Speculative Diffusion Decoding: Accelerating Language Generation through Diffusion

February 10, 2025

University of Virginia **Lawrence Livermore National Laboratory**

Jacob K Christopher, Brian R Bartoldson, Tal Ben-Nun, Michael Cardei, Bhavya Kailkhura, Ferdinando Fioretto

Summary

Speculative Diffusion Decoding (SpecDiff) accelerates Large Language Model inference by replacing the autoregressive drafter with a parallel discrete diffusion model, achieving up to 7.2x speedup over standard autoregressive decoding and 1.75x over existing speculative decoding while preserving output quality. This approach reduces computational overhead and enables the use of longer draft lengths.

Table of Contents

1. Introduction

Large Language Models (LLMs) have achieved unprecedented capabilities in natural language processing, but their widespread deployment faces a significant bottleneck: the immense computational cost of inference. Current techniques like quantization and pruning offer speedups but often sacrifice output quality. Speculative decoding has emerged as a promising approach that maintains quality while achieving 2-3x speedups by using a smaller "draft" model to generate candidate tokens that are then verified by the target LLM.

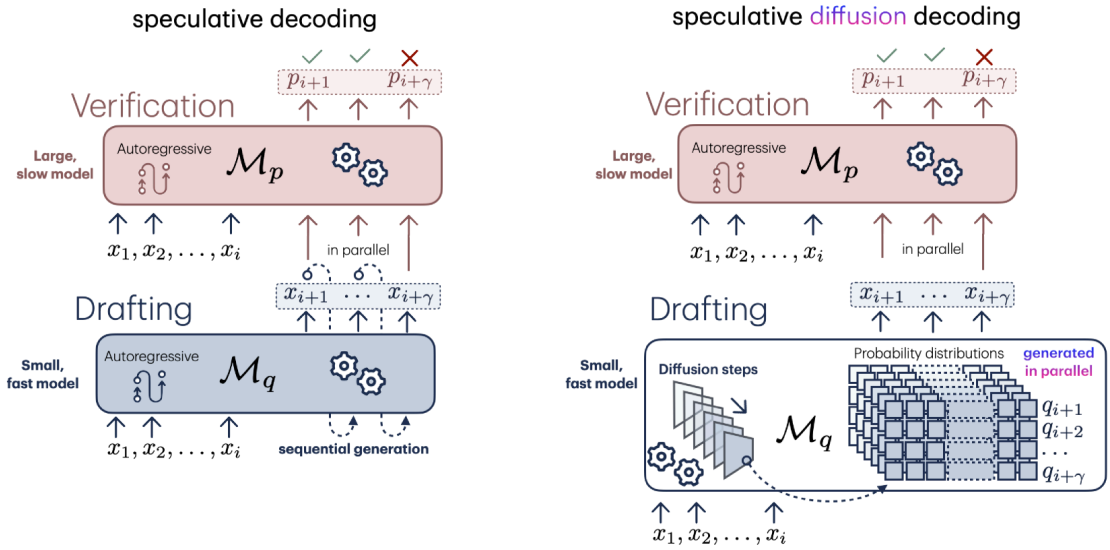

Figure: Architectural comparison between traditional speculative decoding (left) and the proposed speculative diffusion decoding (right). The key difference is replacing the sequential autoregressive drafter with a parallel discrete diffusion model that generates entire token sequences simultaneously.

However, existing speculative decoding approaches are fundamentally limited by their reliance on autoregressive draft models, which generate tokens sequentially. This creates a bottleneck where only the verification step can be parallelized, not the drafting itself. This paper introduces **Speculative Diffusion Decoding (SpecDiff)**, which replaces the autoregressive drafter with a discrete diffusion model capable of generating entire token sequences in parallel, enabling parallelization in both drafting and verification stages.

2. Core Methodology

The fundamental innovation of SpecDiff lies in leveraging discrete diffusion models as the draft generator within the speculative decoding framework. Unlike autoregressive models that generate tokens one by one, diffusion models can produce complete sequences simultaneously through a parallel denoising process.

