



[https://github.com/
Despoinakk/babylm-diffusion](https://github.com/Despoinakk/babylm-diffusion)

Masked Diffusion Language Models with Frequency-Informed Training

Despoina Kosmopoulou ♠♡ Efthymios Georgiou ♦
 Vaggelis Dorovatas ♡ Georgios Paraskevopoulos ♣ Alexandros Potamianos ♠♡
 ♠National Technical University of Athens ♡Archimedes RU, Athena RC ♦University of Bern
 ♣Institute of Language and Signal Processing, Athena RC
 despoinakkosmopoulou@gmail.com efthymios.georgiou@unibe.ch

The *BabyLM Challenge* addresses a fundamental question:
 Can language models achieve human-like learning efficiency?

- **The Challenge:** State-of-the-art LMs train on trillions of tokens, while humans learn from <100M words by age 12
- **BabyLM 2025:** Train models on 100M words for 10 epochs under strict data constraints
- **Our Approach:** use Masked Diffusion Language Models (MDLMs) [1] with key masking strategy innovations
- **Key Insight:** Unified diffusion training objective with well-tuned noise schedules can match hybrid approaches [2]

Contributions

1. **Bimodal Noise Schedules:** Combine low and high masking rates to emulate traditional MLM and enhance generative capabilities
2. **Frequency-Informed Masking:** Prioritize rare tokens with curriculum learning to boost efficiency

1 Introducing Bimodal Noise Schedules

Goal: Balance MLM (low masking) and AR (high masking) benefits in one framework

Schedule	EWoK	BLiMP	BLiMP	Sup.
Uniform	51.98	77.91	67.63	
Cosine	52.44	79.05	70.74	
Bimodal Gauss. ($\gamma = 0.0$)	52.95	78.28	73.13	

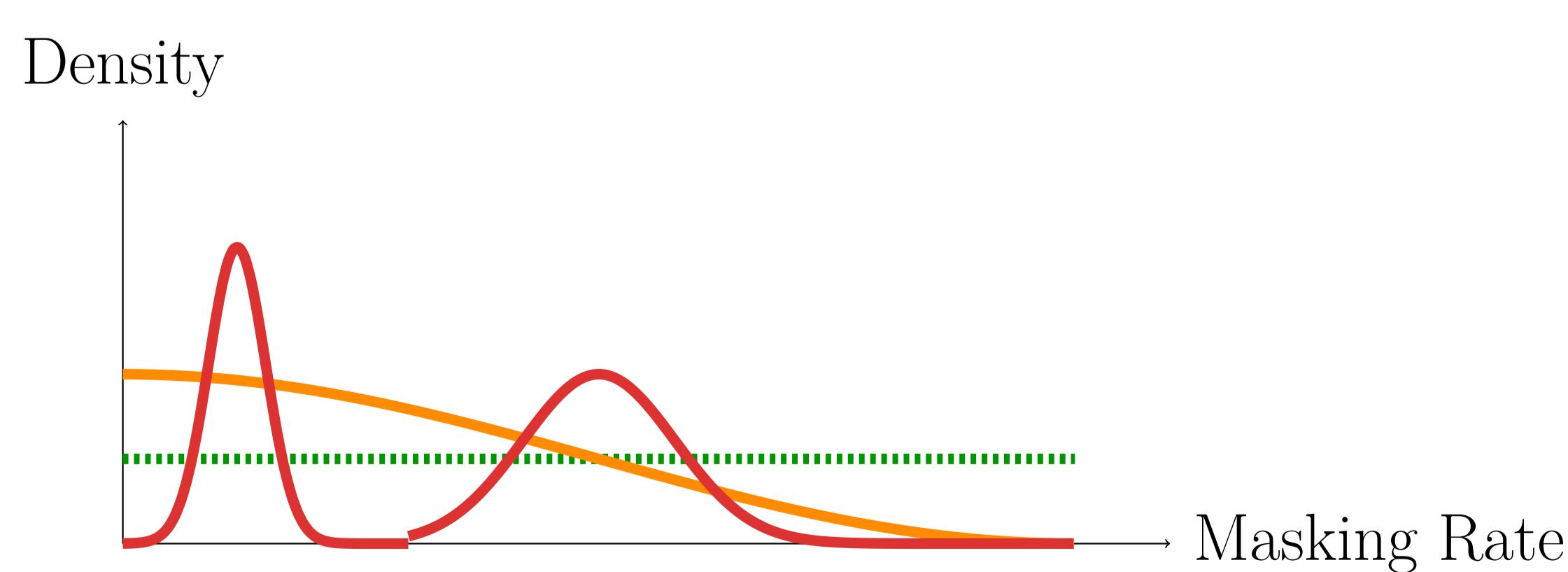
Training Objective:

$$\mathcal{L} = \mathbb{E}_q \int_{t=0}^{t=1} \frac{\alpha'_t}{1 - \alpha_t} \sum_{\ell=1}^L \log \langle x_\theta^\ell(Z_t), x^\ell \rangle dt \quad (1)$$

Bimodal Schedule:

$$p(1 - \alpha_t) = w\mathcal{N}(\mu_1, \sigma_1^2) + (1 - w)\mathcal{N}(\mu_2(\tau), \sigma_2^2) \quad (2)$$

Example values: $\mu_1 = 0.12$, $\mu_2(0) = 0.4$, growing with time



Empirical Finding: Full derivative term α'_t causes performance to degrade

Solution: Use $(\alpha'_t)^\gamma$ with $\gamma < 1.0$

Configuration	EWoK	BLiMP	BLiMP	Sup.
Unimodal ($\gamma = 0.1$)	50.65	64.34	59.32	
Bimodal ($\gamma = 1.0$)	51.10	68.13	63.0	
Bimodal ($\gamma = 0.1$)	52.46	79.49	72.81	

2 Frequency-Informed Masking

Core Idea: Rare tokens are more informative than common words
Masking Weight Formula:

$$w_{new} = \begin{cases} w^p \frac{1 - \alpha_t}{\mu} & \text{if } \mu > 1 - \alpha_t \\ -(1 - w^p) \frac{\alpha_t}{1 - \mu} + 1 & \text{otherwise} \end{cases} \quad (3)$$

Curriculum Learning: Gradually increase p from 0 to 0.02 across training epochs

Configuration	EWoK	BLiMP	BLiMP	Sup.
Cosine	52.44	79.05	70.74	
+ Frequency Masking	52.63	78.92	71.77	

3 Conclusions | Discussion

Our proposed framework matches hybrid baselines with a single objective, achieving competitive performance across diverse tasks.

- **Bimodal schedules** yield the best results by unifying MLM and generation objectives
- **Frequency-informed masking** boosts performance on harder tasks
- **MDLMs are viable** for data-constrained language modeling
- **A more fitting Evaluation Backend** can further improve performance

References

- [1] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models, 2024.
- [2] Lucas Georges Gabriel Charpentier and David Samuel. Gpt or bert: why not both?, 2024.

Acknowledgements

This work has been supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. We acknowledge EuroHPC JU for awarding the project ID EHPC-AI-2024A04-051 access to the EuroHPC supercomputer LEONARDO hosted by CINECA (Italy).