generalize standard diffusion curricula by subjecting different token positions i to different noise schedules, distinguishing the number of generation steps T from the number of noise levels $\mathcal{T}$. Concretely, $q^i_{t|t+1}(x'|x)$ is given by the x-th column of a $Q^i_{t|t+1}$ obtained by substituting each instance of σ_{T-t} or α_{T-t} in $Q_{t|t+1}$ by $\sigma_{\tau^i_t}$ or $\alpha_{\tau^i_t}$, respectively, where the *hyperschedule* $\tau^i_t \in \{0, 1, \cdots, \mathcal{T}\}$ satisfying $\mathcal{T} = \tau^i_0 \geq \tau^i_1 \geq \cdots \geq \tau^i_T = 0$ for all positions $i \in \{0, \cdots, d-1\}$. In effect, the noise schedule ($\bar{\boldsymbol{\sigma}}$ or $\boldsymbol{\alpha}$) unfolds differently at different positions. Figure 2 provides some examples.

We introduce two characterizations of an hyperschedule. First, we define the window width ω as the largest (among all steps t) number of positions i for which (τ^i_t, τ^i_{t+1}) is neither $(0, 0)$ nor $(\mathcal{T}, \mathcal{T})$. All other things being equal, a lower ω offers more opportunities to improve inference time (see Sec. 3.3 for examples). Standard AR models use $\boldsymbol{\tau}_{\text{Quench}}$ with value 1 where $i \geq t$ and 0 elsewhere, and thus have $\omega = 1$ by construction (Fig. 2a). Standard diffusion models use $\boldsymbol{\tau}_{\text{Flat}}$ with value $\mathcal{T} - t$ (often called "noise level" or "time"; not to be confused with our generation step t) for all i, using $\mathcal{T} = T$ and $\omega = d$. Two more examples are provided, both parametrized by ω: concurrent work on block diffusion (Arriola et al., 2025) may be understood in terms of $\boldsymbol{\tau}^{\omega}_{\text{Block}}$ (Fig. 2c), and we introduce novel $\boldsymbol{\tau}^{\omega}_{\text{Slide}}$ (Fig. 2d).

Second, we define the token generation rate ρ as the long-sequence limit of the ratio d/T. All hyperschedules explicitly presented in Fig. 2 share the same $\rho = 1$: in the long run, they require one model call to generate one token. However, all but $\boldsymbol{\tau}_{\text{Quench}}$ may be readily adapted to "quick draft" (*i.e.*, $\rho > 1$) or "think hard" (*i.e.*, $\rho < 1$) regimes.

3.2 Hybrid Processes

In the *absorb* process, each step produces the curriculum by replacing some of $\mathbf{y}$'s entries by the special MASK token. Conversely, at generation time, the only action available to the generator is to replace some MASK tokens by non-MASK ones. Notice that, unless the generator is "perfect", it may become apparent late in the generation process that some early token choices were, in retrospect, inherently incompatible. However, there are no action available for the model to "fix" these

token choices: no backsies. Although such "hindsight" situations may occur in any domain, they are particularly relevant to computer code generation and other reasoning-intensive tasks.[3]

Conversely, the *uniform* process can replace any (non-MASK) token by any other one, both at curriculum specification and sequence generation. Thus, at no point in the generation process does the model have any indication whether it has already altered a given token before. We hypothesize that this may cause a "lack of commitment" on the model's part: how much should you "trust" the value of a token? At least with absorb, MASK tokens capture known unknowns.

These observations motivate an *hybrid* forward process, of which we consider two varieties. The first one uses the SEDD framework with Q_{tok} set to $Q^{\gamma}_{\text{Hybrid}} = (1-\gamma)Q_{\text{Absorb}} + \gamma Q_{\text{Uniform}}$, interpolating the Q_{Uniform} and Q_{Absorb} extremes according to an hyperparameter $0 < \gamma < 1$. The evolution operator $\exp\big((\bar{\sigma}_{\tau^i_t} - \bar{\sigma}_{\tau^i_{t+1}})Q^{\gamma}_{\text{Hybrid}}\big)$ can be solved analytically (because Q_{Uniform} and Q_{Absorb} commute, see Appendix A), enabling use in practice with the standard SEDD loss.

Our second hybrid process variety is closer to the MDM framework: unless $\tau^i_t = 0$ (in which case $x^i_t = y^i$), the probability distribution for x^i_t is given by the y^i-th column of $\big(\mathbb{1} + \epsilon Q_{\text{Uniform}}\big)\big(\mathbb{1} + (1-\alpha_{\tau^i_t})Q_{\text{Absorb}}\big)$, where $0 < \epsilon < 1$ is a step-independent probability that the token is substituted by a uniform one, followed by a standard MDM process henceforth. A weighted cross-entropy loss is used (see Appendix B.2) and, like MDMs, the neural network does not require an explicit noise-level dependency. Which of the two variety is used can be inferred from which of γ or ϵ is specified.

3.3 Attention Mask and Efficiency

In SEDD, each call to the transformer predicts each entry of $\hat{\mathbf{x}}_{t+1}$ in view of all entries in $\hat{\mathbf{x}}_t$. In contrast, standard autoregressive models use a causal attention mask to ensure that $\hat{x}^t_{t+1}$ may only depend on $\hat{x}^{:t}_t$. Combined with the fact that $\hat{x}^{:t}_t = \hat{x}^{:t}_T$ at all step t, this causal maskenables inference-time efficiency improvements such as KV-caching. MDMs such as Sahoo et al. (2024) can enable a similar form of caching by relying on the special role of MASK tokens.

However, new opportunities for optimization come up when the hyperschedule follows a certain autoregressive-like regular structure such as the ones seen in Fig. 2d-c. More specifically, at each steps t the hyperschedule $\boldsymbol{\tau}_t$ and the state $\hat{\mathbf{x}}_t$ both break in three components

$$\boldsymbol{\tau}_t = \boldsymbol{\tau}^{\text{settled}}_t \frown \boldsymbol{\tau}^{\text{active}}_t \frown \boldsymbol{\tau}^{\text{worthless}}_t \qquad \hat{\mathbf{x}}_t = \hat{\mathbf{x}}^{\text{settled}}_t \frown \hat{\mathbf{x}}^{\text{active}}_t \frown \hat{\mathbf{x}}^{\text{worthless}}_t \,, \tag{1}$$

where: $\boldsymbol{\tau}^{\text{settled}}_t$ is composed exclusively of zeros and $\hat{\mathbf{x}}^{\text{settled}}_t$ matches the first entries of $\hat{\mathbf{x}}_T = \hat{\mathbf{y}}$; both $\boldsymbol{\tau}^{\text{active}}_t$ and $\hat{\mathbf{x}}^{\text{active}}_t$ have at most $\omega \ll d$ entries; and $\boldsymbol{\tau}^{\text{worthless}}_t$ is composed exclusively of repeated $\mathcal{T}$ while $\hat{\mathbf{x}}^{\text{worthless}}_t$ bears no information about $\hat{\mathbf{y}}$. Thus, when using an autoregressive attention mask on $\hat{\mathbf{x}}^{\text{settled}}_t$, all the conditions are met to use KV-caching on these tokens just as in a standard autoregressive model. We may completely ignore $\hat{\mathbf{x}}^{\text{worthless}}_t$, which leaves a small number ω of positions that densely attend to $\hat{\mathbf{x}}^{\text{settled}}_t \frown \hat{\mathbf{x}}^{\text{active}}_t$ when generating $\hat{\mathbf{x}}^{\text{active}}_{t+1}$. See details in Appendix B.

3.4 Adaptive Correction Sampler

In addition to the theoretically-grounded inference schemes from SEDD and MDLM, we introduce *adaptive correction sampler* (ACS), a novel variation on MDLM's sampler that allows the model to alter the value of already-unmasked tokens, and that has empirically shown to perform particularly well for our hybrid process of the ϵ-variety. We write p^i_{transfer} the probability that MDLM's sampler (adapted to use our hyperschedule) would unmask the i-th token if it is masked, and proceed as usual for the token that are so masked. However, where MDLM would leave already-unmasked tokens as they are, ACS has probability $\eta(1 - p^i_{\text{transfer}})$ to sample a replacement token from the model's prediction. η serves as a hyperparameter modulating the intensity of this correction. Pseudocode for the original sampler and ACS sampler can be found in Appendix D.

Table 1: *Test Perplexity* for various design choices (lower is better), measured on the heldout 100k sample from **OWT** dataset. All ablations use ALIGNED with $d = 1024$.

Mark	Method	Test PPL ↓	γ	τ
(a)	Baseline SEDD-Absorb [12]	24.10	0	$\boldsymbol{\tau}_{\text{Flat}}$
(b)	**(a)** + Hybrid Process	22.30	0.01	$\boldsymbol{\tau}_{\text{Flat}}$
(c)	**(b)** + Weighted token-embedding(= γ-Hybrid)	22.18	0.01	$\boldsymbol{\tau}_{\text{Flat}}$
(d)	**(c)** − Transformer time-conditioning	22.47	0.01	$\boldsymbol{\tau}_{\text{Flat}}$
(e)	**(c)** + $\boldsymbol{\tau}_{\text{Slide}}^{\omega=d}$	21.53	0.01	$\boldsymbol{\tau}_{\text{Slide}}^{\omega=d}$

Table 2: Test perplexities (PPL; ↓) on LM1B. Perplexity values for diffusion models are upper-bound estimations. †Reported in He et al. (2022). ‡Reported in Sahoo et al. (2024). Best diffusion value is bolded.

		Parameters	PPL (↓)
Autoregressive	OmniNet$_T$ (Tay et al., 2021)	100M	21.5
	Transformer (65B tokens) (Sahoo et al., 2024)‡	110M	22.3
Diffusion	SEDD (65B tokens) (Lou et al., 2024)	110M	32.8
	MDLM (65B tokens) (Sahoo et al., 2024)	110M	31.8
	BD3-LMs $L' = 4$ (65B tokens) (Arriola et al., 2025)	110M	28.2
Diffusion (Ours)	γ-Hybrid $[\gamma = 0.02, \boldsymbol{\tau}_{\text{Flat}}, \text{ALIGNED}]$ (56B tokens)	110M	**27.8**
	γ-Hybrid $[\gamma = 0.02, \boldsymbol{\tau}_{\text{Flat}}, \text{SHIFTED}]$ (56B tokens)	110 M	28.3
	γ-Hybrid $[\gamma = 0.02, \boldsymbol{\tau}_{\text{Block}}^{\omega=d/64}, \text{ALIGNED}]$ (65B)	110M	**27.1**
	γ-Hybrid $[\gamma = 0.02, \boldsymbol{\tau}_{\text{Block}}^{\omega=d/4}, \text{ALIGNED}]$ (65B)	110M	**27.0**
	γ-Hybrid $[\gamma = 0.02, \boldsymbol{\tau}_{\text{Block}}^{\omega=d/64}, \text{SHIFTED}]$ (65B)	110M	**27.5**
	γ-Hybrid $[\gamma = 0.02, \boldsymbol{\tau}_{\text{Block}}^{\omega=d/4}, \text{SHIFTED}]$ (65B)	110M	**26.6**

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

Our experiments are framed in terms of three main design choices: (i) one of the hyperschedules shown in Fig. 2; (ii) one of the ALIGNED and SHIFTED configuration shown in Figs. 5ab; and (iii) the Q_{tok} transition matrix, chiefly our $Q_{\text{Hybrid}}^{\gamma}$ parameterized by α. Further nuances are detailed in Appendix B. We term *hybrid diffusion language model* (HDLM) the variations based on $Q_{\text{Hybrid}}^{\gamma}$ sharing our selected backbone.

4.2 LANGUAGE MODEL LIKELIHOOD EVALUATION

For likelihood evaluations, we conduct extensive experiments on two datasets.

OPENWEBTEXT (OWT) (Gokaslan et al., 2019): This dataset does not have a predefined split, so, following the approach of Ou et al. (2024), we reserve the last 100K documents as a held-out test set for reporting test perplexity (see Appendix E for further details). All our model instances use a context length of 1024 tokens with the `GPT2` tokenizer (Radford et al., 2019).

LM1B (Chelba et al., 2014): We use the `bert-base-uncased` tokenizer and report test perplexities on the test split. All models are trained with a context length of 128 tokens.

