# Accelerating Portuguese Masked Diffusion Models through Representation Alignment

**Adalberto Ferreira Barbosa Junior**[1], **Lucas Lima Neves**[1], **and Adriano César Santana**[2]
[1] Center of Excelence in Artificial Inteligence
[2] School of Electrical, Mechanical and Computer Engineering
jrberto.01@gmail.com, lucas.neves@egresso.ufg.br, adriano@ufg.br

## Abstract

Masked Diffusion Language Models (MDLM) have recently demonstrated that discrete diffusion can achieve competitive performance in text generation. However, training these models remains computationally expensive, particularly for lower-resourced languages like Portuguese. In this work, we adapt REPresentation Alignment (REPA), a technique originally proposed for vision, to the textual domain. We systematically evaluate the impact of aligning the internal representations of a Portuguese MDLM with those of pretrained teacher encoders (e.g., Qwen, BERTimbau). Our experiments show that REPA significantly accelerates training and improves final perplexity by 28.6% compared to a baseline without alignment. We also identify optimal hyperparameters, finding that mid-level alignment with large or domain-specific teacher encoders yields the best results.

## 1 Introduction

Diffusion models have recently emerged as a promising alternative to autoregressive language models, offering parallel decoding and competitive perplexity (Sahoo et al., 2024). In the textual domain, Masked Diffusion Language Models (MDLM) realize discrete diffusion by progressively masking tokens, but they remain costly to train because the model must learn meaningful representations across many timesteps.

Representation alignment aims to speed up diffusion by matching its hidden states to a pretrained encoder. REPresentation Alignment (REPA) adds a lightweight loss that encourages a diffusion model to share the embedding space of an external teacher without changing the architecture or sampler (Yu et al., 2025), yet it has not been explored for text.

In this work, we study REPA for masked diffusion language models in Brazilian Portuguese. We compare teacher encoders, alignment layers, and loss weights, measuring their impact on diffusion ELBO perplexity. To our knowledge, this is the first application of REPA to text diffusion and to Portuguese, and we show that it substantially accelerates training and improves final perplexity.

## 2 Related Work

### 2.1 Diffusion Models for Text Generation

Text diffusion models operate in continuous (Li et al., 2022) or discrete spaces (Austin et al., 2021; Lou et al., 2024). We build on Masked Diffusion Language Models (MDLM) (Sahoo et al., 2024), which formulate diffusion as a weighted average of masked language-modeling losses. This yields a Rao–Blackwellized variational bound that improves stability and perplexity over general discrete diffusion. We adopt this objective as our backbone.

### 2.2 Representation Alignment

Representation Alignment (Yu et al., 2025) was proposed for diffusion transformers in vision as an auxiliary loss that aligns internal hidden states with a pretrained teacher encoder, stabilizing optimization and speeding up convergence without altering inference. Its effectiveness for discrete text diffusion and lower-resource languages remains unexplored.

### 2.3 Portuguese Language Modeling

Portuguese NLP benefits from pretrained encoders such as BERTimbau (Souza et al., 2020), RoBERTaCrawlPT (Garcia et al., 2024), and multilingual Qwen models (Qwen et al., 2025; Yang et al., 2025). We reuse these models as teachers for REPA when training MDLMs for Brazilian Portuguese.

## 3 Methodology

Our approach follows the Masked Diffusion Language Model formulation (Sahoo et al., 2024), in which a clean sequence $x_0$ is gradually corrupted through a masking process. At each timestep $t \in \{1, \ldots, T\}$, tokens are independently replaced



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 1*, pages 968–973
April 13-16, 2026 

by a mask symbol according to a predefined noising schedule. The model is trained to reverse this corruption by predicting $x_0$ from the partially masked sequence $x_t$.

