Scaffold-and-Fill Diffusion (SF-Diff): A Hybrid Architecture for Accelerated Language Model Inference

Hilal Limo
Independent Researcher (Self-Taught, Age 15)
hilallimotl@gmail.com

June 13, 2025

Abstract

Autoregressive transformer models, the dominant architecture for modern Large Language Models (LLMs), are fundamentally constrained by high inference latency due to their sequential generation process. In this paper, I propose Scaffold-and-Fill Diffusion (SF-Diff), a novel hybrid architecture designed to significantly accelerate text generation by deconstructing the task into two parallelizable stages. The core hypothesis is that natural language can be separated into a semantic "scaffolding" of keywords and a grammatical "filler" of structural words. SF-Diff first utilizes a non-autoregressive diffusion model to generate the complete semantic scaffold, a sequence of keyword vector embeddings, in a fixed number of highly parallelizable steps. Subsequently, a lightweight autoregressive transformer decoder performs a "grammatical infilling" task, weaving the structural words around the pre-generated semantic core. This approach aims to combine the holistic, parallel generation strengths of diffusion models with the grammatical precision of transformers, offering a substantial reduction in inference latency while maintaining high-quality, coherent output.

1 Introduction

The generative capabilities of Large Language Models (LLMs) based on the Transformer architecture [8] have established a new paradigm in artificial intelligence. Models such as the GPT series [5] and Llama [7] have demonstrated an impressive ability to generate fluent and coherent text. However, this fluency is achieved via an autoregressive process, where each token is generated sequentially based on all previously generated tokens. This creates a direct, linear relationship between output length and inference time, $O(n)$, which is a significant barrier for real-time, interactive applications.

While techniques like speculative decoding [3] aim to mitigate this, they remain within the autoregressive framework. In this paper, I introduce a theoretical framework for a new architecture, Scaffold-and-Fill Diffusion (SF-Diff), that reimagines the generation process itself.

My approach is based on the linguistic observation that sentences are composed of two distinct components:

1. **Semantic Scaffolding:** A core set of keywords (nouns, verbs, adjectives, adverbs) that carry the primary meaning.
2. **Grammatical Filler:** A set of functional words (determiners, prepositions, conjunctions) that provide grammatical structure.

SF-Diff leverages this separation by using two specialized models: a non-autoregressive diffusion model to generate the entire semantic scaffold in parallel, and a fast autoregressive decoder to weave the grammatical filler around it.

2 Related Work and Architectural Context

My work builds upon several key areas of modern machine learning research.

- **Autoregressive Models:** The "filler" stage of my proposed architecture is a standard Transformer decoder, leveraging its proven ability to handle local grammatical dependencies with high fidelity.

- **Diffusion Models & Hybrid Architectures:** Denoising Diffusion Probabilistic Models (DDPMs) [2] have set a high bar for high-fidelity generation. While they are best known in image synthesis [6], adapting diffusion to discrete data like text is an active, evolving field [4]. Recent work highlights the power of hybrid models combining Transformers with diffusion mechanisms. For instance, Google DeepMind's Gemini Diffusion integrates a diffusion process with a Transformer backbone to deliver state-of-the-art performance in text and code synthesis, enabling fast, block-wise generation and iterative refinement, quite distinct from diffusion designs intended for image output. This model exemplifies the principle of using deep semantic conditioning via a Transformer before applying a diffusion step. My SF-Diff proposal draws on this hybrid strategy but inverts the workflow: using diffusion for intermediate semantic structure scaffolding and relying on the Transformer for the final grammatical decoding in accelerated text-to-text generation.

- **Non-Autoregressive Transformers (NAT):** Early attempts to parallelize text generation, such as NAT [1], tried to generate all tokens simultaneously. These models were fast but often produced incoherent text due to the "multimodality problem." SF-Diff mitigates this by only generating the high-level semantic core non-autoregressively, leaving the fine-grained grammatical details to a more suitable autoregressive model.