4.2.1 ABLATIONS

As a first step, we investigate the impact of the design choices that led us to our HDLM model. In Table 1, we report the *Test Perplexity* – computed as described in Section C.1 – on **OWT**. Our baseline configuration, denoted as **(a)**, is SEDD-Absorb Lou et al. (2024), where we adopt the original network architecture, graph structure, and initialization hyperparameters. Building on this baseline, we observe an immediate improvement when replacing the conventional absorbing transition matrix Q_{Absorb} with our hybrid process $Q_{\text{Hybrid}}^{\gamma}$; this configuration is marked as **(b)**. Inspired by the findings of Ou et al. (2024), we further enhance the model by incorporating a weighted token

[3] As an extreme example, consider the graph coloring of a particularly nasty instance.

Table 3: Zero-shot unconditional perplexity on seven benchmark datasets from Lou et al. (2024) and Sahoo et al. (2024) and Arriola et al. (2025). ‡Reported in Arriola et al. (2025). All models are trained for 524B tokens unless otherwise stated. All diffusion models are upper bounds; the best diffusion value is **bolded**. See Appendix F.1 for complete results.

Method	PTB	WikiText	LM1B	Lambada	AG News	Pubmed	Arxiv
Transformer (Sahoo et al., 2024)	82.05	25.75	51.25	51.28	52.09	49.01	41.73
SEDD Absorb‡ (Lou et al., 2024)	96.33	35.98	68.14	48.93	67.82	45.39	40.03
MDLM‡ (Sahoo et al., 2024)	90.96	33.22	64.94	48.29	62.78	43.13	37.89
BD3-LM $L' = 4$ (Arriola et al., 2025)	96.81	31.31	**60.88**	50.03	**61.67**	42.52	39.20
γ-Hybrid (444B) [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Flat}}$, ALIGNED]	**89.94**	**30.02**	61.01	**45.38**	67.51	46.57	40.62
ϵ-Hybrid (444B) [$\epsilon = 0.01, \boldsymbol{\tau}_{\text{Flat}}$, ALIGNED]	**90.89**	32.53	68.91	50.23	64.61	**41.18**	**37.85**
γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Slide}}^{\omega=d/4}$, ALIGNED]	**90.67**	31.73	73.71	**50.03**	68.27	**41.49**	**37.89**
γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Block}}^{\omega=d/64}$, SHIFTED]	95.22	32.64	63.68	**44.75**	62.18	**42.01**	**37.33**

embedding layer to scale the standard token embeddings (configuration **(c)**; see Appendix B.5). Following Sahoo et al. (2024), we also experiment with removing the timestep conditioning from the transformer backbone (configuration **(d)**). However, this modification results in a slight degradation in performance, therefore we retain the original timestep conditioning layers.[4]

Additionally, we examine the effect of the $\boldsymbol{\tau}_{\text{Slide}}^{\omega=1024}$ curriculum, yielding a modest improvement. This could be a "genuine" advantage of progressively settling the tokens in a left-to-right manner. However, although the comparison is "fair" in terms of the number of times each position gets updated, the neural of calls to the neural network is doubled.[5] We note that these curricula are provided as a proof-of-concept; their relative effectiveness depends critically on the precise values of ρ and ω, and will be investigated further in future work.

4.2.2 Language Modeling Analysis

Now that we have established our design choices, we compare our model's test perplexity with other baseline diffusion and autoregressive models of similar scales. Table 2 shows that on the **LM1B** dataset our model outperforms all previous baselines, achieving an improvement of approximately 19% in comparison with SEDD (Lou et al., 2024) and 3% in comparison with the best diffusion language model in test perplexity. For a fair comparison, all models except our base HDLM are trained for 33B tokens.

4.3 Zero-Shot Likelihood Evaluation

We also assess the ability of our models to generalize to unseen data. We evaluate on the seven benchmark datasets proposed in Lou et al. (2024) and Sahoo et al. (2024). As shown in Table 3, our models outperform all discrete diffusion models on 6 out of the 7 benchmarks. Our best configuration Moreover, our models reduce the performance gap between autoregressive and diffusion-based language models while outperforming them on 2 out of 7 benchmarks.

4.4 Sequence Generation Trade-offs

To assess the balance between quality and diversity in our generated sequences, we analyze two Pareto frontiers. In the left panel of Figure 4, generative perplexity is plotted against token-level entropy. Here, lower perplexity indicates more fluent and coherent generation, while higher entropy reflects greater diversity in the output. In the right panel, generative perplexity is plotted against MAUVE Pillutla et al. (2021)—a metric that measures the similarity between model-generated and

[4]Note that the conclusions of Sahoo et al. (2024) are conditioned on using Q_{Absorb}.

[5]The statement that $\rho = 1$ for $\boldsymbol{\tau}_{\text{Slide}}^{\omega}$ is only meaningful for $\omega \ll d$, whereas here $\omega = d = 1024$.

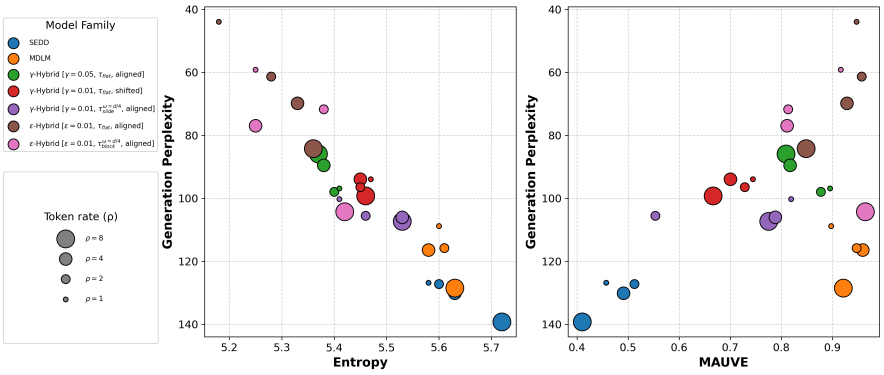


Figure 4: Left: Generative perplexity as a function of token-level entropy. Right: Generative perplexity versus MAUVE score. Our models consistently outperform baselines, achieving lower perplexity at comparable levels of diversity and fluency.

human-written text—to further assess the quality of generated sequences. MAUVE is measured against the held-out OWT documents. Our proposed models consistently occupy a superior region of the Pareto frontiers, achieving lower perplexity for a given level of diversity and higher MAUVE scores for a comparable perplexity level relative to baseline approaches. These findings suggest that our approach successfully balances generation quality and diversity, thereby advancing the state of the art in diffusion-based sequence generation. (see Suppl. F.3 for more details)

4.5 Additional Results

Additional qualitative conditional and unconditional generation examples from each model family are provided in Appendix G. We observe that our models are capable of extending beyond the nominal context length—particularly on **OWT**, which contains longer samples

Our extensive ablation studies, detailed in Appendices 4.2.1, F, and F.4, confirm the effectiveness of our design choices. In particular, we show that a small α (around 0.01 to 0.1) is critical for balancing token commitment and flexibility, as evidenced by improved test perplexity (see Table 7 and Fig. 11). We also show that the adjustment of the generation rate ρ affects both the quality and the speed of generation (Table 8), and our proposed $\boldsymbol{\tau}^{\omega}_{\text{Block}}$ models benefit from reduced inference time, especially when combined with KV-caching (Table 10).

5 Conclusion

Diffusion-based language models offer some unique opportunities – including theory-supported guidance strategies and the native ability to iteratively improve their answer – but these benefits are no substitutes for raw language modeling performances. Staggering resources are continuously spent in scaling up AR models, engineering tools and techniques specialized to the AR paradigm. How could diffusion models even dream of catching up?

This work takes significant step toward a bold strategy: starting from an already-great AR language model, we wish to convert it (*e.g.*, fine-tuning) into an even better diffusion-based sequence generation model. This plan demands a SHIFTED configuration, an hyperschedules generalizing the AR concept (*e.g.*, $\boldsymbol{\tau}_{\text{Slide}}$ or $\boldsymbol{\tau}_{\text{Block}}$), and a curriculum (such as our hybrids) teaching the model how to generate quality sequences without painting itself into a corner.

Much of the design space opened by our innovations remain to be explored.We have merely glanced at the realm of possible hyperschedules, and our success with ϵ-Hybrid illustrates that more involved curricula can pay off, without the need to provide explicit noise levels to the model.

Our innovations also open the path for more fundamental work. Indeed, the limit $1 < \omega \ll d$ presents opportunities for tractable approximations of the joint distribution over ω tokens. On a different front, while our current approach employs a uniform distribution for replacing tokens, further improvements in diversity and quality may be achieved with distributions that more accurately reflect plausible, "honest" and/or "on policy" errors (rather than purely random token substitutions).

REFERENCES

Marianne Arriola, Subham Sekhar Sahoo, Aaron Gokaslan, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Justin T Chiu, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tyEyYT267x.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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. In *Conference of the International Speech Communication Association (Interspeech)*, 2014.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus, 2019. URL http://Skylion007.github.io/OpenWebTextCorpus. Accessed: [DATE].

Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022.

Ari Holtzman, Jacob Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Ce Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2742–2751, 2019.

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

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. *arXiv preprint arXiv:2410.06264*, 2024.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=CNicRIVIPA.

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.

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. *arXiv preprint arXiv:2406.03736*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Sherwood Yao, Jarrid Rector-Brooks, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling. *arXiv preprint arXiv:2502.03540*, 2025.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yi Tay, Mostafa Dehghani, Vamsi Aribandi, Jai Gupta, Philip M Pham, Zhen Qin, Dara Bahri, Da-Cheng Juan, and Donald Metzler. Omninet: Omnidirectional representations from transformers. In *International Conference on Machine Learning*, pp. 10193–10202. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. *arXiv preprint arXiv:2503.00307*, 2025.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

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. *arXiv preprint arXiv:2409.02908*, 2024.

Kun Zhou, Yifan Li, Wayne Xin Zhao, and Ji-Rong Wen. Diffusion-nat: Self-prompting discrete diffusion for non-autoregressive text generation. *arXiv*, 2023.

A EVOLUTION OPERATORS

The two Q_{tok} considered in SEDD may be rewritten as

$$Q_{\text{Uniform}} = \begin{bmatrix} \frac{2-|\mathcal{Y}|}{|\mathcal{Y}|-1} & \cdots & \frac{1}{|\mathcal{Y}|-1} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \frac{1}{|\mathcal{Y}|-1} & \cdots & \frac{2-|\mathcal{Y}|}{|\mathcal{Y}|-1} & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix} \qquad Q_{\text{Absorb}} = \begin{bmatrix} -1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & -1 & 0 \\ 1 & \cdots & 1 & 0 \end{bmatrix}, \tag{2}$$

where $|\mathcal{Y}|$ is the number of tokens in the set $\mathcal{Y}$, including the special MASK token associated with the last dimension of Q_{tok}.

Although our SEDD-based curricula are *a priori* defined in terms of an arbitrary Q_{tok}, actually using a model in practice demands that we can analytically solve the evolution operator $\exp(\Delta Q_{\text{tok}})$ for $\Delta \in \mathbb{R}^+$. This section re-derives the solutions for Q_{Uniform} and Q_{Absorb}, then extends the results to $Q^{\gamma}_{\text{Hybrid}}$.

