SpecDiff-2: Scaling Diffusion Drafter Alignment For Faster Speculative Decoding

November 4, 2025

Jameson Sandler, Jacob K. Christopher, Thomas Hartvigsen, Ferdinando Fioretto

Summary

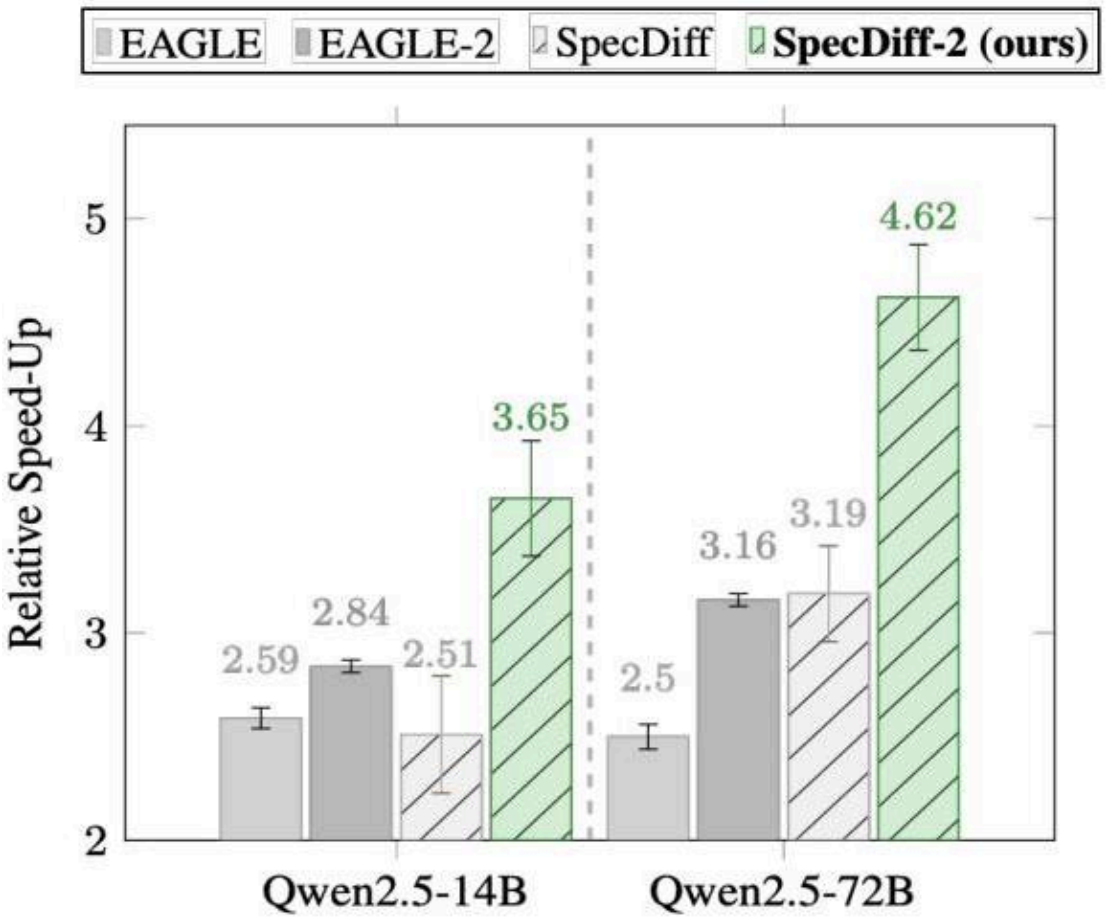
SPECDIFF-2 integrates discrete diffusion models as non-autoregressive drafters for speculative decoding, combined with novel alignment mechanisms. This framework achieves average 4.22x speed-ups in LLM inference without quality loss, enhancing throughput by 55% over prior baselines.

Table of Contents

1. Introduction

Large Language Models (LLMs) have achieved remarkable capabilities in text generation, reasoning, and problem-solving, but their autoregressive nature creates significant latency bottlenecks. These models generate text one token at a time, making inference increasingly slow as sequence lengths grow. This limitation is particularly problematic for complex reasoning tasks that require extensive "thinking time," such as mathematical problem-solving or code generation.

