CS 6120: Comparing Text Diffusion with Autoregressive Generation using Transformers

Dhruv Puri, Jatan Patel, Sam Selvaraj

puri.dh@northeastern.edu, patel.jatan@northeastern.edu, selvaraj.sam@northeastern.edu

Abstract

We compare autoregressive transformers with masked diffusion models for English-to-French machine translation. Using identical 44M-parameter architectures trained on 127K OPUS Books sentence pairs, we evaluate translation quality and training efficiency. Autoregressive models achieve substantially higher BLEU scores (30.42 vs. 13.77) and faster convergence, though this advantage may reflect scale limitations rather than fundamental algorithmic superiority. Our findings suggest diffusion models require larger scale and extended training to realize their potential benefits in bidirectional reasoning and iterative refinement.

1. Introduction

Autoregressive transformers dominate text generation through sequential left-to-right token prediction, achieving strong performance but suffering from inference latency and exposure bias. Diffusion models offer an alternative through parallel iterative refinement with bidirectional context. We systematically compare these paradigms for neural machine translation using matched 44M-parameter architectures on English-to-French translation. Our autoregressive baseline substantially outperforms masked diffusion (BLEU 30.42 vs. 13.77), though this gap likely reflects training efficiency advantages at small scale rather than fundamental limitations of diffusion approaches.

2. Background/Related Work

Autoregressive Generation. Autoregressive models decompose sequence probability as $P(x_1,...,x_n) = \prod_i P(x_i \mid x_1,...,x_{i-1})$. The sequential process requires O(n) forward passes, prohibiting parallelization.

Diffusion Models. Ho et al. (2020) introduced diffusion models that gradually corrupt data with noise and learn reversal. Chen et al. (2023) proposed Analog Bits with self-conditioning for discrete data. Li et al. (2022) developed Diffusion-LM embedding tokens in continuous space for controllable generation. Gong et al. (2023) proposed DiffuSeq achieving comparable performance to autoregressive baselines on translation and summarization with higher diversity.

Parallel Generation. Chang et al. (2022) introduced MaskGIT using bidirectional transformers, achieving 64x speedup. Sahoo et al. (2024) showed masked diffusion can approach autoregressive perplexity. Nie et al. (2025) introduced LLaDA, scaling masked diffusion to 8B parameters with competitive performance versus LLaMA3, notably solving the reversal curse and surpassing GPT-4o in bidirectional reasoning.

3. Data

We use the OPUS Books English-French parallel corpus with 127,000 sentence pairs (95% train and 5% test). We developed a custom BPE tokenizer with 20,000 vocabulary trained jointly on both languages for consistent subword segmentation.

4. Methods

Architecture. Both models use identical 8-layer transformers (43.99M parameters): token embeddings (vocab=20K, d_model=512), learned positional embeddings (max_len=266), 8 decoder layers with multi-head attention (n_heads=8, d_ff=4096), RMSNorm, SwiGLU activation, dropout=0.1, and output projection to vocabulary. Key hyperparameters: lr=1e-4, batch_size=16, accumulation_steps=4, weight_decay=0.01, warmup=1000 steps.

CS 6120: Comparing Text Diffusion with Autoregressive Generation using Transformers

Dhruv Puri, Jatan Patel, Sam Selvaraj
puri.dh@northeastern.edu, patel.jatan@northeastern.edu, selvaraj.sam@northeastern.edu

Autoregressive Model. Uses causal masking preventing future token attention. Training minimizes cross-entropy on next-token prediction. Generation proceeds left-to-right with cost O(N).

Masked Diffusion Model. Follows masked discrete diffusion with bidirectional attention. Forward process randomly masks tokens via linear schedule (0% at t=0 to 100% at t=T). Training predicts original tokens at masked positions with cross-entropy loss. Inference iterates from t=T to t=0: model predicts all tokens simultaneously, retains high-confidence predictions, re-masks low-confidence tokens for refinement. Generation cost O(N × S / B) where S=50 steps, B=block size.

Training. Both trained 50 epochs on RTX 4000 Ada (20GB) using SFT mode (loss computed only on French translations).

5. Results

Translation Quality. Autoregressive substantially outperforms diffusion: BLEU 30.42 vs. 13.77, ROUGE-1 0.66 vs. 0.61, ROUGE-L 0.65 vs. 0.61, BERTScore 0.89 vs. 0.80, indicating better n-gram overlap and semantic similarity.