Using the definitions in Eq. (2), we can verify

$$(Q_{\text{Absorb}})^2 = -Q_{\text{Absorb}} \tag{3a}$$
$$(Q_{\text{Uniform}})^2 = -Q_{\text{Uniform}} \tag{3b}$$
$$Q_{\text{Uniform}}Q_{\text{Absorb}} = -Q_{\text{Uniform}} \tag{3c}$$
$$Q_{\text{Absorb}}Q_{\text{Uniform}} = -Q_{\text{Uniform}} \quad . \tag{3d}$$

Notice that, for any matrix Q_* such that $(Q_*)^2 = \lambda Q_*$, we have

$$\begin{aligned} \mathrm{e}^{\phi Q_*} &= \sum_{k=0}^{\infty} \frac{(\phi Q_*)^k}{k!} = \mathbb{1} + \lambda^{-1} Q_* \sum_{k=1}^{\infty} \frac{(\lambda\phi)^k}{k!} = \mathbb{1} + \lambda^{-1} Q_* \left[-1 + \sum_{k=0}^{\infty} \frac{(\lambda\phi)^k}{k!} \right] \\ &= \mathbb{1} - \lambda^{-1}(1 - \mathrm{e}^{\lambda\phi}) Q_* \quad . \end{aligned} \tag{4}$$

Together with Eq. (3), we re-obtain the evolution operators used in SEDD

$$\mathrm{e}^{\Delta Q_{\text{Absorb}}} = \mathbb{1} + (1 - \mathrm{e}^{-\Delta}) Q_{\text{Absorb}} \tag{5a}$$
$$\mathrm{e}^{\Delta Q_{\text{Uniform}}} = \mathbb{1} + (1 - \mathrm{e}^{-\Delta}) Q_{\text{Uniform}} \quad . \tag{5b}$$

Now notice that Q_{Absorb} and Q_{Uniform} commute

$$[Q_{\text{Absorb}}, Q_{\text{Uniform}}] = Q_{\text{Absorb}}Q_{\text{Uniform}} - Q_{\text{Uniform}}Q_{\text{Absorb}} = 0 \quad , \tag{6}$$

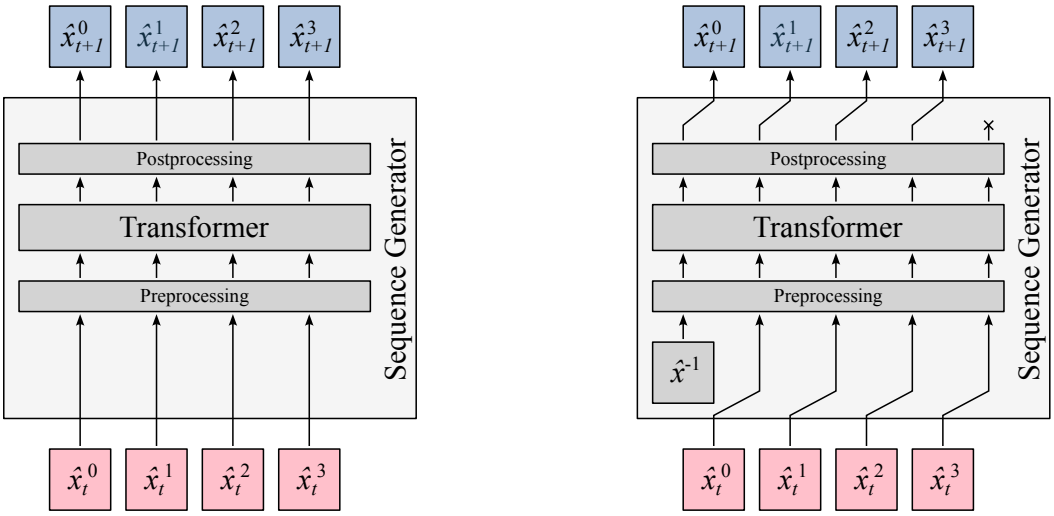
which enables the analytical solution for $Q^{\gamma}_{\text{Hybrid}}$

$$\mathrm{e}^{\Delta Q^{\gamma}_{\text{Hybrid}}} \tag{7a}$$
$$= \mathrm{e}^{\Delta((1-\gamma)Q_{\text{Absorb}} + \gamma Q_{\text{Uniform}})} \tag{7b}$$
$$= \mathrm{e}^{(1-\gamma)\Delta Q_{\text{Absorb}}} \, \mathrm{e}^{\gamma \Delta Q_{\text{Uniform}}} \tag{7c}$$
$$= \left[\mathbb{1} + (1 - \mathrm{e}^{-(1-\gamma)\Delta}) Q_{\text{Absorb}}\right] \left[\mathbb{1} + (1 - \mathrm{e}^{-\gamma\Delta}) Q_{\text{Uniform}}\right] \tag{7d}$$
$$= \mathbb{1} + (1 - \mathrm{e}^{-(1-\gamma)\Delta}) Q_{\text{Absorb}} + (1 - \mathrm{e}^{-\gamma\Delta}) Q_{\text{Uniform}} + (1 - \mathrm{e}^{-(1-\gamma)\Delta})(1 - \mathrm{e}^{-\gamma\Delta}) Q_{\text{Absorb}} Q_{\text{Uniform}} \tag{7e}$$
$$= \mathbb{1} + (1 - \mathrm{e}^{-(1-\gamma)\Delta}) Q_{\text{Absorb}} + (1 - \mathrm{e}^{-\gamma\Delta}) Q_{\text{Uniform}} - (1 - \mathrm{e}^{-(1-\gamma)\Delta})(1 - \mathrm{e}^{-\gamma\Delta}) Q_{\text{Uniform}} \tag{7f}$$
$$= \mathbb{1} + (1 - \mathrm{e}^{-(1-\gamma)\Delta}) Q_{\text{Absorb}} + (\mathrm{e}^{-(1-\gamma)\Delta} - \mathrm{e}^{-\Delta}) Q_{\text{Uniform}} \quad . \tag{7g}$$

Equation (7g) is the desired analytical solution for the evolution operator.

B IMPLEMENTATION NUANCES

This section discusses several implementation details that affect both our model training and evaluation procedures.



(a) ALIGNED config. (b) SHIFTED config.

Figure 5: Two transformer-based sequence generators for $d = 4$. (a) The ALIGNED configuration of standard diffusion models is reminiscent of masked language models. (b) The SHIFTED configuration is closer to autoregressive language models. Here $\hat{x}^{-1}$ represent a token solely part of the conditioning (*i.e.*, not generated), and may or may not be constant (*e.g.*, BOS). Similarly, ↟ represents that the output associated with the last token is discarded. Our position-based indexing abstracts away these details.

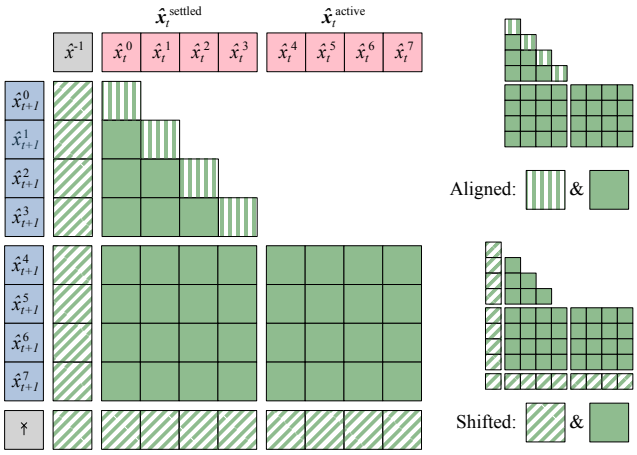


Figure 6: Example of attention mask for ALIGNED and SHIFTED configurations. Although these naive masks are appropriate for inference, directly training on them would be inefficient; see Figures 7–10 for training-ready masks examples.

B.1 TRAINING SETUP

We now detail our experimental training procedure, structured into two distinct stages. In Stage 1, we initially train our base models in both ALIGNED and SHIFTED configurations using the hybrid noising process paired with the flat hyperschedule τ_{flat}. Specifically, we adopt different modeling strategies depending on the chosen variant. For γ-variant models, we extend the existing discrete diffusion framework from Lou et al. (2024), integrating our proposed hybrid transition operator $Q^{\gamma}_{\text{Hybrid}}$. For ϵ-variant models, we instead employ the newly proposed hybrid diffusion cross-entropy (HDCE) loss, detailed in Equation 8. Stage 1 models are trained separately on two standard datasets: *OpenWebText* and *LM1B*. Each configuration undergoes training for approximately 850K gradient

updates with a batch size of 512, processing roughly 444B tokens for OpenWebText and 56B tokens for LM1B.

In Stage 2, we fine-tune the Stage 1 models under alternative hyperschedules—specifically, τ_{block} and τ_{slide}. These experiments involve custom-designed attention masks tailored to each hyperschedule. Due to this customization, highly optimized attention kernels such as Flash-Attention are not applicable, necessitating reliance on the standard PyTorch attention mechanism, which incurs higher computational costs. Stage 2 training continues for an additional 150K gradient steps, resulting in a cumulative training volume of 524B tokens for OpenWebText and 65B tokens for LM1B.

B.2 Loss Function

Our training objective consists of two primary loss components: (i) a standard cross-entropy loss computed on the *settled* tokens, and (ii) a diffusion-weighted loss calculated on the *active* tokens. We propose the *Hybrid Diffusion Cross-Entropy* (HDCE) as our diffusion-weighted loss, which blends a per-token cross-entropy loss with a specialized weighting strategy contingent upon whether tokens are masked, shuffled, or unchanged.

Formally, the HDCE loss is defined as:

$$\mathcal{L}_{\text{HDCE}}(\theta) = \frac{1}{Nd} \sum_{i=1}^{N} \sum_{t=1}^{d} w_{i,t} \Big[-\log p_\theta\big(y_{i,t} \mid x_{i,t}\big) \Big], \tag{8}$$

where N denotes the batch size, d is the sequence length, and the per-token loss corresponds to the conventional cross-entropy formulation. The token-specific weights $w_{i,t}$ are defined as:

$$w_{i,t} = \begin{cases} \dfrac{1}{p_{\text{mask}}(i,t)} & \text{if } x_{i,t} \text{ is masked,} \\ \dfrac{\lambda(1-\epsilon)}{1-p_{\text{mask}}(i,t)} & \text{if } x_{i,t} \text{ is unmasked and shuffled,} \\ \dfrac{\lambda\epsilon}{1-p_{\text{mask}}(i,t)} & \text{if } x_{i,t} \text{ is unmasked and not shuffled,} \end{cases} \tag{9}$$

where $p_{\text{mask}}(i,t)$ is the masking probability for token $x_{i,t}$, and λ and ϵ represent hyperparameters controlling the relative importance of shuffled versus unshuffled tokens.

In practice, we distinguish two model variants based on the employed diffusion-weighted loss. For γ-variant hybrid models, we adopt the diffusion-weighted denoising score entropy (DWDSE) loss proposed by Lou et al. (2024), denoted as $\mathcal{L}_{\text{DWDSE}}$. Conversely, for our ϵ-variant hybrid models, we use the proposed HDCE loss as defined in Equation 8.

Letting $\mathcal{L}_{\text{CE}}$ represent the cross-entropy loss computed over $\hat{\mathbf{x}}_t^{\text{settled}}$, our overall loss function is thus expressed as:

$$\mathcal{L} \;=\; \beta_1\, \mathcal{L}_{\text{CE}}\big(\hat{\mathbf{x}}_t^{\text{settled}}\big) \;+\; \begin{cases} \beta_2\, \mathcal{L}_{\text{DWDSE}}\big(\hat{\mathbf{x}}_t^{\text{active}}\big), & \text{for } \gamma\text{-Hybrid,} \\ \beta_2\, \mathcal{L}_{\text{HDCE}}\big(\hat{\mathbf{x}}_t^{\text{active}}\big), & \text{for } \epsilon\text{-Hybrid,} \end{cases} \tag{10}$$

where $\beta_1, \beta_2 \in \mathbb{R}$ are hyperparameters balancing these two components.

Additionally, since early positions in the sequence tend to become *settled* sooner, we apply a reweighting strategy to normalize the contribution of *settled* tokens at different positions. Specifically, we partition the sequence of length d into blocks of width ω, assigning each token at position i a weight:

$$w(i) \;=\; \frac{\lfloor i/\omega \rfloor}{\lceil d/\omega \rceil - 1}, \quad i = 0, 1, \ldots, d-1, \tag{11}$$

with the convention that if $\lceil d/\omega \rceil = 1$, then $w(i) = 1$ for all i.

