Block Diffusion: Interpolating Between Autoregressive and Diffusion Language Models

Ntountounakis Georgios, Markoulidakis Georgios, Vitalis Petros, Makras Ilias, Kritharidis Konstantinos, Kordas Nikolaos

Pattern Recognition, ECE
National Technical University of Athens

February 2026

Table of Contents

Table of Contents

Introduction to the Problem-Motivation

Two main approaches for Language Models:

Autoregressive (AR):

- Token-by-token generation
- High quality
- KV caching
- Variable length

Diffusion:

- Parallel generation
- Better controllability
- Fixed length (limitation)
- Lower quality (Perplexity Gap)

Question

Can we combine the advantages of both approaches?

Core Idea: Block Diffusion

Block 1 → Block 2 → Block 3 $\cdots$

Diffusion within each block(parallel)
Autoregressive over blocks

Parameterization: Trade-off through block size L':

- $L' = 1 \rightarrow$ Pure AR
- $L' = L \rightarrow$ Pure Diffusion

Technical Contribution:

- Optimized training and sampling algorithms
- Introduced clipped noise schedules for reduced gradient variance during training
- SoTA PPL among diffusion models + Variable length generation capabilities

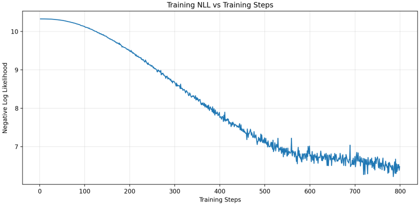
Table of Contents

Table of Contents

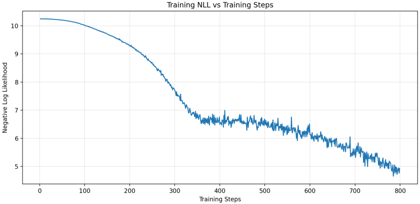
AR vs BD3LM with L'=1

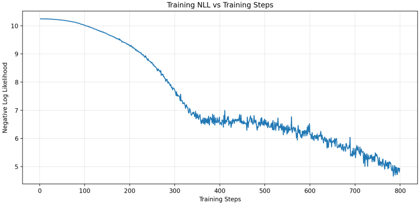
Test Perplexities for single token generation on LM1B dataset (800 Training Steps)

	PPL (↓)
Autoregressive	**1893**
BD3LM L'=1	2231
BD3LM L'=1 + Tuned Schedule	2220



AR





The Effect of Clipped Noise Schedules

Test Perplexities for single token generation on LM1B dataset
(400 Pretraining Steps + 100 Fine-tuning Steps)

L'	Clipping	PPL	Var. NELBO
128	U [0, 0.5]	1000	10.00
	U [0, 1]	1000	10.00
16	U [0.3, 0.8]	**1278**	**10.50**
	U [0, 1]	1279	10.51
4	U [0.5, 1]	**1226**	**44.41**
	U [0, 1]	**1226**	**44.41**

Commentary (placeholder)

Key takeaways:

- *???*
- *???*
- *???*

Conclusion...

BD3LMs vs ARs vs Diffusion Models on LM1B

Test perplexities (PPL; ↓) of models on LM1B.
(400 Pretraining Steps + 100 Fine-tuning Steps)

		PPL (↓)
Autoregressive		
Transformer		3042
Diffusion		
SEDD		1447
MDLM		1616
Block diffusion (Ours)		
BD3-LMs	$L' = 16$	1278
	$L' = 8$	1734
	$L' = 4$	**1226**

Commentary (placeholder)

Key takeaways:

- *???*
- *???*
- *???*

Conclusion...

BD3LMs vs ARs vs Diffusion Models on OWT

Test perplexities (PPL; $\downarrow$) of models on OWT.
(800 Pretraining Steps + 800 Fine-tuning Steps)

		PPL ($\downarrow$)
Autoregressive		
Transformer		2036
Diffusion		
SEDD		2120
MDLM		2101
Block diffusion (Ours)		
BD3-LMs	$L' = 16$	1939
	$L' = 8$	1941
	$L' = 4$	**1935**

Commentary (placeholder)

Key takeaways:

- *???*
- *???*
- *???*

Conclusion...

Performance on other Datasets

Zero-shot validation perplexities (↓) of models trained on OWT.(?training steps?)

	LM1B	Lambada	Wikitext
AR	2388	1550	**2875**
SEDD	2742	1562	3335
MDLM	2722	1556	3283
BD3-LM $L' = 4$	**2196**	**1438**	3143

Commentary (placeholder)

Key takeaways:

- *??? (e.g., BD3-LM improves on …compared to diffusion baselines)*
- *??? (e.g., note where AR is still best / gap remains)*
- *??? (e.g., mention that diffusion values are upper bounds if relevant)*

Conclusion…

Variable-Length Sequence Generation

Generation length statistics from sampling 10 documents from models trained on OWT.(?training steps?)