### 3.1 MDLM ELBO Objective

MDLM optimizes the Evidence Lower Bound (ELBO) on the log-likelihood of the data. In practice, we minimize its negative counterpart, the Negative ELBO (N-ELBO), which under the SUBS parameterization of Sahoo et al. (2024) can be written as:

$$\mathcal{L}_{\text{N-ELBO}}^{\infty} = \mathbb{E}_q \left[ \frac{\alpha_t'}{1-\alpha_t} \sum_{\ell=1}^{L} \log \langle x_\theta^\ell(z_t),\ x^\ell \rangle \right] \tag{1}$$

where $\mathbb{E}_q$ is taken over diffusion trajectories $(z_t, x)$, $\alpha_t$ denotes the cumulative masking rate at timestep $t$, and the inner logarithm corresponds to a token-level cross-entropy between the model predictions $x_\theta^\ell(z_t)$ and the ground-truth tokens $x^\ell$. Minimizing this N-ELBO is equivalent to maximizing the original ELBO; therefore, lower N-ELBO values imply lower (better) diffusion perplexity.

### 3.2 REPA Alignment Loss

Training diffusion models from scratch requires the model to learn semantically meaningful representations at every timestep, which makes optimization slow and unstable. To improve convergence, we incorporate REPA as an auxiliary objective. Figure 1 gives a high-level overview of our training setup.

Given the representation $f_{\text{T}}(x_0)$ from a pretrained teacher encoder and an internal hidden state $h_t^l \in \mathbb{R}^{d_{\text{model}}}$ extracted from layer $l$ of the diffusion model at timestep $t$, REPA aligns these two representations by minimizing a loss function.

We evaluate two alignment functions: cosine similarity and mean squared error (MSE). While cosine similarity has been effective in vision-based REPA, we observe higher variance and reduced stability when applied to textual diffusion. In contrast, MSE provides smoother gradients, faster convergence, and more consistent improvements in reconstruction loss. Therefore, we adopt the MSE formulation for alignment.

Because the teacher representation generally lies in a different vector space, the diffusion hidden state is first passed through a learned linear projection $W \in \mathbb{R}^{d_{\text{T}} \times d_{\text{model}}}$ to match the dimensionality of the teacher embedding. For each sample $i$ in the batch, the projected diffusion vector is $Wh_t^{l,(i)}$,

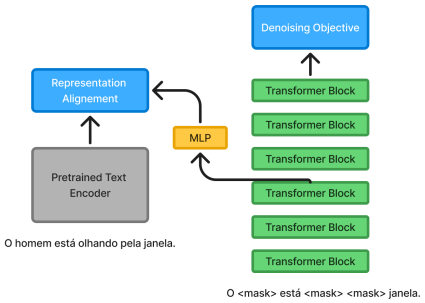


Figure 1: Overview of our REPA-enhanced MDLM. A pretrained text encoder provides semantic representations that are aligned, via a learned projection, to the hidden states of a masked diffusion language model trained with the diffusion ELBO objective.

while $f_{\text{T}}(x_0^{(i)})$ denotes the pooled representation obtained by feeding the clean input sequence $x_0^{(i)}$ into the external encoder.

The resulting REPA loss is:

$$\mathcal{L}_{\text{REPA}} = \mathbb{E}_i \left\| Wh_t^{l,(i)} - f_{\text{T}}(x_0^{(i)}) \right\|_2^2. \tag{2}$$

This loss is applied only to selected transformer layers, typically at intermediate depth, where semantic representations are most stable and alignment has shown the strongest effect.

### 3.3 Final Training Objective

The complete training objective combines the MDLM ELBO and the REPA alignment term:

$$\mathcal{L} = \mathcal{L}_{\text{N-ELBO}}^{\infty} + \lambda\, \mathcal{L}_{\text{REPA}}, \tag{3}$$

where $\lambda$ controls the influence of the alignment loss. Importantly, REPA does not modify the sampling algorithm nor the architecture; it only provides additional semantic guidance during training.

### 3.4 Forward Noising Process

Following MDLM, the corruption process factorizes across tokens. At timestep $t$, each token in $x_0$ is independently replaced by the mask symbol with probability $\alpha_t$, producing $x_t$. Thus, the forward diffusion distribution is:

$$q(x_t^\ell = [M] \mid x_0^\ell) = \alpha_t,\ q(x_t^\ell = x_0^\ell \mid x_0^\ell) = 1-\alpha_t.$$

### 3.5 Reverse Denoising Model

As in MDLM, the denoising network predicts the clean sequence $x_0$ directly from $(x_t)$. The model outputs a distribution over the vocabulary for each token position, $p_\theta(x_0^\ell \mid x_t)$, enabling a single shared transformer to perform denoising at all diffusion steps.