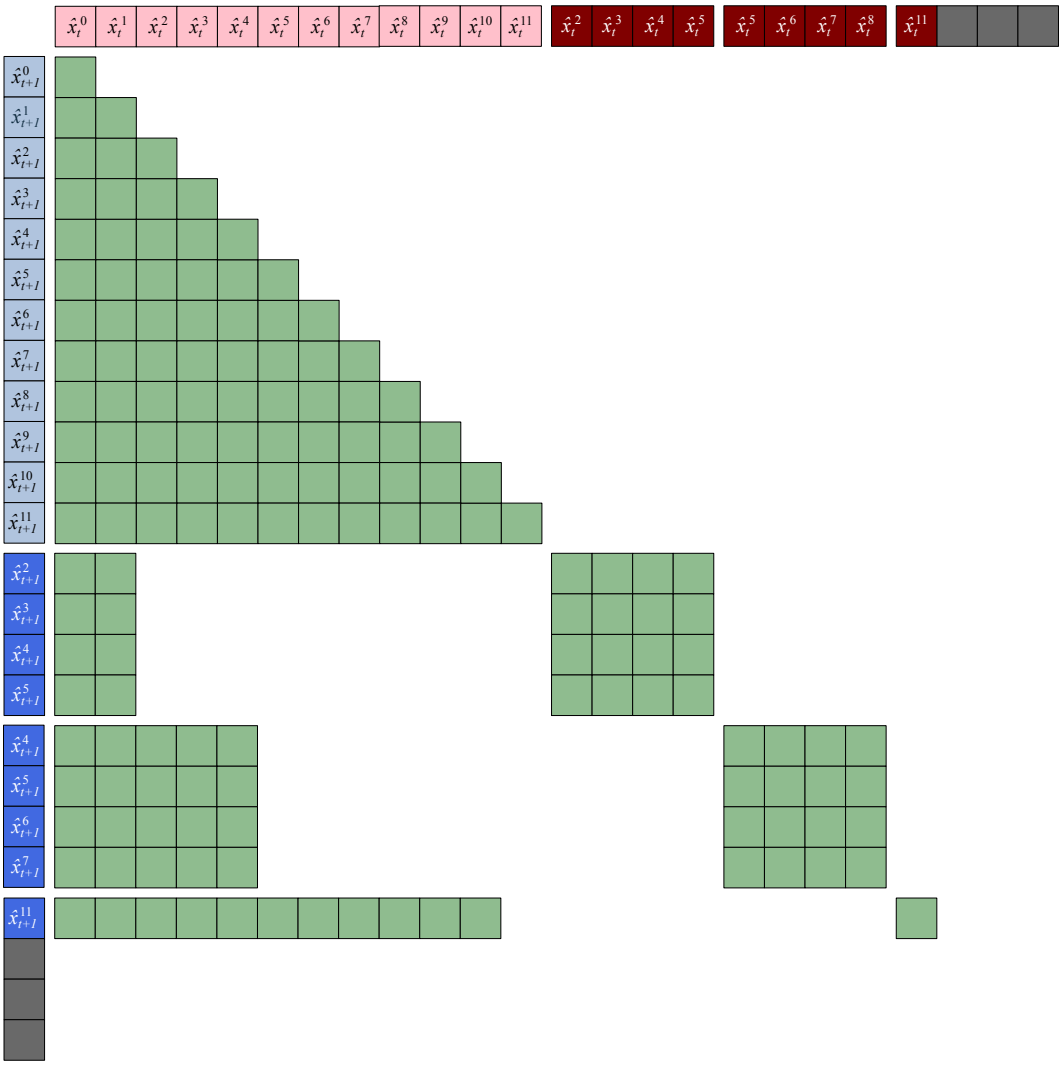


Figure 7: Example training attention mask for ALIGNED configuration for use with $\boldsymbol{\tau}_{\text{Slide}}^{\omega=4}$.

B.3 EFFICIENT TRAINING AND INFERENCE

As mentioned in Sec. 3.3, we take particular care in crafting our attention matrices to enable KV-caching at inference time. These scheme are particularly beneficent when $\omega \ll d$, but naively using an attention matrix such as Fig. 6 would train the diffusion head on only ω positions while demanding to process on average $d/2$ context tokens. Here we present how we may train the diffusion head on about approximately half the positions, increasing the training-time efficiency by a factor d/ω. Note that, under these efficient schemes, the reweighing of active tokens as given in Eq. (11) is no longer required.

Figures 7–10 provides examples of attention masks that are compatible with the KV-caching scheme presented in Sec. 3.3, while dedicating about half the positions to the denoising task. Light red/blue squares represent positions that are settled, whereas dark red/blue represent positions that are active. In all cases, the top-left part of the matrix has an autoregressive structure, and the production of dark blue positions attends densely on the corresponding dark red inputs as well as the light red inputs that precede them. All cases presume $d = 12$ and $\omega = 4$.

The ALIGNED cases are easier to understand. For $\boldsymbol{\tau}_{\text{Slide}}$, Fig. 7 presents a situation where it was randomly determined that the denoising will be performed on the intervals $j \leq i < \min(j + \omega, d)$ for $j \in \{2, 5, 11\}$. For $\boldsymbol{\tau}_{\text{Block}}$, these starting points j are always multiples of ω, here $j \in \{0, \omega, 2\omega\}$.

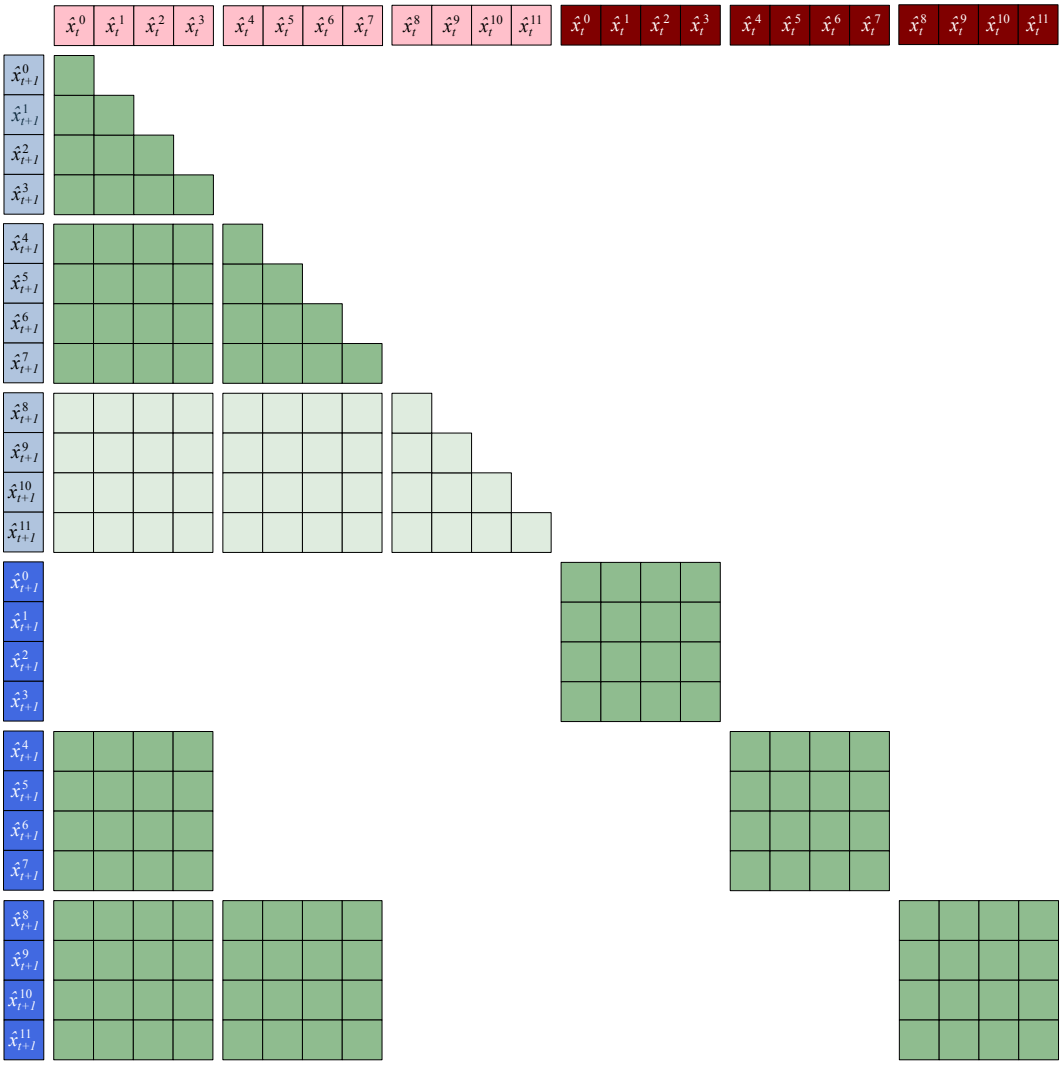


Figure 8: Example training attention mask for ALIGNED configuration for use with $\boldsymbol{\tau}^{\omega=4}_{\text{Block}}$.

The light green blocks indicate entries that are not actually involved in the denoising and could thus potentially be eschewed.

Figures 9 and 10 present the corresponding matrices for the SHIFTED configuration. Notice how settled tokens (light red or the gray $\hat{x}^{-1}$) are repeated as the first input of an interval to denoise in the second half of the matrix, and how the last output of each such interval is discarded.

B.4 INFERENCE AND KV-CACHING

At inference, both Euler and τ-leaping analytical solutions are available; however, our empirical results suggest that τ-leaping is the de facto superior choice. As a result of the presence of $\hat{\mathbf{x}}_t^{\text{settled}}$ tokens, we can leverage KV-caching to accelerate inference. Specifically, during each forward pass of the transformer, only the $\hat{\mathbf{x}}_t^{\text{active}}$ tokens are updated while the cached keys and values for $\hat{\mathbf{x}}_t^{\text{settled}}$ remain unchanged.

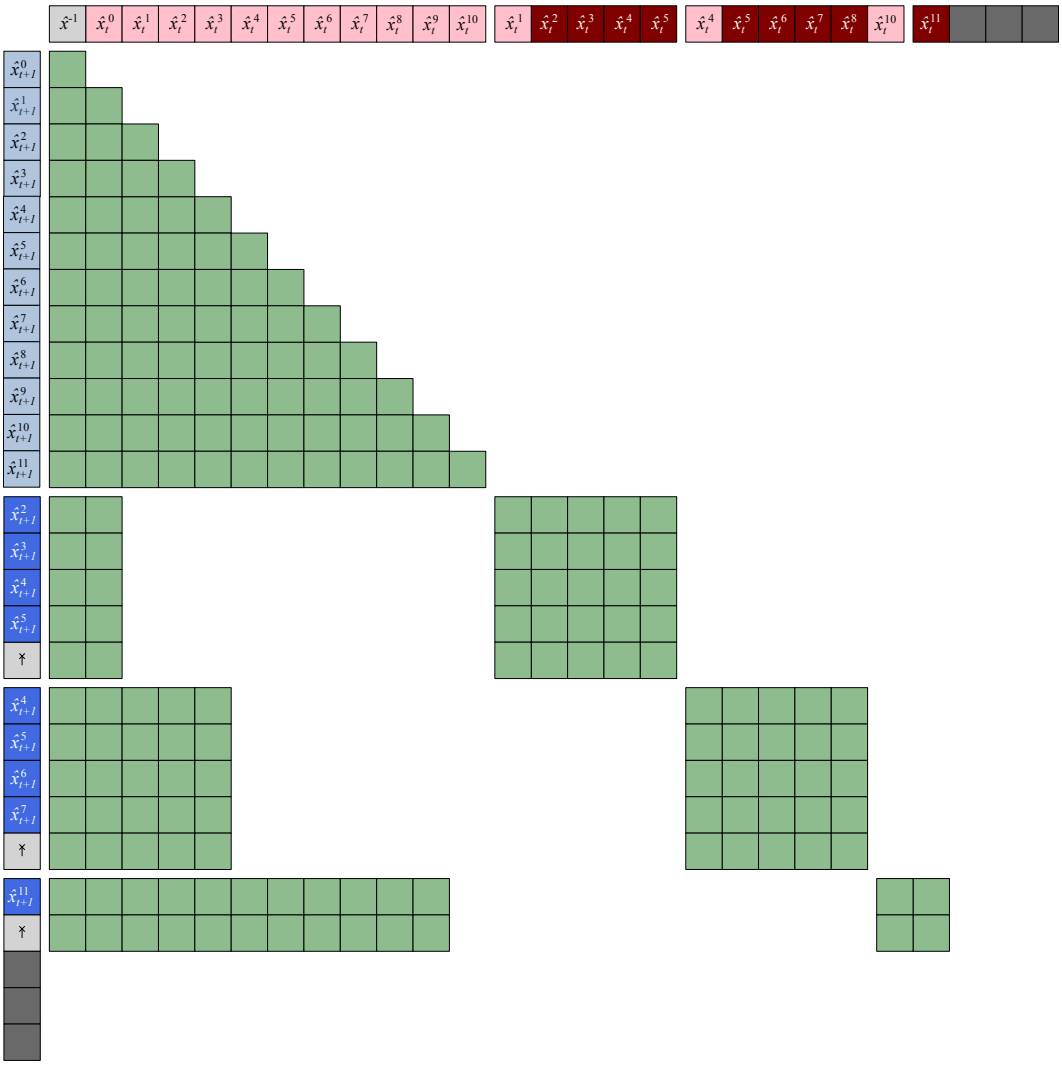


Figure 9: Example training attention mask for SHIFTED configuration for use with $\boldsymbol{\tau}^{\omega=4}_{\text{Slide}}$.

