

# 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

- 1 Paper Overview
- 2 Our Results
  - Reproduction
  - Extensions
- 3 Conclusion & Future Work
- 4 Team Organization
- 5 Retrospection

# Table of Contents

## 1 Paper Overview

## 2 Our Results

- Reproduction
- Extensions

## 3 Conclusion & Future Work

## 4 Team Organization

## 5 Retrospection

# Introduction to the Problem-Motivation

## Two main approaches for Language Models:

### Autoregressive (AR):

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

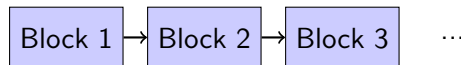
### Diffusion:

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

### Question

Can we combine the advantages of both approaches?

# Core Idea: Block Diffusion



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

1 Paper Overview

2 Our Results

- Reproduction
- Extensions

3 Conclusion & Future Work

4 Team Organization

5 Retrospection

# Table of Contents

1 Paper Overview

2 Our Results

- Reproduction
- Extensions

3 Conclusion & Future Work

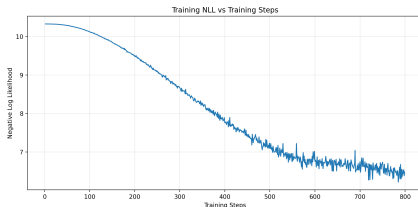
4 Team Organization

5 Retrospection

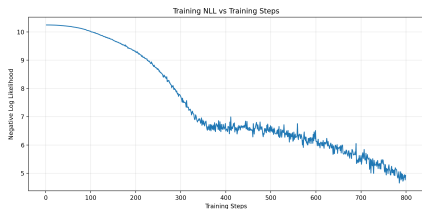
# AR vs BD3LM with $L'=1$

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

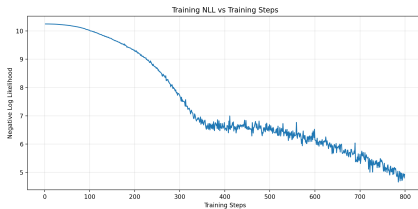
	PPL (↓)
<b>Autoregressive</b>	<b>1893</b>
<b>BD3LM <math>L'=1</math></b>	<b>2231</b>
<b>BD3LM <math>L'=1</math> + Tuned Schedule</b>	<b>2220</b>



AR



BD3LM



BD3LM + Tuned Schedule





# 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	$\mathcal{U}[0, 0.5]$	1000	10.00
	$\mathcal{U}[0, 1]$	1000	10.00
16	$\mathcal{U}[0.3, 0.8]$	<b>1278</b>	<b>10.50</b>
	$\mathcal{U}[0, 1]$	1279	10.51
4	$\mathcal{U}[0.5, 1]$	<b>1226</b>	<b>44.41</b>
	$\mathcal{U}[0, 1]$	<b>1226</b>	<b>44.41</b>

## 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 (↓)
<b>Autoregressive</b>	
Transformer	3042
<b>Diffusion</b>	
SEDD	1447
MDLM	1616
<b>Block diffusion (Ours)</b>	
BD3-LMs $L' = 16$	1278
$L' = 8$	1734
$L' = 4$	<b>1226</b>

## Commentary (placeholder)

### Key takeaways:

- ???
- ???
- ???

*Conclusion...*

# BD3LMs vs ARs vs Diffusion Models on OWT

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

	PPL (↓)
<b>Autoregressive</b>	
Transformer	2036
<b>Diffusion</b>	
SEDD	2120
MDLM	2101
<b>Block diffusion (Ours)</b>	
BD3-LMs $L' = 16$	1939
$L' = 8$	1941
$L' = 4$	<b>1935</b>

## Commentary (placeholder)

### Key takeaways:

- ???
- ???
- ???

*Conclusion...*

# Performance on other Datasets

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

	LM1B	Lambada	Wikitext
<b>AR</b>	2388	1550	<b>2875</b>
SEDD	2742	1562	3335
MDLM	2722	1556	3283
BD3-LM $L' = 4$	<b>2196</b>	<b>1438</b>	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?)