2.1. Discrete Diffusion Models for Language

SpecDiff employs Masked Diffusion Language Models (MDLM), which adapt the diffusion paradigm for discrete text data. The forward process gradually masks tokens in a sequence, while the reverse process learns to reconstruct the original text. The model optimizes a continuous-time Negative Evidence Lower BOund (NELBO) objective:

$$L = \mathrm{E}_{t,x_0}\left[\| f_\theta(x_t, t) - x_0 \|^2\right]$$

where f_θ is the denoising network, x_t represents the noisy state at time t, and x_0 is the clean text sequence.

2.2. SpecDiff Algorithm

The SpecDiff process modifies traditional speculative decoding as follows:

1. **Parallel Draft Generation**: The diffusion model M_q generates a candidate sequence of γ tokens through T denoising steps, producing probability distributions for each position simultaneously.
2. **Target Model Verification**: The target LLM M_p evaluates the entire drafted sequence in parallel, computing its own probability distributions.
3. **Token Acceptance**: For each drafted token x_j, it is accepted if the draft probability $q(x_j) \leq$ target probability $p(x_j)$. Otherwise, it's rejected with probability $1 - \frac{p(x_j)}{q(x_j)}$.
4. **Sequential Processing**: Accepted tokens are returned along with one additional token sampled from the adjusted target distribution.

2.3. Key Advantages

The parallel nature of diffusion models enables SpecDiff to:

- Scale draft length γ to much larger values (15-20 tokens) with minimal overhead
- Reduce sensitivity to γ while becoming more sensitive to the number of diffusion steps T
- Achieve computational efficiency through fewer required model evaluations

3. Experimental Results

The evaluation demonstrates SpecDiff's superior performance across multiple metrics and model configurations:

3.1. Performance Gains

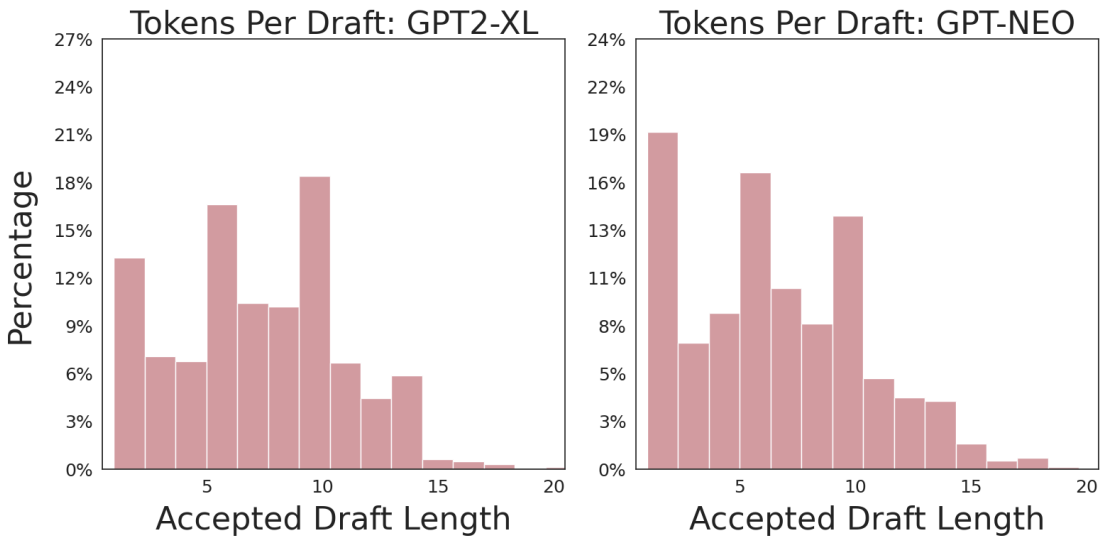

Figure: Distribution of accepted draft lengths for GPT2-XL and GPT-NEO, showing SpecDiff's ability to achieve longer accepted sequences compared to traditional approaches.