B.5 WEIGHTED TOKEN-EMBEDDING

Ou et al. (2024) demonstrated that when employing the absorbing transition matrix Q_{Absorb}, scaling the model's score by the analytic, time-dependent factor

$$\frac{\exp(-\bar{\sigma}(t))}{1-\exp(-\bar{\sigma}(t))}$$

causes the remaining score to be independent of t, eliminating the need to explicitly condition on time within the network. However, when using the γ-Hybrid process $(1-\gamma)Q_{\text{Absorb}} + \gamma\, Q_{\text{Uniform}}$, this factor remains present but is insufficient for capturing all temporal dependencies. In particular, under the hybrid process, an unmasked token is perturbed with probability $e^{-\gamma\bar{\sigma}(t)}$ (whereas under Q_{Absorb} alone, a non-mask token remains unchanged).

In other words, when the model encounters an unmasked input token x_t^i subject to cumulative noise $\bar{\sigma}(t)$, it should treat that token as if it were unperturbed with probability $\mathrm{e}^{-\gamma\bar{\sigma}(t)}$, and as if it were masked with probability $1-\mathrm{e}^{-\gamma\bar{\sigma}(t)}$. One natural way to embed this inductive bias into the model is to interpolate the token's embedding accordingly. Denoting by $\mathbf{f}(x_t^i) \in \mathbb{R}^{d_{\text{model}}}$ the standard embedding of token x_t^i, we replace it with

$$\mathrm{e}^{-\gamma\bar{\sigma}(t)}\ \mathbf{f}(x_t^i)\ +\ \left(1-\mathrm{e}^{-\gamma\bar{\sigma}(t)}\right)\mathbf{f}(\text{MASK})\,. \tag{12}$$

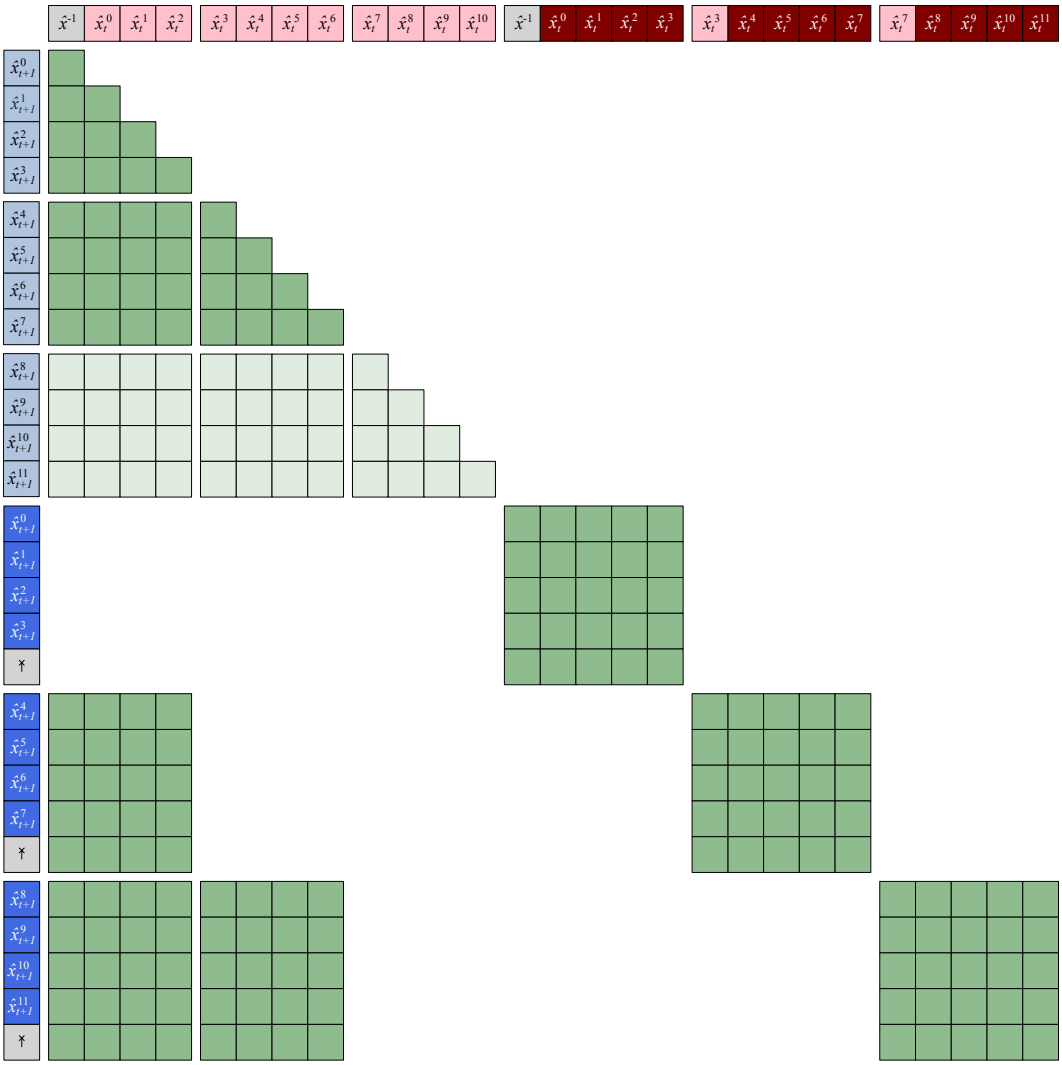


Figure 10: Example training attention mask for SHIFTED configuration for use with $\boldsymbol{\tau}^{\omega=4}_{\text{Block}}$.

C MODEL EVALUATION METRICS

C.1 UPPER BOUND ESTIMATION OF PERPLEXITY

Evaluating perplexity for diffusion-based language models is challenging because the model's likelihood involves an integration over a continuum of noise levels. In our work, we estimate the negative log-likelihood (NLL) via a Monte Carlo (MC) approximation over a discrete set of diffusion timesteps. In particular, given a trained model $p_\theta(\mathbf{y})$ and a diffusion process that perturbs a sequence $\mathbf{y}$ into latent states $\hat{\mathbf{x}}_t$ (with $t \in [0, T]$), our goal is to estimate the per-token loss that, when exponentiated, yields an *upper bound* on the true perplexity.

Let

$$\log p_\theta(\mathbf{y}) = \mathbb{E}_{t\sim q(t)}\Big[\log p_\theta\big(\mathbf{y} \mid \hat{\mathbf{x}}_t\big)\Big] - D_{\mathrm{KL}}\Big(q_t(\mathbf{x}_t \mid \mathbf{y}) \,\|\, p(\mathbf{x}_t)\Big).$$

In practice, we approximate the expectation with M Monte Carlo samples:

$$\hat{L}(\mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} \log p_\theta\big(\mathbf{y} \mid \hat{\mathbf{x}}_{t_i}\big), \quad t_i \sim \mathrm{Uniform}(0, 1).$$

Since our loss function returns the *total* NLL over a sequence of length d (i.e., it produces a tensor of per-token losses whose sum over tokens yields the total loss for a sequence), we define the average per-token loss as

$$\ell = \frac{\hat{L}(\mathbf{y})}{d}.$$

The estimated perplexity is then given by

$$\mathrm{PPL} = \exp(\ell).$$

Because the MC approximation truncates the integration to M samples, the resulting perplexity is an upper bound on the true perplexity. In our experiments, we typically set $M = 1000$, and we observe that increasing M further reduces the variance of the estimate.

For generators with semi-autoregressive or autoregressive configurations (with a fixed window width ω), we calculate the NLL only over the $\boldsymbol{\tau}_t^{\text{active}}$ tokens (i.e., those tokens that are actively being updated during generation). This ensures that the perplexity computation is fair and reflects the model's performance on the tokens whose values are uncertain, rather than being diluted by tokens that are already settled.

We summarize the estimation procedure in Algorithm 1.

Algorithm 1 Monte Carlo Upper Bound Estimation of Perplexity

Require: Model parameters θ, loss function $\mathcal{L}$, evaluation dataset $\mathcal{D}$, number of MC samples M, sequence length d

Initialize total loss $\mathcal{L}_{\text{total}} \leftarrow 0$ and token count $N_{\text{total}} \leftarrow 0$
for each batch $\mathbf{y} \in \mathcal{D}$ **do**
    Initialize batch loss $\mathcal{L}_{\text{batch}} \leftarrow 0$
    for $i = 1$ to M **do**
        Sample $t_i \sim \mathrm{Uniform}(0, 1)$
        Compute per-token loss $\ell_i \leftarrow \mathcal{L}(\theta, \mathbf{y}, t_i) \in \mathbb{R}^{B\times d}$
        $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \ell_i$
    end for
    $\mathcal{L}_{\text{batch}} \leftarrow \frac{1}{M}\,\mathcal{L}_{\text{batch}}$ ▷ Average over MC samples
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{total}} + \sum \mathcal{L}_{\text{batch}}$
    $N_{\text{total}} \leftarrow N_{\text{total}} + B \times d$
end for
Compute average per-token loss: $\ell = \mathcal{L}_{\text{total}}/N_{\text{total}}$
Compute perplexity: $\mathrm{PPL} \leftarrow \exp(\ell)$ **return** PPL

Remarks 1. *In our setting, the loss function $\mathcal{L}(\theta, \mathbf{y}, t)$ returns a tensor of shape $[B, d]$. If the generator is semi-autoregressive with a fixed active window of width ω, then the loss is computed only on the $\boldsymbol{\tau}_t^{active}$ tokens. The expectation over t is approximated by averaging over M independent samples. Since the integral is truncated, the computed perplexity serves as an* upper bound *on the true value. Finally, dividing the total loss by N_{total} gives the average per-token loss, so that the perplexity is computed as* $\exp(\textit{avg loss per token})$.

C.2 Generative Perplexity Evaluation

In addition to the intrinsic perplexity estimation described in Section C.1, we also assess our generator via *generative perplexity*. In this approach, a pretrained autoregressive language model — in our case, GPT2-Large Radford et al. (2019) — serves as an external judge of the generated sequences. This method has been used in prior work Keskar et al. (2019); Holtzman et al. (2020) as a proxy for fluency and coherence when direct likelihood evaluation is intractable.

Concretely, our procedure is as follows: We first sample sequences from our diffusion generator using the analytical solution Lou et al. (2024). Since our generator and GPT2-Large may employ different tokenization schemes, the generated samples are retokenized using the GPT2 tokenizer. The retokenized sequences are then fed into the GPT2-Large model, which computes the negative log-likelihood (NLL) for each token; this value quantifies the *surprise* of the judge regarding the generated text. Finally, by averaging the NLL over all tokens and exponentiating the result, we obtain the generative perplexity:

$$\mathrm{PPL}_{\mathrm{gen}} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} -\log p_{\mathrm{GPT2}}(x_i)\right),$$

where N is the total number of tokens in the generated text and $p_{\mathrm{GPT2}}(x_i)$ denotes the probability assigned by GPT2-Large to token x_i.

In the case of semi-autoregressive or autoregressive models with a fixed active window of width ω, we compute the NLL only over the tokens corresponding to the active portion $\boldsymbol{\tau}_t^{\text{active}}$. This ensures that the perplexity is estimated fairly by focusing on those positions where the model is actually making nontrivial predictions.

The full procedure is summarized in Algorithm 2.

Algorithm 2 Generative Perplexity Estimation via External Judge

Require: Generator G with parameters θ, pretrained judge model J (GPT2-Large), number of samples S

1: Generate a set of samples $\{\mathbf{y}^{(s)}\}_{s=1}^{S}$ using the analytical solution of G
2: Retokenize each generated sample using the GPT2 tokenizer:

$$\tilde{\mathbf{y}}^{(s)} \leftarrow \mathrm{Tokenize}_{\mathrm{GPT2}}\big(\mathbf{y}^{(s)}\big)$$

3: **for** each retokenized sample $\tilde{\mathbf{y}}^{(s)}$ **do**
4: Compute per-token negative log-likelihood $\ell^{(s)} \leftarrow -\log p_J\big(\tilde{\mathbf{y}}^{(s)}\big)$
5: **end for**
6: Compute the overall average per-token loss:

$$\bar{\ell} = \frac{1}{N}\sum_{s=1}^{S} \ell^{(s)}$$

7: Compute generative perplexity:

$$\mathrm{PPL}_{\mathrm{gen}} = \exp\big(\bar{\ell}\big)$$

return $\mathrm{PPL}_{\mathrm{gen}}$

D ADAPTIVE CORRECTION SAMPLER (ACS) PSEUDO CODE

For completeness, we present pseudo code for both the original sampling procedure and our proposed Adaptive Correction Sampler (ACS).