3 Proposed Architecture: SF-Diff

The SF-Diff architecture is a multi-stage pipeline designed for fast inference.

3.1 Data Preparation: Deconstructing Language

A crucial prerequisite is a method to separate text into keywords and filler. I propose using a standard Part-of-Speech (POS) tagger. For a sentence S, I generate a sequence of keyword tokens $W_k = \{k_1, k_2, ..., k_n\}$ and a corresponding structural pattern $P = \{p_1, p_2, ..., p_m\}$, where each p_i is either a placeholder token `<KEYWORD>` or a functional "filler" word.

3.2 Stage 1: The Diffusion Scaffolder

This non-autoregressive model is responsible for generating the semantic core of the response. It is a conditional denoising diffusion model trained on sequences of keyword vector embeddings. The forward process q adds Gaussian noise over T timesteps with a variance schedule β_t. The model ϵ_θ is a neural network trained to predict this noise from a noisy input E_k^t at timestep t, conditioned on the user's prompt C_p and the structural pattern P.

The training objective is to minimize the L2 loss:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, E_k^0, \epsilon}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}E_k^0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, C_p, P)||^2\right] \tag{1}$$

where $\bar{\alpha}_t$ is the cumulative product of $(1-\beta_t)$. At inference, this model starts with pure noise and iteratively denoises it in a fixed number of parallel steps to produce the final semantic scaffold E_k^0.

3.3 Stage 2: The Transformer Filler

This lightweight autoregressive model renders the semantic scaffold into fluent text. It is a decoder-only Transformer modified with cross-attention layers.

- **Self-Attention:** Attends to the previously generated filler words to maintain grammatical fluency.
- **Cross-Attention:** Attends to the entire sequence of keyword embeddings E_k^0 from the diffusion model, allowing every filler word to "see" the full semantic plan.

The model is trained to maximize the log-probability of the correct filler words W_f given the ground-truth keywords:

$$\mathcal{L}_{\text{fill}} = -\sum \log p(w_i | w_{<i}, E_k^0, P) \tag{2}$$

4 Potential Advantages and Discussion

1. **Inference Speed:** The primary advantage. The semantically heavy part of generation is done in a fixed number of parallel steps by the diffusion model. This could decouple inference time from output length for the most complex part of the task.

2. **Controllability:** The explicit structural pattern P serves as a powerful control vector. By providing different patterns, a user could guide the model to generate text with a specific style, meter, or complexity.

3. **Hybrid Strengths:** The architecture leverages the best of both worlds: the holistic, parallel nature of diffusion for semantic planning and the high-fidelity, sequential nature of transformers for grammatical precision.

5 Challenges and Future Work

- **Structural Pattern Generation:** The generation of a coherent structural pattern P at inference time is a non-trivial challenge. This may require a third, highly optimized model, or the diffusion model may need to be conditioned on the prompt alone.
- **Error Propagation:** The Transformer Filler must be robust to imperfections in the semantic scaffold generated by the diffusion model. If the keyword embeddings are semantically incoherent, the decoder may struggle to produce a fluent output.
- **Training Pipeline Complexity:** This is a multi-stage system requiring a sophisticated training pipeline to ensure the two models learn to cooperate effectively.

6 Conclusion

SF-Diff presents a theoretical framework for a hybrid text generation architecture that diverges from the purely autoregressive paradigm. By separating semantic generation from grammatical infilling, it has the potential to drastically reduce inference latency. While significant engineering challenges remain, the "Scaffold-and-Fill" approach I have outlined offers a promising new direction for building faster, more controllable language models.

References

[1] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[3] Yossi Leviathan, Matan Kalman, Yair Gray, Maor Geva, Shlomi Chronic, Alon Alon, Alon Jacovi, Gadi Shmila, Amir Globerson, Eran Yahav, et al. Fast inference from transformers via speculative decoding. *arXiv preprint arXiv:2211.17192*, 2022.

[4] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm: Improving controllable text generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Technical Report*, 2018.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[7] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.