- **Speedup over vanilla decoding**: Up to 7.2x acceleration
- **Speedup over standard speculative decoding**: Up to 1.75x improvement
- **Computational efficiency**: Over 33% fewer FLOPs per draft compared to methods like EAGLE and EAGLE-2

3.2. Model-Specific Results

For the Vicuna 33B model with fine-tuned MDLM:

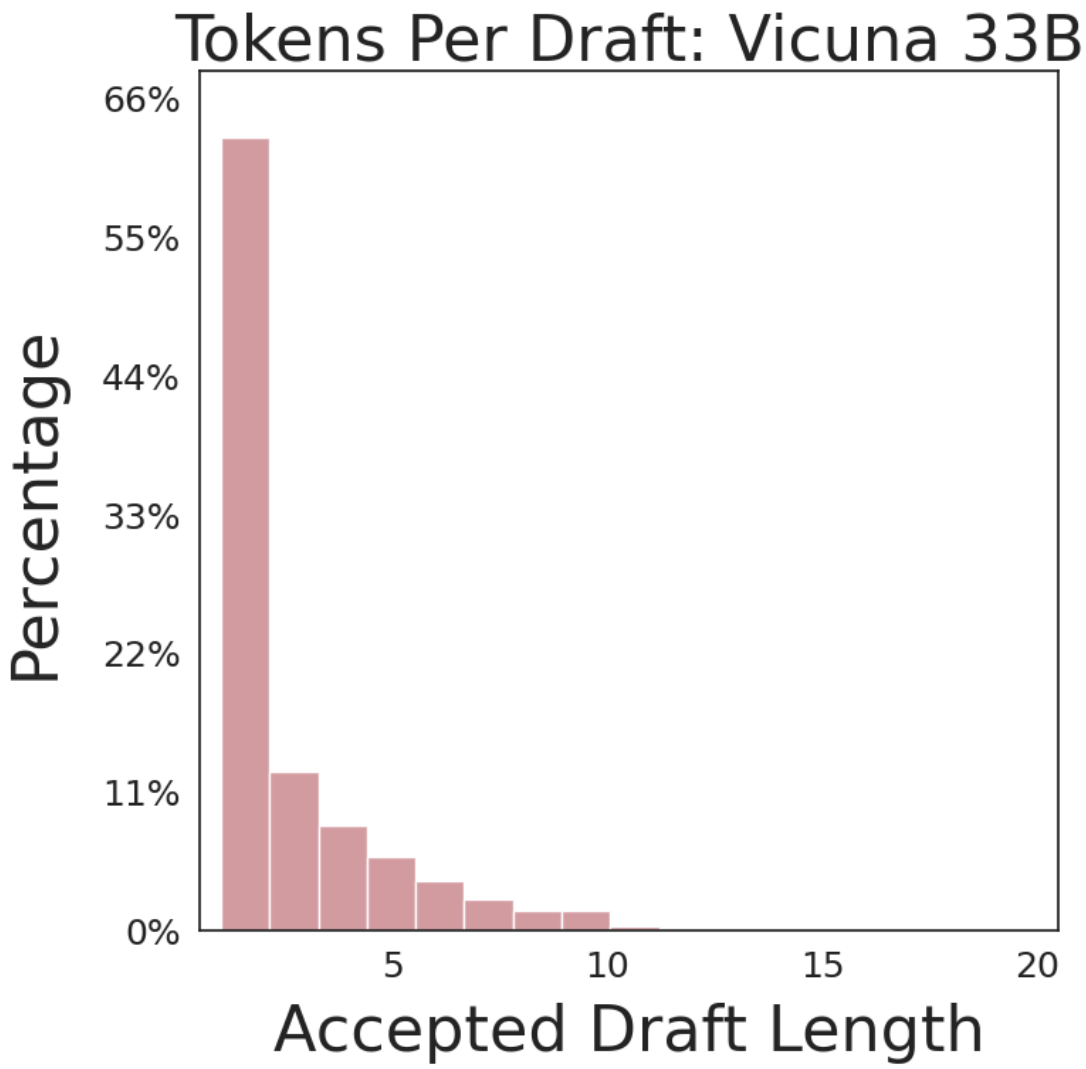

Figure: Vicuna 33B shows exceptional performance with over 60% of cases accepting only 1 token, indicating high alignment between the fine-tuned diffusion drafter and target model.