Algorithm 3 Original Sampler

Input: model, context length L, total steps S, temperature T.
Initialize $x \leftarrow$ a tensor of shape $[B, L]$ filled with mask tokens.
Compute timesteps $\{t_i\}_{i=0}^{S}$ linearly spaced between 1 and 0).
for $i = 0, \ldots, S-1$ **do**
    Set $t \leftarrow t_i$ and $s \leftarrow t_{i+1}$
    Compute transfer probability $p_{\text{transfer}} \leftarrow 1 - \frac{s}{t}$
    for each token in x **do**
        if token is masked and a random draw is below p_{transfer} **then**
            Add Gumbel noise to the token's logits and update it via $\arg\max$.
        end if
    end for
    Update x accordingly.
end for
return x

Algorithm 4 Adaptive Correction Sampler (ACS)

Input: model, context length L, total steps S, temperature T, correction parameter η.
Initialize $x \leftarrow$ a tensor of shape $[B, L]$ filled with mask tokens.
Compute timesteps $\{t_i\}_{i=0}^{S}$ linearly spaced between 1 and 0.
for $i = 0, \ldots, S-1$ **do**
    Set $t \leftarrow t_i$ and $s \leftarrow t_{i+1}$
    Compute transfer probability $p_{\text{transfer}} \leftarrow 1 - \frac{s}{t}$
    for each token in x **do**
        if token is masked and a random draw is below p_{transfer} **then**
            Update the token using the standard denoising update (with Gumbel noise).
        else if token is unmasked and a random draw is below $\eta\,(1 - p_{\text{transfer}})$ **then**
            Update the token via a uniform correction mechanism.
        end if
    end for
    Update x accordingly.
end for
return x

E EXPERIMENTAL SETUP

In our experiments, we adopt a training and evaluation protocol similar to that of Sahoo et al. Sahoo et al. (2024). We conduct experiments on two datasets: the One Billion Word Benchmark (*LM1B*; Chelba et al. (2014)) and *OpenWebText* (*OWT*; Gokaslan et al. (2019)). For models trained on LM1B, we employ the `bert-base-uncased` tokenizer with a context length of d = 128 tokens, and report perplexities on the test split of *LM1B*. In contrast, models trained on *OWT* use the `GPT2` tokenizer Radford et al. (2019) with a context length of d = 1024 tokens.

Since the *LM1B* corpus predominantly consists of single-sentence examples, a straightforward padding scheme for reaching a fixed context length may not be optimal. Accordingly, following Sahoo et al. Sahoo et al. (2024), we concatenate and wrap sequences to fit a context window of 128 tokens. Similarly, for *OWT* we concatenate and wrap sequences to 1024 tokens, rather than simply truncating or padding, thereby ensuring that our evaluation is performed on coherent text segments. In the case of *OWT*, which lacks a designated validation split, we reserve the final 100K documents for validation purposes.

Method	PTB	WikiText	LM1B	Lambada	AG News	Pubmed	Arxiv
GPT-2 (WebText)*	138.43	41.60	75.20	45.04	–	–	–
Transformer (Sahoo et al., 2024)	82.05	25.75	51.25	51.28	52.09	49.01	41.73
D3PM† (Austin et al., 2021)	200.82	75.16	138.92	93.47	–	–	–
Plaid† (Gulrajani & Hashimoto, 2024)	142.60	50.86	91.12	57.28	–	–	–
MD4 (Shi et al., 2024)	102.26	35.90	68.10	48.43	–	–	–
SEDD Absorb‡ (Lou et al., 2024)	96.33	35.98	68.14	48.93	67.82	45.39	40.03
MDLM‡ (Sahoo et al., 2024)	90.96	33.22	64.94	48.29	62.78	43.13	37.89
BD3-LM $L' = 4$ (Arriola et al., 2025)	96.81	<u>31.31</u>	**60.88**	50.03	**61.67**	<u>42.52</u>	39.20
RADD-λ-DCE (Ou et al., 2024)	107.85	37.98	72.99	51.70	–	–	–
γ-Hybrid (444B) [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Flat}}$, ALIGNED]	**89.94**	**30.02**	<u>61.01</u>	**45.38**	67.51	46.57	40.62
ϵ-Hybrid (444B) [$\epsilon = 0.01, \boldsymbol{\tau}_{\text{Flat}}$, ALIGNED]	**90.89**	32.53	68.91	50.23	64.61	**41.18**	**37.85**
γ-Hybrid (444B) [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Flat}}$, SHIFTED]	100.88	37.48	71.51	56.57	70.69	<u>43.06</u>	<u>38.83</u>
γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Slide}}^{\omega=d/4}$, ALIGNED]	**90.67**	31.73	73.71	**50.03**	68.27	**41.49**	**37.89**
γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Block}}^{\omega=d/4}$, ALIGNED]	**95.32**	38.94	70.49	**48.18**	67.32	44.23	42.78
γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Block}}^{\omega=d/64}$, ALIGNED]	**90.74**	35.24	<u>62.64</u>	51.21	69.62	**41.46**	**37.13**
γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Block}}^{\omega=d/64}$, SHIFTED]	<u>95.22</u>	32.64	<u>63.68</u>	**44.75**	<u>62.18</u>	**42.01**	**37.33**

Table 4: Zero-shot unconditional perplexity on seven benchmark datasets from Lou et al. (2024) and Sahoo et al. (2024) and Arriola et al. (2025). †Reported in He et al. (2022). ‡Reported in Arriola et al. (2025). *The GPT-2 numbers are reported for the checkpoint pretrained on WebText and are not a direct comparison. All models are trained for 524B tokens unless otherwise stated. All diffusion models are upper bounds; the best diffusion value is **bolded**, the second best values is <u>underscored</u>.

Our model architecture builds upon the diffusion transformer framework Peebles & Xie (2023), augmented with rotary positional embeddings Su et al. (2024). We instantiate our autoregressive baselines — SEDD, MDLM — with a transformer backbone as described in Sahoo et al. (2024): 12 layers, a hidden dimension of 768, and 128 attention heads.

F Additional Results

This section presents additional results and ablation studies on our family of models.

F.1 Zeroshot Perplexity

In Table 4 we report the full list of results for the family of our models against all the reported models available online. This is an extensive and more complete version of Table 3.

F.2 Generative Perplexity Using a Judge LLM

In this subsection, we report the *generative perplexity* (see Section C.2) of our model. Zheng et al. (2024) were the first to demonstrate that the generative perplexity evaluations of baseline masked dif-

fusion language models are flawed due to imprecise categorical sampling. They show that employing 32-bit floating point precision in categorical sampling via the Gumbel trick induces an artificial temperature-lowering effect, which results in a lower (i.e., seemingly better) generative perplexity at the expense of reduced entropy—a key indicator of generation diversity. Their proposed remedy is to cast the values to 64-bit floating point precision.

To ensure a fair comparison of generative perplexity across baselines, we report both the flawed (32-bit) and the corrected (64-bit) perplexity values in Table 5. All the entries were resampled using full precision. Our results indicate that, irrespective of the artificial temperature effect, our models consistently outperform all diffusion-based counterparts.

Method	**FP32-PPL** $\downarrow$	**FP64-PPL** $\downarrow$
SEDD-Absorb [12]	43.41	105.91
MDLM [19]	43.88	108.88
γ-Hybrid [γ= 0.05, $\boldsymbol{\tau}_{\text{Flat}}$, ALIGNED] (444B tokens)	**39.53**	**89.05**
γ-Hybrid [γ= 0.01, $\boldsymbol{\tau}_{\text{Flat}}$, SHIFTED] (444B tokens)	48.08	110.60
γ-Hybrid [γ= 0.01, $\boldsymbol{\tau}_{\text{Block}}^{\omega=d/4}$, ALIGNED]	**40.48**	**85.01**
γ-Hybrid [γ= 0.01, $\boldsymbol{\tau}_{\text{Slide}}^{\omega=d/4}$, ALIGNED]	61.05	131.45
γ-Hybrid [γ= 0.01, $\boldsymbol{\tau}_{\text{Block}}^{\omega=d/64}$, ALIGNED]	76.12	121.99
γ-Hybrid [γ= 0.01, $\boldsymbol{\tau}_{\text{Block}}^{\omega=d/4}$, SHIFTED]	53.89	111.73

Table 5: Generative perplexities (PPL; lower is better) on OWT. All models were trained for 524B tokens unless otherwise indicated. "FP32" denotes the flawed 32-bit sampling, whereas "FP64" corresponds to the corrected 64-bit precision values. All available models were resampled using their published weights.

F.3 Inference Pareto frontier Results

In Table. 6 we report the results we utilized to report the Pareto frontier plots in the main paper.

Higher ρ values ($\rho = 8, 4$)						
Method	MAUVE ($\uparrow$)		Gen PPL. ($\downarrow$)		Entropy ($\uparrow$)	
	$\rho = 8$	$\rho = 4$	$\rho = 8$	$\rho = 4$	$\rho = 8$	$\rho = 4$
SEDD	0.410	0.491	139.2	130.1	5.72	5.63
MDLM	0.921	0.959	128.5	116.4	5.63	5.58
γ-Hybrid [$\gamma = 0.05$, $\boldsymbol{\tau}_{\text{Flat}}$, ALIGNED]	0.809	0.817	85.9	89.5	5.37	5.38
γ-Hybrid [$\gamma = 0.01$, $\boldsymbol{\tau}_{\text{Flat}}$, SHIFTED]	0.666	0.700	99.2	93.9	5.46	5.45
γ-Hybrid [$\gamma = 0.01$, $\boldsymbol{\tau}_{\text{Slide}}^{\omega d/4}$, ALIGNED]	0.775	0.788	107.3	106.0	5.53	5.53
ϵ-Hybrid [$\epsilon = 0.01$, $\boldsymbol{\tau}_{\text{Flat}}$, ALIGNED]	0.848	0.928	84.2	69.8	5.36	5.33
ϵ-Hybrid [$\epsilon = 0.01$, $\boldsymbol{\tau}_{\text{Block}}^{\omega d/4}$, ALIGNED]	0.964	0.811	104.2	76.9	5.42	5.25
Lower ρ values ($\rho = 2, 1$)						
Method	MAUVE ($\uparrow$)		Gen PPL. ($\downarrow$)		Entropy ($\uparrow$)	
	$\rho = 2$	$\rho = 1$	$\rho = 2$	$\rho = 1$	$\rho = 2$	$\rho = 1$
SEDD	0.512	0.457	127.2	126.8	5.60	5.58
MDLM	0.947	0.897	115.8	108.8	5.61	5.60
γ-Hybrid [$\gamma = 0.05$, $\boldsymbol{\tau}_{\text{Flat}}$, ALIGNED]	0.877	0.895	97.9	96.8	5.40	5.41
γ-Hybrid [$\gamma = 0.01$, $\boldsymbol{\tau}_{\text{Flat}}$, SHIFTED]	0.728	0.744	96.4	93.9	5.45	5.47
γ-Hybrid [$\gamma = 0.01$, $\boldsymbol{\tau}_{\text{Slide}}^{\omega d/4}$, ALIGNED]	0.553	0.819	105.5	100.2	5.46	5.41
ϵ-Hybrid [$\epsilon = 0.01$, $\boldsymbol{\tau}_{\text{Flat}}$, ALIGNED]	0.957	0.947	61.3	43.9	5.28	5.18
ϵ-Hybrid [$\epsilon = 0.01$, $\boldsymbol{\tau}_{\text{Block}}^{\omega d/4}$, ALIGNED]	0.813	0.916	71.7	59.1	5.38	5.25

Table 6: Sample quality of absorbing state discrete diffusion models. Upper block: higher ρ values ($\rho = 8, 4$); Lower block: lower ρ values ($\rho = 2, 1$).

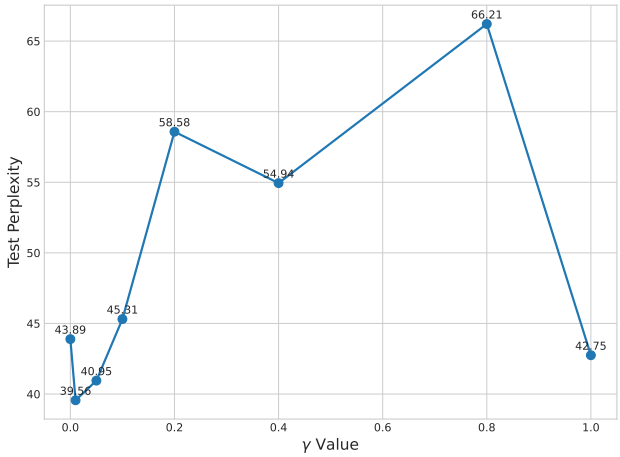


Figure 11: Ablation study illustrating the effect of varying the γ (for γ-Hybrid variants) parameter on perplexity evaluated on the OWT test split at the 26B-token observation point. Lower perplexity values reflect improved model performance. Consistent with our previous observations, γ between 0.01 and 0.1 yields optimal performance.