	Median # tokens	Max # tokens
OWT train set	717	131K
AR	4008	131K
SEDD	1021	1024
BD3-LM $L' = 16$	798	2927

Commentary (placeholder)

Key takeaways:

- *??? (e.g., diffusion models are constrained in max length, BD alleviates this)*
- *??? (e.g., compare median vs max lengths across AR/SEDD/BD3)*
- *??? (e.g., mention practical implication for long-form generation)*

Conclusion...

Sample Quality

Generative Perplexity (Gen.PPL;↓) and number of Function Evaluations (NFEs;↓) of 300 samples. All models are trained on OWT. (?training steps?)

	$L = 1024$		$L = 2048$	
Model	Gen. PPL	NFEs	Gen. PPL	NFEs
AR	14.1	1K	13.2	2K
Diffusion				
SEDD	52.0	1K	–	–
MDLM	46.8	1K	41.3	2K
Block Diffusion				
SSD-LM $L' = 25$	37.2	40K	35.3	80K
281.3	1K	281.9	2K	
BD3-LMs $L' = 16$	32.97	1K	31.42	2K
$L' = 8$	29.35	1K	27.42	2K

Effect of Different Noise Schedules

Effect of noise schedule on PPL and variance of NELBO for different L' on LM1B. (?training steps?)

Noise schedule	PPL	Var. NELBO
$L' = 4$		
Clipped		
$\mathcal{U}[0.45, 0.95]$	**29.21**	**6.24**
$\mathcal{U}[0.3, 0.8]$	29.38	10.33
Linear $\mathcal{U}[0, 1]$	30.18	23.45
Logarithmic	30.36	23.53
Square root	31.41	26.43
$L' = 16$		
Clipped		
$\mathcal{U}[0.45, 0.95]$	31.42	3.60
$\mathcal{U}[0.3, 0.8]$	**31.12**	**3.58**
Linear $\mathcal{U}[0, 1]$	31.72	7.62
Square	31.43	13.03
Cosine	31.41	13.00

Commentary (placeholder)

Key takeaways:

- *???*
- *???*
- *???*

Conclusion...

Table of Contents

Noise Scheduling in Masked Diffusion (Summary)

Continuous time index:

$$t \sim \mathcal{U}[0,1]$$

Noise schedule $\Rightarrow$ masked probability:

$$p(t): \text{ masked probability induced by the noise schedule}$$

Keep (no-mask) probability:

$$a(t) = 1 - p(t)$$

Loss scaling induced by the schedule:

$$\text{loss scaling}(t) \;=\; \frac{a'(t)}{1 - a(t)}$$

Intuition

The noise schedule sets *where* the model learns via $p(t)$ (masking rate). When sampling $t \sim \mathcal{U}[0,1]$, loss_scaling$(t) = \frac{a'(t)}{1-a(t)}$ acts as a weight on the per-token log-likelihood term so the discrete estimator matches the continuous-time integral/ELBO and balances low- vs high-noise regions.

Already Implemented Noise Schedules

Schedule	$p(t)$	**loss_scaling**(t)
LogLinear	t	$-\frac{1}{t}$
Square	t^2	$-\frac{2}{t}$
Square root	$t^{0.5}$	$-\frac{1}{2t}$
Logarithmic	$\frac{\log(1+t)}{\log 2}$	$-\frac{1}{(1+t)\,\log(1+t)}$
Cosine	$1-(1-\varepsilon)\cos\left(\frac{\pi t}{2}\right)$	$-\frac{\left(\frac{\pi}{2}\right)(1-\varepsilon)\sin\left(\frac{\pi t}{2}\right)}{1-(1-\varepsilon)\cos\left(\frac{\pi t}{2}\right)}$

Gaussian & Bimodal Gaussian Noise Schedules

Goal: sample a masked probability $p(t) \in (0,1)$ from $t \sim \mathcal{U}[0,1]$.

Gaussian schedule (truncated to $(0,1)$):
Let $\alpha = \frac{0-\mu}{\sigma}$, $\beta = \frac{1-\mu}{\sigma}$, $\Phi_\alpha = \Phi(\alpha)$, $\Phi_\beta = \Phi(\beta)$. For $t \in (0,1)$:

$$z(t) = \Phi^{-1}\big(\Phi_\alpha + t(\Phi_\beta - \Phi_\alpha)\big), \qquad p(t) = \mu + \sigma z(t) \in (0,1).$$

With $Z = \Phi_\beta - \Phi_\alpha$ and $\varphi(\cdot)$ the standard normal pdf:

$$\text{loss\_scaling}(t) = \frac{a'(t)}{1-a(t)} = -\frac{p'(t)}{p(t)} = -\frac{\sigma Z}{\varphi(z(t))\, p(t)}.$$