SPECDIFF-2 addresses this fundamental challenge by advancing speculative decoding—a framework that accelerates LLM inference without sacrificing output quality. The paper introduces the first principled approach to align diffusion-based drafters with autoregressive verifiers, achieving substantial speed-ups through both parallel token generation and intelligent alignment mechanisms.



2. Background and Motivation

Speculative decoding operates on a "draft-then-verify" principle where a smaller, faster drafter model proposes multiple tokens in parallel, while a larger verifier model validates

these proposals. This approach maintains the exact output distribution of the original model while reducing inference time. However, existing methods face two critical bottlenecks:

Autoregressive Dependency: Traditional drafters are themselves autoregressive, limiting their parallelism and creating sequential dependencies even during the drafting phase.

Drafter-Verifier Misalignment: When the drafter's proposals don't align well with what the verifier would generate, frequent rejections occur, forcing token regeneration and reducing effective speed-ups.

Previous work has attempted to address these issues separately, but SPECDIFF-2 is the first to tackle both simultaneously within a unified framework using diffusion models as non-autoregressive drafters.

3. Methodology

SPECDIFF-2's approach centers on two complementary alignment mechanisms designed specifically for diffusion drafters:

3.1. Diffusion-Based Drafting

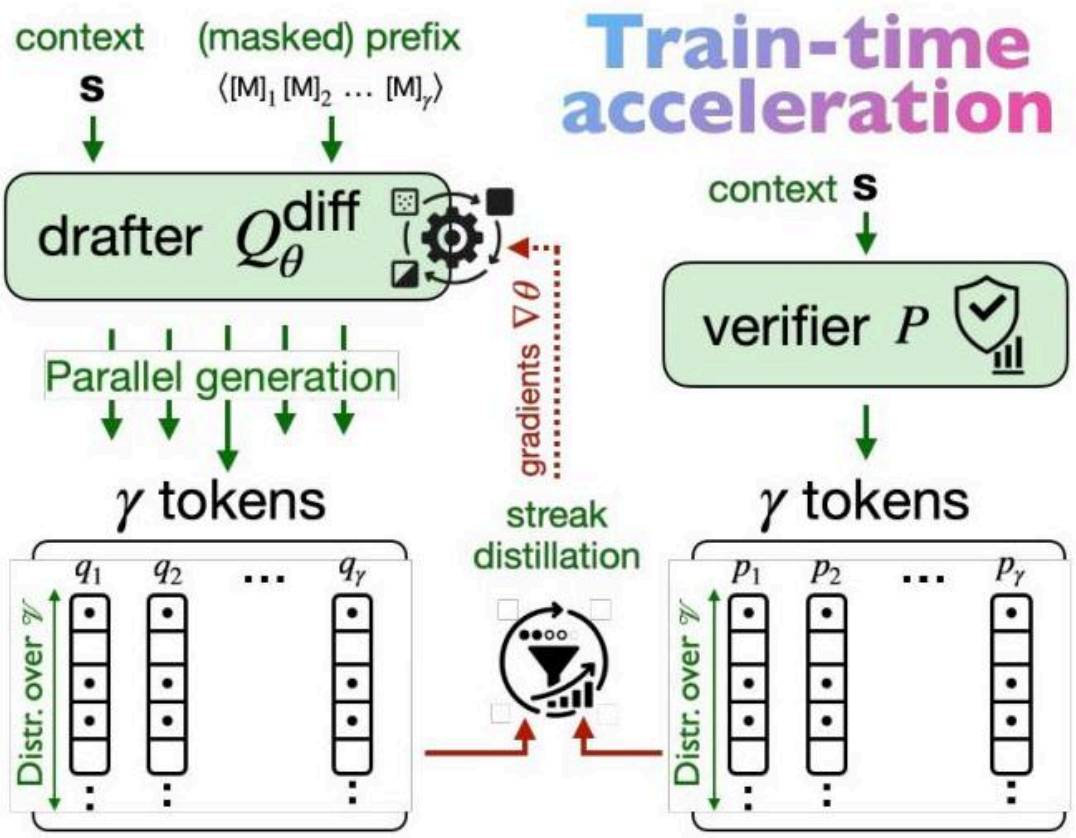
The framework employs Masked Discrete Diffusion Language Models (MDMs) as drafters. Unlike autoregressive models, these diffusion models generate all tokens in a sequence simultaneously through iterative denoising steps. Starting with a sequence of masked tokens, the diffusion drafter predicts token distributions for all positions in parallel, with drafting cost primarily dependent on the number of denoising steps rather than sequence length.

