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

January 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 $\rightarrow$ Block 2 $\rightarrow$ 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

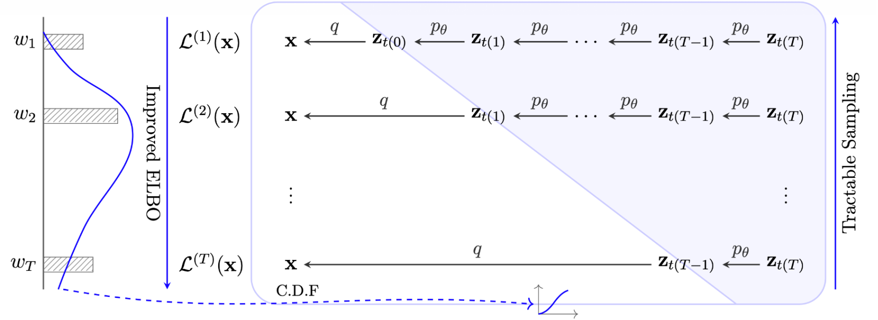
Table of Contents

Table of Contents

Table of Contents

Reweighted Losses for Improved Diffusion Objective

Reweighted Losses are Better Variational Bounds



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

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	$\log \frac{\alpha_t}{1-\alpha_t}$	$\mathrm{sech}(\frac{\lambda}{2})$	$2\sqrt{\alpha_t(1-\alpha_t)}$
Sigmoid		$\mathrm{sigmoid}(-\lambda+k)$	$\frac{1-\alpha_t}{1-(1-e^{-k})\alpha_t}$
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