## 4 Experiments

We conduct a systematic empirical study to evaluate the impact of REPA on masked diffusion language models trained in Portuguese. Our setup follows the original MDLM formulation, extending it with an auxiliary REPA loss applied at specific transformer layers.

### 4.1 Experimental Setup

All models are trained on the `por_Latn` split of the FineWeb-2 corpus (Penedo et al., 2025). This dataset features a diverse domain composition, including official news portals, general pop culture blogs (spanning gaming, beauty, cooking, culture, etc.), social media platforms, advertisements, charitable campaigns, and profiles of public figures or corporations. The data is divided into training, validation, and test splits; we use a subset of the corpus, with our training split comprising 655,360,000 tokens. Given the established quality of the FineWeb-2 corpus, no further pre-processing steps were applied. As the base diffusion model, we use the Qwen2.5-0.5B architecture with causal attention replaced by bidirectional attention. Models are trained for 5,000 steps with a sequence length of 256, using AdamW (8-bit), BF16 precision, and TF32 acceleration. The masking probability is annealed linearly from 0.01 to 1.0 during training.

### 4.2 Grid Search Design

We perform a full grid search over three dimensions:

- **Teacher:** `Qwen3-0.6B`, `Qwen3-1.7B` (Yang et al., 2025), `Qwen2.5-1.5B`, `Qwen2.5-0.5B` (Qwen et al., 2025), `Neuralmind-BERT` (Souza et al., 2020), `RoBERTaCrawlPT` (Garcia et al., 2024).
- **Layer:** $\{3, 7, 11, 17, 23\}$.
- **Weight:** $\{0.25, 0.50, 0.75, 1.00\}$.

This results in 120 trained REPA-enhanced models, plus a baseline MDLM model without REPA.

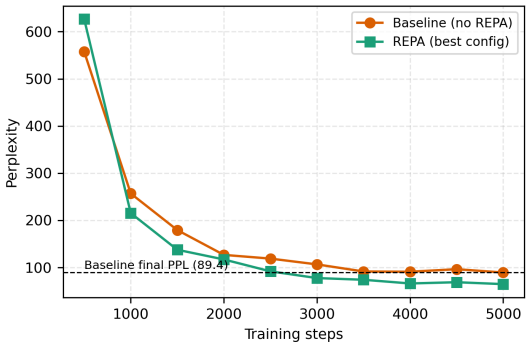


Figure 2: Training curves for the baseline MDLM and the best REPA configuration (Qwen3-1.7B teacher, layer 17, $\lambda = 1.0$). The REPA model matches the baseline's final perplexity in roughly half the number of training steps, evidencing faster convergence.

### 4.3 Evaluation Metric

We evaluate all models using the variational perplexity derived from the Evidence Lower Bound on a held-out test set. It is important to note that, as MDLMs optimize a variational bound, all reported perplexities are upper bounds on the true perplexity (lower is better).

## 5 Results

### 5.1 Performance and Acceleration

Across the full grid, the baseline MDLM without REPA reaches a diffusion perplexity of **86.42**, while our best REPA configuration attains **61.76**, corresponding to a reduction of **28.6%**. Figure 2 shows the training dynamics for the baseline and the best REPA model. The REPA-enhanced model rapidly closes the gap to the baseline and reaches the baseline's final perplexity level after roughly 2,600 steps, indicating that REPA achieves competitive performance with about half the training budget.

### 5.2 Impact of Hyperparameters

We next analyze how teacher choice, alignment weight, and alignment layer affect performance.