F.4 Effect of Varying γ

In this section, we examine the influence of the hyperparameter α, which modulates the contribution of Q_{uniform} in the hybrid process and thereby allows the model to reexamine its predictions after unmasking a token. As shown in Table 7, while the corrective influence of Q_{uniform} is essential, the value of α must remain relatively small. If α is set too high, the model tends to simply reshuffle the tokens and the MASK token, effectively undermining the intended unmasking process.

Another perspective is that increasing α reduces the penalty associated with errors in the unmasking operation, thereby devaluing its corrective impact. Moreover, during the denoising process, each token is influenced not only by its own prediction but also by the context provided by neighboring tokens. Consequently, if a token (e.g., token A) is mispredicted, the resulting change in the overall structure may leave little opportunity for subsequent correction.

Remarks 2. *The analyses presented above are mainly intuitive. Further empirical investigation is necessary to confirm.*

Configuration	**FP64-PPL** $\downarrow$
$\gamma = 0.01$	84.32
$\gamma = 0.05$	82.27
$\gamma = 0.1$	85.15
$\gamma = 0.4$	91.48
$\gamma = 0.8$	90.99

Table 7: Generative perplexities (PPL; lower is better) on OWT. All the models are trained under ALIGNED configuration with $\boldsymbol{\tau}_{\text{Block}}^{\omega=256}$ for 524B tokens. We use the double precision, denoted as "FP64-PPL".

To further examine the effect of α, we evaluate our trained models' test perplexity (on OWT held out set) after processing 26B tokens under various configurations. As shown in Fig. 11, small α values—approximately 0.01 and 0.1—yield the best performance.

F.5 Effect of Varying ρ

In this section, we examine the impact of varying ρ in our γ-Hybrid models using the $\boldsymbol{\tau}^{\omega=256}_{\text{Block}}$ hyperschedule. Recall that the parameter ρ is modified during the generation process and directly only influences the quality of the generated sequence. Consequently, we adopt *generative perplexity* as our evaluation metric. For completeness—and to facilitate comparison with prior baselines—we report the generative perplexity computed under both double precision (FP64) and full precision (FP32), as illustrated in Table 8. As ρ decreases, the generation process becomes slower, thereby entering the "think hard" regime; in this regime, the model tends to produce higher-quality outputs at the cost of increased computational time.

Remarks 3. *This trade-off is a key characteristic of diffusion models, which inherently possess a flexible inductive bias that allows for varying degrees of commitment in generation. In contrast, autoregressive models are restricted to generating one token at a time.*
Moreover, under the flawed 32-bit sampling scheme, increasing the number of sampling steps effectively reduces the artificial temperature, thereby reducing the tokenwise entropy. In contrast, the tokenwise entropy remains mostly unaffected when the Gumbel trick is executed in double precision.

Configuration	**FP32-PPL** $\downarrow$	**FP64-PPL** $\downarrow$
$\rho = 16$ $(T = 64)$	75.57	100.11
$\rho = 8$ $(T = 128)$	64.90	95.38
$\rho = 4$ $(T = 256)$	53.02	81.37
$\rho = 2$ $(T = 512)$	44.15	79.3
$\rho = 1$ $(T = 1024)$	40.48	85.01
$\rho = \frac{1}{2}$ $(T = 2048)$	33.09	87.31
$\rho = \frac{1}{4}$ $(T = 4096)$	25.39	88.75
$\rho = \frac{1}{8}$ $(T = 8192)$	24.05	83.21

Table 8: Generative perplexities (PPL; lower is better) on OWT. The same very model (γ-Hybrid $[\gamma{=}\ 0.01, \boldsymbol{\tau}^{\omega=d/4}_{\text{Block}}, \text{ALIGNED}]$) has been used under different generation regimes. ρ value as well as equivalent T "diffusion steps" are used in the table. "FP32" denotes the flawed 32-bit sampling, whereas "FP64" corresponds to the corrected 64-bit precision values.

F.6 Effect of Varying η in Adaptive Correction Sampler

In this section, we investigate the effect of the hyperparameter η in our proposed Adaptive Correction Sampler. Table 9 illustrates results for our ϵ-variety family of models. As we increase the number of sampling steps (corresponding to a decrease in ρ), our model tends to overcorrect when η is too large, which ultimately harms generation diversity. To mitigate this issue, we find that using smaller η values is beneficial. We further suggest that the optimal choice of η is related to the ϵ value used during training: the more inherently corrective the model is, the smaller the optimal η should be.

Model Family	**Sampler**	MAUVE (↑)				Gen PPL. (↓)				Entropy (↑)			
		ρ=8	*ρ=4*	*ρ=2*	*ρ=1*	*ρ=8*	*ρ=4*	*ρ=2*	*ρ=1*	*ρ=8*	*ρ=4*	*ρ=2*	*ρ=1*
ϵ-Hybrid $[\epsilon{=}\ 0.01, \boldsymbol{\tau}_{\text{Flat}}, \text{ALIGNED}]$	Original Sampler	0.950	0.944	0.848	0.779	130.78	124.75	121.90	129.52	**5.51**	5.47	**5.49**	**5.50**
	ACS ($\eta = 0.25$)	**0.955**	0.821	0.859	0.928	**79.64**	65.06	**55.05**	49.09	5.35	5.28	5.24	5.19
	ACS ($\eta = 0.05$)	0.846	0.99	0.865	0.936	105.94	93.91	83.83	77.16	5.46	5.48	5.31	5.29
	ACS ($\eta = 0.01$)	0.848	0.928	0.957	0.947	84.28	69.84	61.35	**43.98**	5.36	5.33	5.28	5.18
	ACS ($\eta = 0.001$)	0.871	**0.949**	**0.919**	**0.998**	80.37	**64.42**	55.48	45.80	5.35	5.31	5.25	5.15
ϵ-Hybrid $[\epsilon{=}\ 0.01, \boldsymbol{\tau}^{\omega=d/4}_{\text{Block}}, \text{ALIGNED}]$	Original Sampler	0.916	**0.976**	0.778	0.847	148.45	130.16	139.64	142.13	5.39	5.35	5.43	5.46
	ACS ($\eta = 0.25$)	0.962	0.948	0.652	0.653	112.87	**64.01**	**54.67**	**43.32**	5.34	5.13	5.01	4.78
	ACS ($\eta = 0.05$)	0.964	0.811	0.813	0.916	**104.26**	76.98	71.77	59.15	5.42	5.25	5.38	5.25
	ACS ($\eta = 0.01$)	0.568	0.746	0.767	**0.974**	144.71	107.06	101.30	75.91	5.53	5.30	5.38	5.35
	ACS ($\eta = 0.001$)	**0.979**	0.847	**0.977**	0.906	150.41	150.32	139.66	114.12	**5.48**	**5.54**	**5.55**	**5.48**

Table 9: Sample quality of ϵ-variant models using different samplers.

F.7 INFERENCE SPEED UP WITH CACHING

In Table 10, we report the wallclock time required to generate eight samples for various models on a single *NVIDIA A100 80 GB* GPU. Owing to our custom attention masks, we were unable to leverage fast transformer kernels such as Flash-Attention, which in turn results in slower sampling speeds. We note that future work may design specialized kernels that are compatible with our attention masks to accelerate inference.

Our $\boldsymbol{\tau}^{\omega}_{\text{Block}}$ models are intrinsically faster since, during sampling, only the tokens corresponding to the τ_{settled} and τ_{active} components are fed into the transformer (e.g., ω, 2ω, $\ldots$, up to d tokens). Moreover, incorporating KV-caching would further boost the sampling speed.

	$Time(\downarrow)$
SEDD	68
MDLM	59
+ caching	36
γ-Hybrid $_{[\boldsymbol{\tau}_{\text{Flat}}, \text{ ALIGNED}]}$	131
γ-Hybrid $_{[\boldsymbol{\tau}^{\omega=d/4}_{\text{Block}}, \text{ ALIGNED}]}$	79
+ KV-caching	38

Table 10: Wall clock time reported in seconds to generate 8 samples on a single `NVIDIA A100 80 GB` GPU. The same number of diffusion steps were utilized for all the models.

G SAMPLES

The following is a random selection of a few samples from our 3 diffusion language model families.

G.1 UNDONDITONAL SAMPLE

Twenty-something host will soon have a new spot at a main commercial-run drugmakers' clinic in his home state of New Mexico.

Chris Brewster is convinced that it could lead to the world market for Type II's genetically modified metallomics drugs from certain not-too-famous contenders elsewhere in the medical marijuana industry who have partnered up with the cannabis company Little Troy Farms to roll out FDA approval for the widely used CRISPR-free Jacoby-Glynda hybrid product.

At least three manufacturers have been selected for the enterprising position, which also includes Schluge Therapeutics, Porsale, Mosaic, Marin and IE Pharma. 'The total number of drug prices in America in 2013 had the highest in any single year.'

Prasepalan, Dr. Dave Rouzeau's permission to begin selling medicine has long been a crusader for medical marijuana and these days, Dr. Brewster and Dr. Rimelter are taking their criminal flair to the next level.

Rouxau and his fellow New Mexico drug regulatory experts will be making their cases in October, which would complete a 90-page interface presenting the medical world, checks and balances and system reviews for each company's patented drugs involved in overall delivery by patient care providers and regulators.

"We have seen this is a breakthrough that patents have exploded," said Johnson, who represents Prasepalan's chiropractor, Adderall, and his partner, Rick Kelbyck, who specializes in pharmaceutical treatment and drug treatment products at Wellesley and Kroll Laboratories, among other giants, in Austin.

"We do have a case to make," and an additional 12 to 25 miles away from Prasepalan, Dr. Hansen will charge a direct fee to Symantec to make the prescription of Nathur Shemogood's Marin-REX Therapeutics-Lucenti, a New Mexico-based machine that has shipped a billion dollars worth of medicine to every business in the United States and Canada.¡—endoftext—¿No man on the verge of retirement must fear the official automated deportation system, fair trade prosecutor Rafaela Gonzalez thinks.

D.C.—a remote northern island nation governed by rigid immigration rules with 90 million illegal immigrants, 85 million people without adequate food, and 1 million illegal, unauthorized immigrants—that will soon surge around the United States?

Gonzalez got the idea from a journalist at a Caution Center paper to mark the anniversary of the United Nation's signing of the massive Comprehensive Economic and Policy Agreement (COP21) in 1995. Typically it's a minor-level agreement that provides an objective, stable key equation wherein every country, every country can control both the economy and country's external affairs by changing its immigration and customs policy. Democracy and separation of the people and law enforcement.

In March of this year, Gonzalez released a new chapter in her career-spanning memoir Magic of Arrows: The Filibuster of Gang Targets and Promise of Perfection. Modeled after a projection of long-term economic effects from government handouts to its leaflet, her book is dated May 21–24: Four days before America secedes from formally abolishing the six-nation customs union, America's ex-military leaders vow to force America to break off the military response to escalating pro-secularism.

The first chapter is rated as one of the first lists of the U.S. Category A countries for its purposes, code for "unfair advantage," with a stylized listing of 22 whose influence has been ascendant since the dawn of capitalism. As Gonzalez points out, the list displays the position of bodies such as United States Samoa and Croatia, whose remotes are sprinkled with their own market share and have expressed renewed interest in reaching out to countries like the BRICS, a 33-nation grouping that then can take tougher measures against those competitors.

"There are tremendous opportunities," Gonzalez writes. "From a Washington perspective, it's easy to see players–the world's poorest, country's most vulnerable–already feeling the need for a more calculated and conscientious policy in regard to smuggling drugs and other illegal trade into our midst."

D.C. is not alone: The Pan American Conference of Scholars announced that more than 3,000 countries sent almost 1,000 requests for tickets, and more than 17,000 applications for tickets have been denied since March of this year. Last May, the Selena–Ranich Trade Union Confederation—the country's only market, for which agribusiness corporations receive billions of dollars of grants annually—received a "delusion team response" of 6,000 applicants in 81 international training centres.

Robert Dellinger, deputy director for the International Union of Red Cross Office for Latin American Mission who awaits a decision in the file

Figure 12: Unconditional samples generated by γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Flat}}$, ALIGNED] trained on **OWT** dataset.