Model	$L = 1024$		$L = 2048$	
	Gen. PPL	NFEs	Gen. PPL	NFEs
AR	14.1	1K	13.2	2K
<b>Diffusion</b>				
SEDD	52.0	1K	—	—
MDLM	46.8	1K	41.3	2K
<b>Block Diffusion</b>				
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
$L' = 4$	<b>24.74</b>	1K	<b>23.85</b>	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		
$\mathcal{U}[0.45, 0.95]$	1199.43	<b>15.2313</b>
$\mathcal{U}[0.3, 0.8]$	1101.03	27.6118
Linear $\mathcal{U}[0, 1]$	751.15	260.1035
Logarithmic	750.05	125.6219
Square root	<b>718.90</b>	83.832
$L' = 16$		
Clipped		
$\mathcal{U}[0.45, 0.95]$	1056.46	<b>4.3476</b>
$\mathcal{U}[0.3, 0.8]$	797.62	6.4124
Linear $\mathcal{U}[0, 1]$	661.99	44.7271
Square	<b>627.02</b>	27.5325
Cosine	633.59	20.8008

## Table 8 Reproduction

### Key takeaways:

- *Variance Hypothesis*  
*Confirmed:* Clipped schedules reduced NELBO variance by 17x (from 260.1 to 15.2), validating the method's stability.
- *Perplexity trade-off:* Unlike the paper, standard noise schedules achieve better PPL. Those noise schedules are probably easier to be learned faster but lack on convergence.

*The method effectively stabilizes training (low variance) immediately, but PPL gains from this stability likely require a longer training horizon*

# Table of Contents

1 Paper Overview

2 Our Results

- Reproduction
- Extensions

3 Conclusion & Future Work

4 Team Organization

5 Retrospection



# Noise Scheduling in Masked Diffusion (Summary)

**Continuous time index:**

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

**Noise schedule  $\Rightarrow$  masked probability:**

$p(t)$  : 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  $t \sim \mathcal{U}[0, 1]$ ,  $\text{loss\_scaling}(t) = \frac{a'(t)}{1 - a(t)}$  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  $t \in (0, 1)$ :

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

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  $(\mu_1, \sigma_1)$ , otherwise use  $(\mu_2, \sigma_2)$ :

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

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:

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

# New Schedules' Results

- Perplexities (PPL) and Var. NELBO for implemented vs new schedules on LM1B (400 Pretraining Steps + 100 Fine-tuning Steps).
- B.G. stands for Bimodal Gaussian.
- We use constant values:  $\sigma^2 = x$  (Gaussian) and  $\sigma_1^2 = x, \sigma_2^2 = x$  (B.G.) for all experiments. Training under varying  $\mu$  and  $\mu_1, w_1$  respectively.

Block Size	Already Implemented			Newly Implemented		
	Noise Schedule	PPL	Var. NELBO	Noise Schedule	PPL	Var. NELBO
128	Loglinear	1000	10.00	Gaussian ( $\mu = 0.5$ )	1000	10.00
	Loglinear $\mathcal{U}[0, 0.5]$	1000	10.00	Gaussian ( $\mu = 0.6$ )	1000	10.00
	Cosine	1000	10.00	B.G. ( $\mu_1 = 0.3, w_1 = 0.6$ )	1000	10.00
	Cosine $\mathcal{U}[0, 0.5]$	1000	10.00	B.G. ( $\mu_1 = 0.1, w_1 = 0.6$ )	1000	10.00
16	Loglinear	1279	10.50	Gaussian ( $\mu = 0.5$ )	<b>1234</b>	<b>10.28</b>
	Loglinear $\mathcal{U}[0.3, 0.8]$	1278	10.51	Gaussian ( $\mu = 0.6$ )	1235	10.31
	Cosine	1236	10.30	B.G. ( $\mu_1 = 0.3, w_1 = 0.6$ )	1254	10.42
	Cosine $\mathcal{U}[0.3, 0.8]$	1235	10.29	B.G. ( $\mu_1 = 0.1, w_1 = 0.6$ )	1295	10.59
4	Loglinear	<b>1226</b>	<b>44.41</b>	Gaussian ( $\mu = 0.5$ )	1250	46.76
	Loglinear $\mathcal{U}[0.5, 1]$	<b>1226</b>	<b>44.41</b>	Gaussian ( $\mu = 0.7$ )	1252	46.76
	Cosine	1228	45.28	B.G. ( $\mu_1 = 0.3, w_1 = 0.6$ )	1253	46.84
	Cosine $\mathcal{U}[0.5, 1]$	1229	45.26	B.G. ( $\mu_1 = 0.1, w_1 = 0.6$ )	1243	46.49