**Teacher selection.** Figure 3 compares the mean perplexity of different teacher models. Qwen3-1.7B yields the best results, followed closely by BERTimbau, which surprisingly outperforms larger modern models like Qwen2.5-1.5B. This suggests that domain specificity (as in BERTimbau's Portuguese pretraining) can be as important as raw model capacity for effective alignment. Conversely,

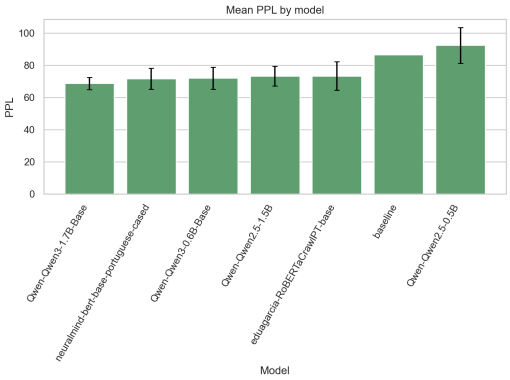


Figure 3: Mean diffusion perplexity by teacher model, averaged across layers and weights. Strong teachers (Qwen3-1.7B) and domain-specific encoders (BERTimbau) provide the best alignment targets.

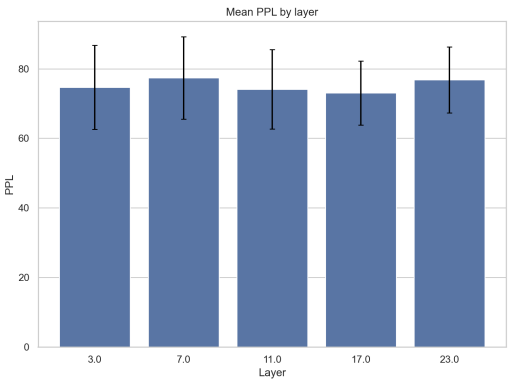


Figure 4: Mean diffusion perplexity by alignment layer, aggregated across all weights and teachers. Mid-level layers perform best on average.

weak teachers such as Qwen2.5-0.5B degrade performance relative to the baseline.

**Layer and weight.** Figures 4 and 5 show the mean perplexity aggregated by alignment layer and weight, respectively. Mid-level layers, especially layer 17, yield the lowest perplexity on average, confirming that they provide the most useful semantic representations for REPA. Smaller alignment weights ($\lambda \in [0.25, 0.5]$) offer the best trade-off between stability and performance across all teachers; however, the optimal weight is teacher-dependent: while our best configuration uses $\lambda = 1.0$ with Qwen3-1.7B, other strong teachers like BERTimbau peak at $\lambda = 0.5$. Thus, while low weights are robust on average, tuning $\lambda$ per teacher is necessary to unlock the absolute best performance.

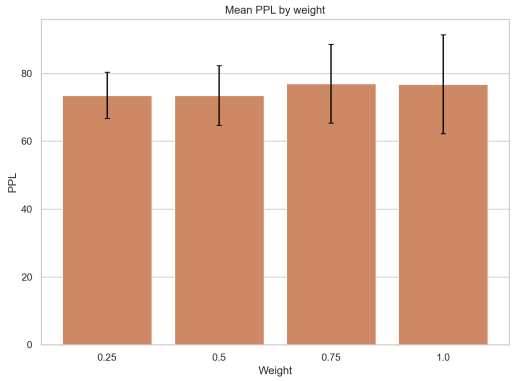


Figure 5: Mean diffusion perplexity by REPA weight, aggregated across all layers and teachers. Smaller weights are more stable overall.

### 5.3 Efficiency and Cost-Benefit Analysis

While REPA significantly accelerates convergence in terms of training steps, it introduces computational overhead per step due to the additional forward pass through the teacher model and the projection layer. This trade-off is crucial for evaluating the method's practicality, particularly concerning energy consumption and training costs.

To quantify this, we define an *Efficiency Score* as the ratio between the percentage reduction in perplexity (PPL Red.) and the percentage increase in training time relative to the baseline (Time Inc.):

$$\text{Efficiency Score} = \frac{\text{PPL Reduction \%}}{\text{Training Time Increase \%}} \tag{4}$$

A score greater than 1.0 indicates that the performance gain outpaces the added computational cost.

It is important to note that the perplexity values used in this efficiency analysis are *averages* across all experimental configurations (various layers and weights) for each teacher, rather than the optimal configuration reported in earlier sections. This provides a more robust estimate of the expected performance gain versus cost for a given teacher architecture.