Bimodal Gaussian schedule (mixture):
Choose a split weight $w \in (0,1)$ (denote $w_1 = w$, $w_2 = 1-w$). With probability w use (μ_1, σ_1), otherwise use (μ_2, σ_2):

$$t_1 = \frac{t}{w_1} \;\; (t < w_1), \qquad t_2 = \frac{t - w_1}{w_2} \;\; (t \geq w_1), \qquad p(t) = \begin{cases} \mu_1 + \sigma_1 z_1(t_1), & t < w_1 \\ \mu_2 + \sigma_2 z_2(t_2), & t \geq w_1 \end{cases}$$

where each $z_i(\cdot)$ is defined as above (with its own $\alpha_i, \beta_i, \Phi_{\alpha_i}, \Phi_{\beta_i}$ and Z_i). The resulting scaling is piecewise:

$$\text{loss\_scaling}(t) = -\frac{1}{p(t)} \begin{cases} \frac{1}{w_1} \frac{\sigma_1 Z_1}{\varphi(z_1(t_1))}, & t < w_1 \\ \frac{1}{w_2} \frac{\sigma_2 Z_2}{\varphi(z_2(t_2))}, & t \geq w_1 \end{cases}$$

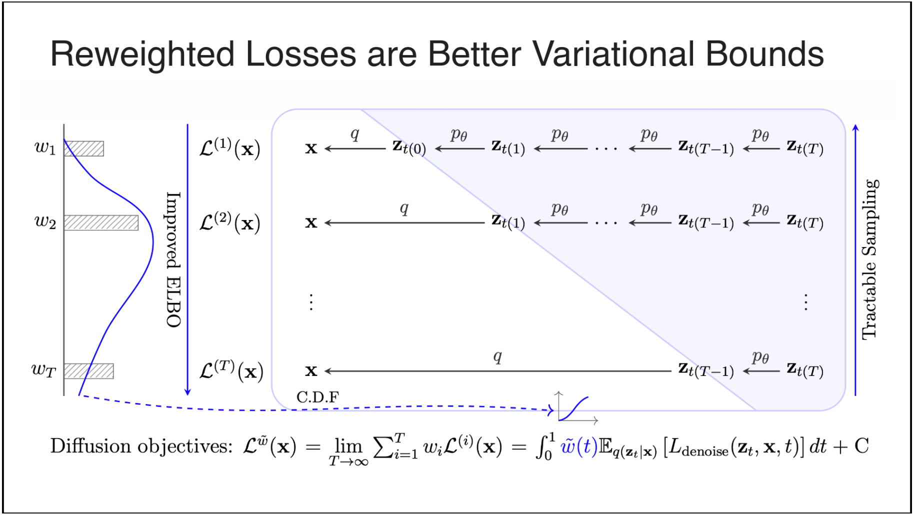
Exploring New Schedules

- Motivation for designing alternative noise schedules
- Trade-off between stability and generation quality
- Impact on gradient variance
- Compatibility with block diffusion framework

New Schedules' Results

- Comparison with cosine and clipped schedules
- Effects on perplexity
- Training stability observations
- Sampling behavior differences

Reweighted Losses for Improved Diffusion Objective



Reweighted Losses for Masked Diffusion

- Initial Reweighted NELBO:

$$\mathcal{L}^{\tilde{w}}(\mathbf{x}) = -\int_0^1 \tilde{w}(t) \frac{\alpha'_t}{1-\alpha_t} \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})} \left[\delta_{\mathbf{z}_t, m} \cdot \mathbf{x}^\top \log \mu_\theta(\mathbf{z}_t) \right] \mathrm{d}t$$

- Reparameterization trick: $\lambda(t) = \log \frac{\alpha_t}{1-\alpha_t}$:

$$\mathcal{L}^{\hat{w}}(\mathbf{x}) = -\int_0^1 \hat{w}(\lambda(t)) \frac{\alpha'_t}{1-\alpha_t} \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})} \left[\delta_{\mathbf{z}_t, m} \cdot \mathbf{x}^\top \log \mu_\theta(\mathbf{z}_t) \right] \mathrm{d}t$$

Name	$\lambda(t)$	$\hat{w}(\lambda)$	$\tilde{w}(t)$
EDM		$p_{\mathcal{N}(2.4, 2.4^2)}(\lambda) \frac{e^{-\lambda}+0.5^2}{0.5^2}$	$w(\lambda(t))$
IDDPM		$\mathrm{sech}(\frac{\lambda}{2})$	$2\sqrt{\alpha_t(1-\alpha_t)}$
Sigmoid	$\log \frac{\alpha_t}{1-\alpha_t}$	$\mathrm{sigmoid}(-\lambda + k)$	$\frac{1-\alpha_t}{1-(1-e^{-k})\alpha_t}$
FM		$e^{-\frac{\lambda}{2}}$	$\sqrt{\frac{1-\alpha_t}{\alpha_t}}$
Simple		-	$-\frac{1-\alpha_t}{\alpha'_t}$

Reweighted Losses Results

Extended Table 3: Test Perplexities

	PPL (↓)					
Autoregressive						
Transformer	1221					
Diffusion						
SEDD	1403					
MDLM	1370					
Block diffusion	**Base**	**IDDPM**	**EDM**	**Sigmoid (k = 0)**	**FM**	**Simple**
BD3-LMs $L' = 16$	1345	252	49.88	36.06	76213	53070
$L' = 8$	1210	249	49.14	35.79	109169	36010741760
$L' = 4$	1176	246	49.01	**35.08**	67332	2396260

Table of Contents

Table of Contents

Table of Contents