3.2. Train-Time Alignment: Streak-Distillation

Streak-distillation optimizes the diffusion drafter to maximize the expected length of accepted token sequences. The key insight is that acceptance probability depends on the entire draft window, not just individual tokens. The objective function is:

$$L_{\text{streak}} = E_{s,x_{1:\gamma}}\left[\sum_{j=1}^{\gamma}\prod_{i=1}^{j} P(x_i \mid s, x_{1:i-1})\right]$$

This formulation encourages the drafter to produce long, contiguous streaks of tokens that align with the verifier's distribution, considering the sequential dependencies inherent in autoregressive verification.

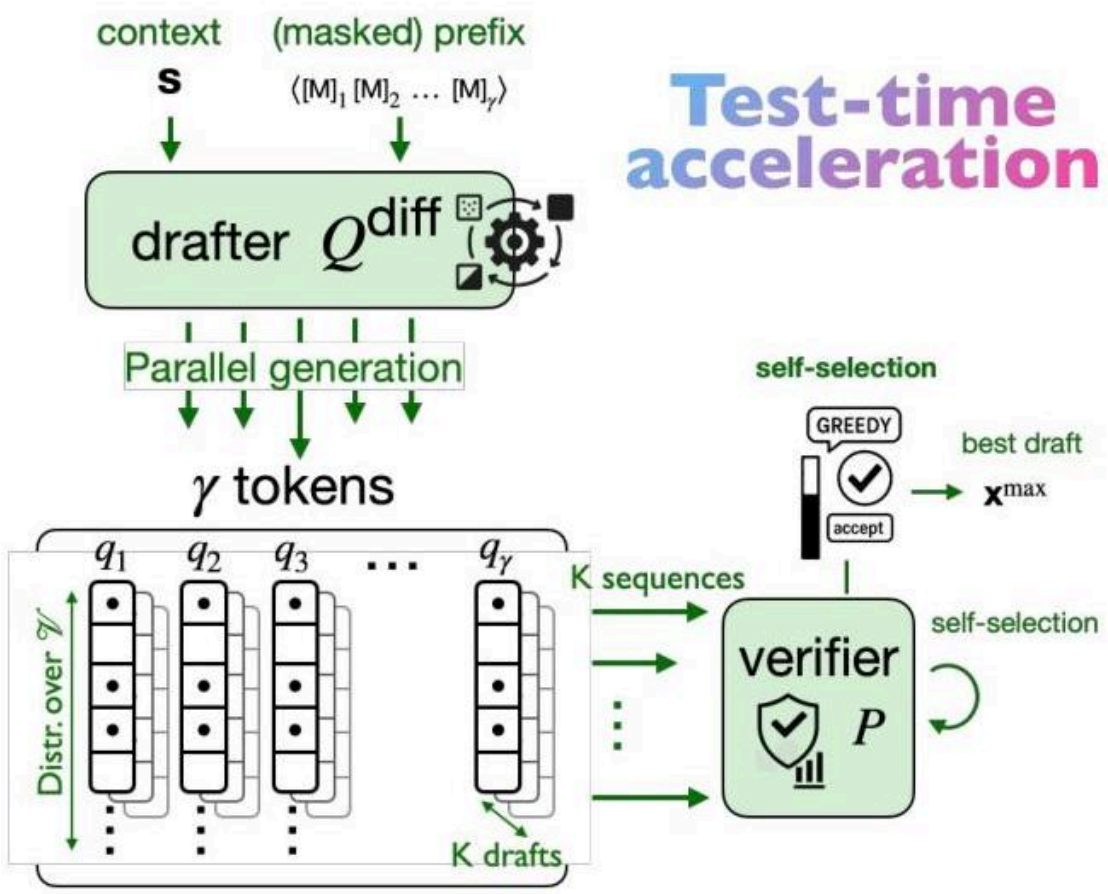


3.3. Test-Time Alignment: Self-Selection Acceptance

At inference time, self-selection acceptance leverages the diffusion drafter's ability to generate multiple draft candidates efficiently. From a single forward pass yielding position-wise marginal distributions, multiple joint drafts can be sampled in parallel. The verifier then evaluates each draft's expected throughput:

$$\text{Throughput}(x, s) = \sum_{j=1}^{|\gamma|} \prod_{i=1}^{j} P(x_i | s, x_{1:i-1})$$