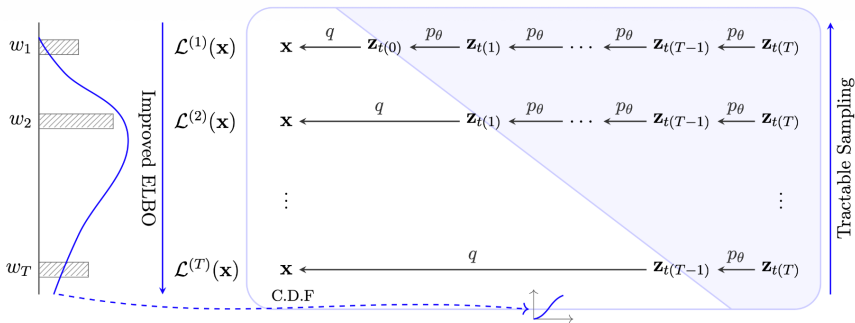
# Using Bimodal Gaussian on Pretraining

**We can also apply the Bimodal Gaussian noise schedule during pretraining.**

The table reports, for each pretraining noise schedule, the *best* fine-tuning noise schedule we found, with the corresponding PPL and Var. NELBO (400 Pretraining Steps + 100 Fine-tuning Steps).

Block Size	Pretraining Schedule	Fine-tuning Schedule	PPL	Var. NELBO
128	Loglinear	?????	1000	10.00
	Bimodal Gaussian	?????	1000	10.00
16	Loglinear	Gaussian ( $\mu = 0.5$ )	1234	10.28
	Bimodal Gaussian	Cosine $\mathcal{U}[0.3, 0.8]$	<b>1227</b>	<b>10.22</b>
4	Loglinear	Loglinear	1226	44.41
	Bimodal Gaussian	Loglinear	<b>1213</b>	<b>44.39</b>

## Reweighted Losses are Better Variational Bounds



$$\text{Diffusion objectives: } \mathcal{L}^{\bar{w}}(\mathbf{x}) = \lim_{T \rightarrow \infty} \sum_{i=1}^T w_i \mathcal{L}^{(i)}(\mathbf{x}) = \int_0^1 \tilde{w}(t) \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})} [L_{\text{denoise}}(\mathbf{z}_t, \mathbf{x}, t)] dt + C$$

# Reweight Losses for Masked Diffusion

- Initial Reweighted NELBO:

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

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

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

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	$\log \frac{\alpha_t}{1 - \alpha_t}$	$\text{sech}(\frac{\lambda}{2})$	$2\sqrt{\alpha_t(1 - \alpha_t)}$
Sigmoid		$\text{sigmoid}(-\lambda + k)$	$\frac{1 - \alpha_t}{1 - (1 - e^{-k})\alpha_t}$
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 (↓)							
<b>Autoregressive</b>							
Transformer		1221					
<b>Diffusion</b>							
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	<b>35.08</b>	67332	2396260



# Table of Contents

1 Paper Overview

2 Our Results

- Reproduction
- Extensions

3 Conclusion & Future Work

4 Team Organization

5 Retrospection

# Table of Contents

- 1 Paper Overview
- 2 Our Results
  - Reproduction
  - Extensions
- 3 Conclusion & Future Work
- 4 Team Organization
- 5 Retrospection

# Table of Contents

- 1 Paper Overview
- 2 Our Results
  - Reproduction
  - Extensions
- 3 Conclusion & Future Work
- 4 Team Organization
- 5 Retrospection