3.3. Hyperparameter Analysis

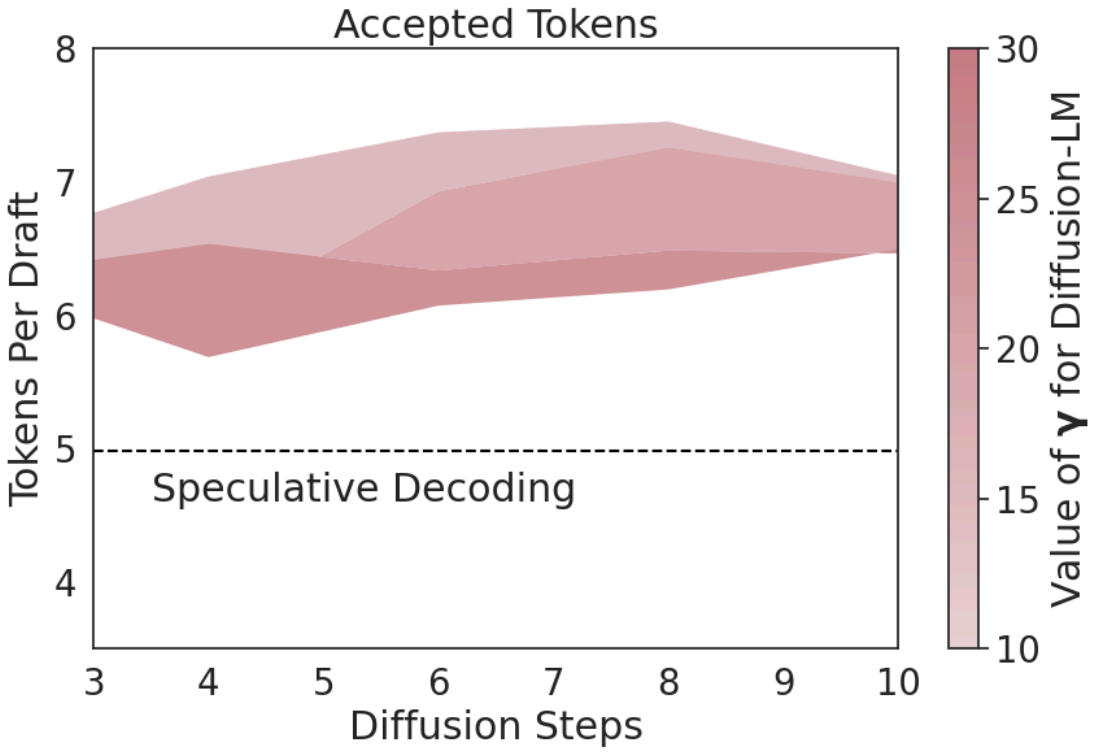

Figure: The relationship between number of diffusion steps and accepted tokens, showing optimal performance around 4-6 steps for the OpenWebText dataset.