SALT LAKE CITY — The Utah Legislature committee chairwoman says it is a problem to persuade the state to legalize drug behavior. Wroeff Bugler reports for the Salt Lake Tribune Jan. 7, 2014 The bill's sponsor, Rep. Mike Tate, chairman of the Utah Business Improvement Association, says they would legalize the use of drugs "in a variety of policy areas." A Republican lawmaker called them to say a ban on the use of marijuana is the best way to tackle the problems of drug rights. (Photo: Cmdr. Roger Sultanousi)

According to a lawmakers 16-30 majority, the bill has the support of animal rights groups and equal representation. The bill would allow adult humans from nearly all biological categories to get arrested if they haven't already committed crimes and other vulnerable individuals who could be facing significant financial penalties. Utah's enactur percent has only 40 percent of all parental consent for an adult; that's not as far as human children go, but nearly half of teenagers are aborted at 18 months of age and illegal background checks don't exist for any other concern.

"Now focus on the serious crime aspect of the bill," spokeswoman Heather Chien said at the time. "Our believe hearing the issue of having the guy using the drug would be beneficial not just for the child, but the offender."

It comes as more and more farmers and livestock owners are scheduling change reviews. The Senate voted on Thursday to either let best-information request a vote on the pass or create a freedom of information request prioritized specifically by the state's enforcement motion, 54-12, to allow the vote by a vote of more than two to one.

NEWSLETTERS Get the AZ Memo newsletter delivered to your inbox We're sorry, but something went wrong Get the pulse of Arizona — Local news, in-depth state coverage and what it all means for you Please try again soon, or contact Customer Service at 1-800-332-6733. Delivery: Mon-Fri Invalid email address Thank you! You're almost signed up for AZ Memo Keep an eye out for an email to confirm your newsletter registration. More newsletters

In answer to a question posed by friend, lobbyist, and political activist Ed Priderout, this bill ultimately failed.

Though it remains to be seen what will happen in the future, a resolution Friday would call for the use of fewer than 5 percent of all adult diapers in a single year for the state's menial and surgical courier services.

The measure was also a tamper-resistant rabble fruit without trans fats.

"Despite some objections we need to make the safety mechanism work in a future we don't believe is most effective, and that's just who they are as I've been trying to get our labor. We have to use a combination of dog tags and earplugs, so that we can improve our conditions," said Rep. David Cale-Dero.

Copyright 2015 The Salt Lake Tribune.¡—endoftext—¿CALGARY — It was the two-day drive north through the republic – Ontario and the provinces – over the past three months that signs giving freemen extraordinary tourism in the province appear to have slipped away, leaving many of the seasonal magnetes at a bottom.

Really?

Rogue tards are wasting time in Canadians and anxieties are getting better.

Conspiracists have become "emerged," as experts characterize them, in a new sense, either as congenial naïvetés – quasi-religious adventurous sluts – rather than numb, territorial isolates like northern Canadians who are already born in the same spot, or simply as less flexible and maneuvering players that can deliver on their hard economic-mindedly defined tenets. Certain politicians are reading the Pearls Bible and happy that they lack cable, satellite TV or otherwise space-time worthy support in the province's often futile re-election campaigns. Yet many Canadians see the Alberta premier as more philosophical than conservative.

Some have joined the "unreliable" view that there is another side to the tale.

Amplification over the provincial election left Coyote Falls and Tire Centre, Conservative and Labor alike, shared the view a month ago on the periphery of the serenity of the Alberta electorate that the drums were driving on the Alberta government's provincial campaign. But the NDP MLA was also shaking that opinion with equal dismay when Ford's demand that union representation be halted by the federal government for the next two years yielded a somewhat sympathetic "no."

Article Continued Below

And say many of the former leaders want more safeguards in the build-up of unions. One, minister Clifton Sanderson, acknowledged his former colleagues had brought with them new rules, regulations and so on that could be in their path. But, it remains a matter of interpretation.

"We welcome that a promise of transparency and clarity – that

Figure 13: Unconditional samples generated by γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}_{\text{Slide}}^{\omega = d/4}$, ALIGNED] trained on **OWT** dataset.

White House chief of staff Reince Priebus traveled to Washington, D.C., on Wednesday to be briefed by Vice President Mike Pence and other top advisers about the nature of the crisis in Ukraine and its potential impact on the United States.
The new official, who was not authorized to speak publicly by the White House, met with senior American Russian policy experts and other world leaders as part of a trip to Moscow to meet with Ukrainian leader Mykola Azarov. The trip comes as tensions peak over President Donald Trump's handling of the refugee crisis and its potential for conflict with Russia.
Jared Kushner's efforts to pump up the Dubrovnik deal have raised concerns that it could backfire. Kushner Trump is accused of colluding with the Kremlin, according to a New York Times report .
White House officials said that press secretary Sean Spicer called Kushner a "little stranger" when asked if he was familiar with the concerns swirling around the discussions, according to an account published by NBC News .
"The president-elect agreed to be briefed by my team on Russian meddling in our election process," Press Secretary Sean Spicer told NBC News in a telephone call, adding that he didn't have an authority to disclose details of the discussions.
Versions of the meeting have drawn criticism from U.S. politicians, military officials and human rights advocates.
In a separate report, former acting FBI Director Andrew McCabe said he was not aware of any conversations with any Russian officials. He has said he did not know why the advisers were so closely involved in this scheduling conflict.
"I use very preliminary, telephone calls with colleagues both within our own department and overseas, and I have not had conversations with any of them," McCabe said of the meetings, which included senior officials and a White House official.
The meetings involve joint efforts by the White House and individual individuals with the understanding that such a meeting would not occur, McCabe said, adding that no other member of the senior staff was involved.
"Please, contact the people you know with whom you have the knowledge in this room. Basically, I am asking you to do the work of reaching out to each one of the top officials associated with your Department of State," McCabe said.
"Since this is tense, I have trying to contact both the President and his staff and conduct the interview," McCabe said.
Guidance
White House advisers held a series of meetings with top officials of the Department of State and its federal counterparts. However, the White House's meetings were themselves influenced by events to help the new administration rule out greater involvement in an escalating crisis that has killed more than 200,000 lives since last August.
In speeches delivered to the nation, Trump touted Friday's meeting on June 20 as another reminder of the importance of new sanctions against Russia. Trump said Russia's nuclear weapons "are insulting weak countries" and hoped "for a new normal" after a nosedive in the U.S. elections.
"We have a great relationship with Russia, and we continue to look forward to strengthening our relationship with them," said Trump.
The "warm" Russia policy has deepened the already long-standing divisions over the Obama administration, and Putin's recent ally Russia, who built a buffer zone along the Black Sea in 2013, ordered an invasion in 2014 of eastern Ukraine.
White House spokesman Sean Spicer said he was "careful" with ties with Russia.¡—endoftext—¿On Tuesday night, LeBron James received news he was going — so often — from the Hall of Fame. Cleveland's most respected individual, who most known for his love of basketball, was getting a heart-to-heart with one of the most prestigious prospects in pro basketball history, sitting in the stands as the Knicks' first guard of the year. The official photo for the Knicks' current superstar, George Conte, was taken by the longtime reality television legend Wade Boggs.
We demand James get a heart. Exactly what he deserves is hard to guess. People to the point a man taken to heart who never weeks ago was the kind of person a champion of truth would want to meet or make close friends with without his inner purpose. After all, people feel very similar if they haven't seen Wade Boggs' latest film, Clump, in which he takes matters into his own hands and seizes his wife — the same wife that never had a child up for adoption.

This event has the potential to be a sour revenge for the Kevin Durant coronation that began a week ago – a winner's assuredly standard season that had little impact on James' looming future. But that's not the case here, either; the only one he is saying is that the uninspiring

Figure 14: Unconditional samples generated by γ-Hybrid [$\gamma = 0.01, \boldsymbol{\tau}^{\omega=d/64}_{\text{Block}}$, SHIFTED] trained on **OWT** dataset.

be violent might be protected for one of the reasons he challenged.
Tsherberg is talking about former Republican Rep. Joe Burns who, in many ways tried to do better politically, canceled my appearance in 2012. She said that when the Republican president and we were playing with one other's decisions, it was the Chechen deal.
"Every time I think, there's nothing in the issue," she said. "There comes a time when I faced with reality in a very difficult way. It's not something I want to put the side of the my team past me."
During O'Reilly Factor in December 2011, when I dealt with most of my viewers, I watched the Obama Britain show again.
"I was watching very, very closely, knowing that there was an opportunity, I knew I didn't want to be able to be able to help the Americans. I couldn't help that," I said.
Cack Obama also said several people without any authority, just due diligence, have commented on his plan to return to Chechnistan and Russia. T. Obama said he's not against Putin, and referred to himself as Putin-PM, who said he's taking "America's values forward."
"I think that's a turning point," Iza said in 2008 just before the start of my comments, referring to Iza's parent party, Moscow's Communist Party, in 1997, according to The Associated Press.
Asked whether Iza believes the economy and its security should run well into the Chechen agreement, and the leader ISIS was this month, Iza said that Trump is a far to go, more conservative agenda.
"He can't be a politician — he's not a tough guy who doesn't take a person," she said. "My people have respect too. My children are not civil andRoll but they do not [acc believe me]."
After speaking at Trump's Palm Beach, a bus stop in Texas on Sunday, his attorney, Republican Frank Agoodz, said neither he nor Iza accused me of arguing at the "Maxby Restaurant" in Texas.
"Williams, I want them to help, I don't want them to be able to help me," Iza was recorded as saying.
"He said, 'You have an opportunity in doing this and then you have to do this, you can do this; it's starting, that's getting the right people to do it, and he said, 'They're going and did, yeah,,'" she tells The Associated Press on Friday evening. Sokherberg had not prior known Iza's release — since his public hearing, he hasn't has been expecting anything since then.
But Iza said this is a situation that will cause "Great pain" for all his family.
"It's important to my well-being, to care for others, about peace and just to be strong," she said. "As long as I don't, I won't. I think if I won't, I will have no apologies."
Their representatives before they gather on Saturday.
"It is a man with the very strong side of things, that has a long way to go," she said. "He's been singled out by the President. He has turned even more extreme off the map and is being led by the Republican party and the economy that are making a debt-laden economy out of his dream.
"The public is a mind-set, and his decision to speak to the public was a mind-set."
In a CNN video excerpt that includes her closing remarks, the then-U.S. Senate wasift to safety to stand in front of Republican House-time House Speaker and former White House special counsel Paul Pro Romanoff.¡—endoftext—¿New Villa in the midst of not making a cause in Sunday's clash at Villa, who started toestern to the title last season.
With Villa attacking you no certainly doubt it, if it was the second title they had won in the most recent five years, over the likes of Marc Canado, Robbie, Fernandes and CFOi.
But Villa think an answer to that battle over Canado is the help of the Vy's and in speeding up the relationship with the Portuguese forward in ways he hasn't shown in his past season.
West Villa turned their attention to defence in with Canado, who represented the England national team even last season after he became the first ever to score four goals.
He went on the same defensive side though, is now holding up front to rival one of the top midfielders and attack in the second half.
West Villa's second-half form last season however, has been done again in favour to substitute Scott-Fidane Jr, who later moved to Paris City on Wednesday.

John Gomez and João Aguero have moved to

Figure 15: Unconditional samples generated by ϵ-Hybrid [$\epsilon = 0.01, \boldsymbol{\tau}_{\text{Flat}}$, ALIGNED] trained on **OWT** dataset.

G.2 Conditional Samples

The cat sat on the mat and looked up, staring at the ceiling as if something interesting were up there. It made no sound, but occasionally flicked its tail back and forth.
"Is he always like this?" I asked.
She smiled and nodded. "Pretty much. Sometimes he stares out the window too."
The cat glanced at me, seemed to consider something deeply philosophical for a moment, and then resumed staring upward.
"Do you think he sees things we don't?" I asked, half-joking.
"Probably," she replied, laughing softly. "Cats always seem to have a foot in another world, don't they?"
I chuckled, taking a sip from my coffee. "Maybe we should take notes."
"Maybe we should," she said, still smiling. "We might learn something."

The cat yawned lazily, stretched, and settled even more comfortably onto the mat, clearly deciding that whatever secrets the universe held could wait a little longer.

Figure 16: Conditional (conditioned on first 6 tokens) samples generated by ϵ-Hybrid [$\epsilon = 0.01, \boldsymbol{\tau}_{\text{Flat}}$, ALIGNED] trained on **OWT** dataset.