Table 1 summarizes the cost-benefit analysis for different teacher models. We observe two distinct regimes:

1. **High Efficiency Encoders:** Encoder-only models like `RoBERTaCrawlPT` (Garcia et al., 2024) and `BERTimbau` (Souza et al., 2020) achieve the highest efficiency scores ($> 1.0$). They provide substantial perplexity reductions ($15 - 17\%$) with only a minor increase in training time ($7 - 15\%$). Since training time

on a fixed hardware setup correlates linearly with energy consumption and CO2 emissions, these models represent the most environmentally sustainable approach for alignment.

2. **High Performance, High Cost LLMs:** Large generative teachers like `Qwen3-1.7B` yield the best absolute performance (24.6% average reduction) but come with a significant computational cost, more than doubling the training time (+109.8%). Their efficiency score is consequently lower ($\approx 0.22$), reflecting a less favorable cost-benefit trade-off. While they maximize quality, they are less resource-efficient than smaller encoders.

Importantly, using a weak teacher (`Qwen2.5-0.5B`) is detrimental, resulting in worse perplexity while still incurring a 37% time penalty. This confirms that teacher selection is critical not only for performance but also for computational efficiency.

| Teacher Model | PPL Red. (%) | Time Inc. (%) | Score |
|---|---|---|---|
| `RoBERTaCrawlPT-Base` | 15.2 | 7.3 | **2.08** |
| `BERTimbau-Base` | 17.2 | 15.4 | 1.12 |
| `Qwen3-0.6B` | 16.8 | 83.7 | 0.20 |
| `Qwen2.5-1.5B` | 15.2 | 82.1 | 0.19 |
| `Qwen3-1.7B` | **24.6** | 109.8 | 0.22 |
| `Qwen2.5-0.5B` | -6.9 | 37.4 | -0.18 |

Table 1: Efficiency analysis of REPA with different teacher models. **PPL Red.** denotes the percentage reduction in mean perplexity (averaged across all runs for that teacher) compared to the baseline. **Time Inc.** denotes the percentage increase in training time relative to the baseline. **Score** is the ratio PPL Red. / Time Inc.

## 6 Conclusion

We performed the first large-scale study of REPA for masked diffusion language models in Portuguese, evaluating over 120 trained configurations. Our experiments show that REPA is a highly effective auxiliary signal that improves perplexity by up to 28.6% compared to the baseline.

Three key findings emerge: (1) **Mid-level layers** (especially layer 17) provide the best alignment targets. (2) **Low to moderate REPA weights** (0.25–0.50) yield the best average performance, though the optimal weight varies by teacher (e.g., $\lambda = 1.0$ for Qwen3-1.7B). (3) **Teacher quality and domain match** are crucial: Qwen3-1.7B achieves the best results, but the domain-specific BERTimbau also performs competitively.

These results confirm that REPA is an effective and lightweight mechanism for accelerating and stabilizing diffusion-based language models. Future work should explore dynamic weighting schedules, multilingual REPA, and autoregressive–diffusion hybrid models.

## 7 Limitations

Our study focuses exclusively on Brazilian Portuguese, and the findings may not fully generalize to other languages with different morphological characteristics. We also fix a relatively short sequence length of 256 tokens in order to make the 120-model grid search computationally feasible; we expect the observed benefits of REPA to transfer to longer contexts, which we leave for future work. Additionally, we explored a fixed set of teacher models and a specific range of hyperparameters; other architectures or wider grid searches could yield different optimal configurations. Finally, while REPA improves convergence, the computational cost of the additional projection layer and teacher model inference during training remains a practical concern. Our efficiency analysis reveals that while larger teachers yield better perplexity, they can double the training time per step, potentially offsetting the benefits in strictly resource-constrained environments.

## Acknowledgments

This work has been partially funded by the project Research and Development of Computational Techniques for Security and Privacy of Second-Generation Multimodal Data, supported by the Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI, grant number 057/2023, signed with EMBRAPII.

## References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Eduardo A. S. Garcia, Nadia F. F. Silva, Felipe Siqueira, Hidelberg O. Albuquerque, Juliana R. S. Gomes, Ellen Souza, and Eliomar A. Lima. 2024. RoBERTaLexPT: A legal RoBERTa model pretrained with

deduplication for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 374–383, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2025. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*.