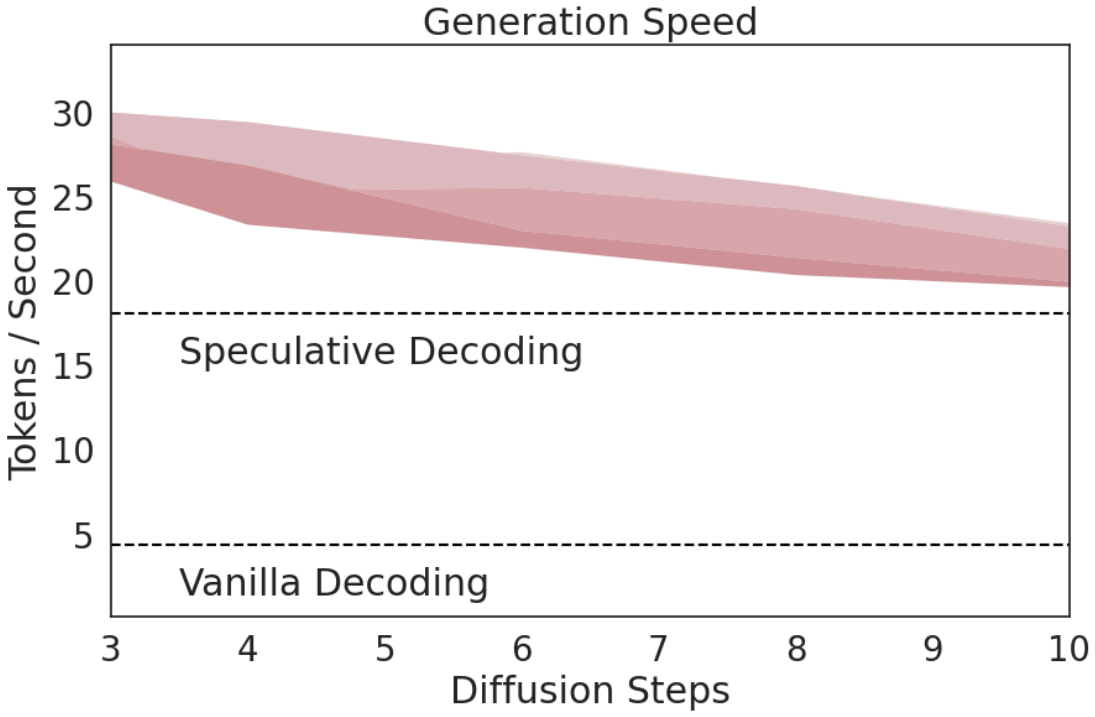

Figure: Generation speed decreases with more diffusion steps due to increased computational overhead, demonstrating the need for careful optimization of the T parameter.

The experiments reveal that SpecDiff achieves optimal performance with:

- Draft lengths (γ) of 10-20 tokens (robust across this range)
- Diffusion steps (T) of 4-6 (requiring careful tuning to balance quality and speed)
- Initialization with standard speculative decoding for the first few tokens (beneficial for pretrained models)

4. Technical Contributions and Innovation

4.1. Architectural Flexibility

SpecDiff demonstrates that speculative decoding can effectively utilize drafters with completely different architectures than the target model. This breaks the conventional paradigm of using smaller versions of the same autoregressive architecture and opens new possibilities for heterogeneous model combinations.

4.2. Computational Efficiency Analysis

Using the Work-Depth parallel computation model, the paper shows that SpecDiff's depth (longest dependency chain) is primarily determined by T rather than γ. Since T is empirically smaller than typical draft lengths in traditional speculative decoding, this enables greater parallelization.

4.3. Quality Preservation Mechanism

The speculative decoding framework inherently preserves output quality by having the target LLM ultimately verify and correct all generated tokens. This allows SpecDiff to leverage the speed advantages of diffusion models without inheriting their quality limitations when used for standalone generation.

5. Significance and Impact

SpecDiff represents a paradigm shift in LLM inference acceleration by successfully bridging two distinct generative modeling approaches. The method's ability to achieve substantial speedups while maintaining output quality addresses critical challenges in making powerful LLMs more accessible and cost-effective.

The research opens several important avenues for future work, including optimization of diffusion model probability calibration for stochastic sampling, broader tokenizer compatibility, and integration with other acceleration techniques. The planned release of models on HuggingFace will enable broader community adoption and further development.

By demonstrating that discrete diffusion models can serve as effective drafters despite their architectural differences from autoregressive targets, this work expands the toolkit available for LLM optimization and suggests new directions for hybrid generative systems that combine the strengths of multiple model paradigms.

6. Relevant Citations

Fast inference from transformers via speculative decoding

This is a foundational paper that introduced the modern concept of speculative decoding. The SpecDiff method presented in the main paper is a direct extension of the framework described by Leviathan et al. and is used as the primary baseline for comparison.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In International Conference on Machine Learning, pages 19274–19286. PMLR.

Simple and effective masked diffusion language models

This citation is crucial as it introduces the Masked Diffusion Language Model (MDLM), which is the exact discrete diffusion model used as the drafter in the main paper's SpecDiff method. The core technical innovation of SpecDiff relies directly on the model and techniques described in this work.

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. arXiv preprint arXiv:2406.07524.

Discrete diffusion modeling by estimating the ratios of the data distribution

This paper is cited as a key motivation for the SpecDiff approach, demonstrating that recent discrete diffusion models can be significantly faster than traditional autoregressive models. It establishes the premise that replacing an autoregressive drafter with a diffusion-based one can lead to substantial speedups.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In Forty-first International Conference on Machine Learning.

Speculative decoding: Exploiting speculative execution for accelerating seq2seq generation

This work is identified as the earliest literature to adapt a non-autoregressive model (a masked language model) to act as the drafter in speculative decoding. It serves as an important conceptual precursor to SpecDiff, validating the core idea of moving beyond purely autoregressive drafters.

Heming Xia, Tao Ge, Peiyi Wang, Si-Qing Chen, Furu Wei, and Zhifang Sui. 2023. Speculative decoding: Exploiting speculative execution for accelerating seq2seq generation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 3909–3925.