The draft with maximum expected throughput is selected for verification, increasing the likelihood of longer accepted sequences.



4. Results and Effectiveness

SPECDIFF-2 demonstrates substantial improvements across diverse benchmarks, achieving up to 5.5× speed-ups over standard decoding and consistent gains over state-of-the-art methods like EAGLE-2.

4.1. Speed-Up Performance

The method excels particularly on structured tasks requiring precise reasoning:

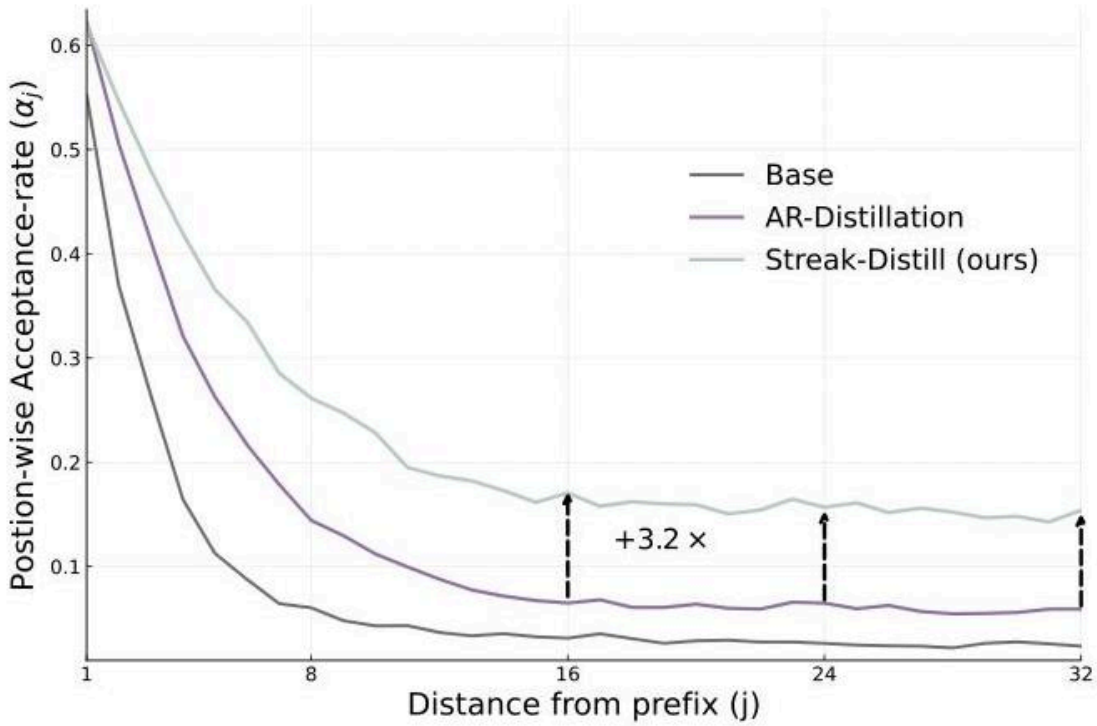
- **Mathematical Reasoning (Math-500)**: 4.71× average speed-up vs. EAGLE-2's 3.43×
- **Code Generation (HumanEval)**: Similar substantial improvements
- **Open-Ended Tasks (GPQA)**: 3.24× speed-up vs. EAGLE-2's 2.80×