Training Dynamics. Autoregressive converges faster to lower loss (~1.5 vs. ~2.5) and perplexity (~50 vs. ~200-300), showing stronger predictive confidence. Both exhibit stable training without overfitting.

Qualitative Analysis. For simple sentences, both produce acceptable translations, though diffusion introduces artifacts (extra punctuation, repetition). For complex sentences, autoregressive generates coherent translations while diffusion produces corrupted output with semantic errors and poor long-range dependencies.

Analysis. Diffusion underperformance stems from: (1) training efficiency gap—diffusion requires more steps to converge, (2) scale limitations—emergent capabilities appear only at 7B+ parameters, (3) architectural constraints—lacking RoPE/GQA used in state-of-the-art diffusion models.

6. Conclusions

Autoregressive models substantially outperform masked diffusion at small scale (44M parameters, 50 epochs), achieving 2.2× higher BLEU with faster convergence. However, results reflect training efficiency advantages rather than fundamental limitations. Future work should explore: (1) scaling to 1B+ parameters where diffusion shows emergent capabilities, (2) extended training (5-10× more steps), (3) advanced architectures (RoPE, GQA), (4) tasks benefiting from bidirectional context (infilling, constrained generation), and (5) hybrid approaches. While autoregressive remains practical for translation currently, diffusion offers promise for applications requiring global coherence or controllable generation at larger scales.

References

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *NeurIPS*.

Li, X., Thickstun, J., Gulrajani, I., Liang, P., & Hashimoto, T. (2022). Diffusion-LM Improves Controllable Text Generation. *ICLR*.

Gong, S., Li, M., Feng, J., Wu, Z., & Kong, L. (2023). DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. *ICLR*.

Chen, T., Zhang, R., & Hinton, G. (2023). Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. *ICLR*.

CS 6120: Comparing Text Diffusion with Autoregressive Generation using Transformers

Dhruv Puri, Jatan Patel, Sam Selvaraj

puri.dh@northeastern.edu, patel.jatan@northeastern.edu, selvaraj.sam@northeastern.edu

Chang, H., Zhang, H., Jiang, L., Liu, C., & Freeman, W. T. (2022). MaskGIT: Masked Generative Image Transformer. *CVPR*.

Sahoo, S. S., Doucet, A., & Grathwohl, W. (2024). Simple and Effective Masked Diffusion Language Models. *NeurIPS*.

Nie, S., et al. (2025). Large Language Diffusion Models. *arXiv:2502.09992*.

TABLES/FIGURES:

Table 1: Translation Metrics (Epoch 40)

Metric	Diffusion	ARM
BLEU	13.77	30.42
ROUGE-1	0.6134	0.6619
ROUGE-L	0.6068	0.6549
BERTScore	0.8013	0.8905

Table 2: Example Translations

Input	Google translate	Diffusion translation	Transformer translation
The sun is shining.	Le soleil brille.	Il soleil un beau soleil »	Le soleil est brillant .
He opened the door.	Il ouvrit la porte.	Il ouvrit la porte :	Il ouvrit la porte .
She looked at him.	Elle le regarda.	Elle la ' d ' regardait :	Elle le regarda .
I am happy.	Je suis heureux.	J suis heureux ….	Je suis heureux .

Figure 1: Training/validation loss curves

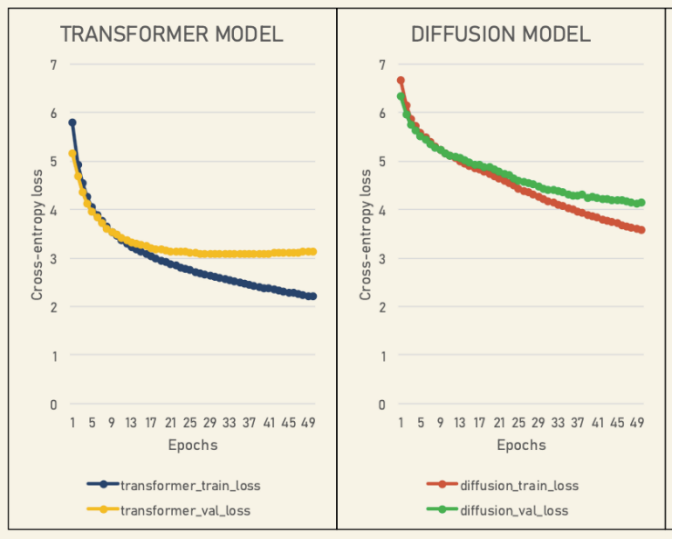

Figure 2: Perplexity comparison