4.2. Alignment Mechanism Validation

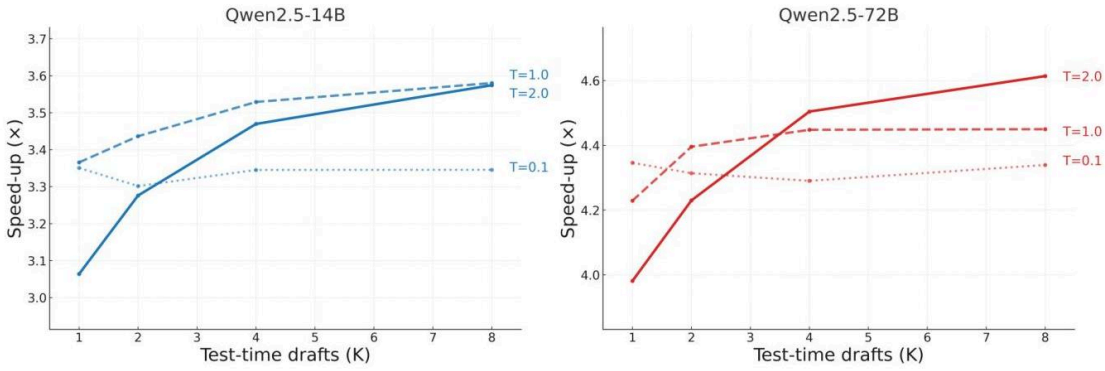
Empirical analysis reveals the effectiveness of both alignment mechanisms:

Streak-Distillation Impact: Position-wise acceptance rates show dramatic improvements throughout the draft window. While traditional autoregressive alignment degrades rapidly

for later tokens, streak-distillation maintains strong alignment, improving average acceptance at later positions by 3.2× over AR approaches.



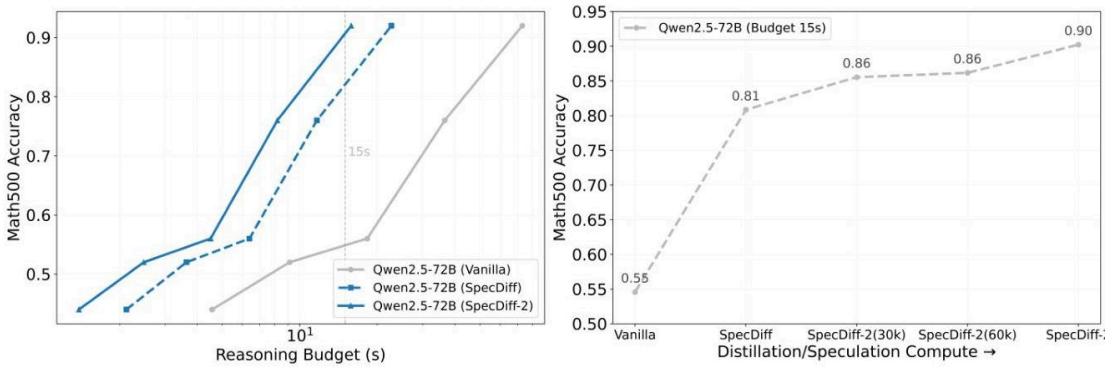
Self-Selection Scaling: The method exhibits smooth scaling with the number of parallel drafts, achieving up to +20% additional throughput when using 8 parallel drafts with appropriate temperature settings.



4.3. Acceleration-Compute Scaling

Perhaps most significantly, SPECDIFF-2 validates the concept of "acceleration-compute" scaling, where inference speed improvements directly translate to better task performance under fixed time constraints. On mathematical reasoning tasks with a 15-second budget, the

method achieved a +63% accuracy improvement over vanilla decoding and +11% over unaligned diffusion drafting.



5. Significance and Future Directions

SPECDIFF-2 makes several important contributions to LLM inference optimization:

Methodological Innovation: The work establishes the first principled framework for aligning diffusion drafters with autoregressive verifiers, opening new architectural possibilities for speculative decoding systems.

Performance Advancement: By achieving state-of-the-art speed-ups while maintaining lossless generation quality, the method makes LLMs more practical for latency-sensitive applications.

Conceptual Framework: The introduction of "acceleration-compute" scaling provides a new lens for understanding the relationship between inference speed and model capability, particularly relevant for reasoning-intensive tasks.

Broader Impact: The superior performance on structured tasks like mathematical reasoning and code generation makes LLMs more viable tools for scientific computing, automated problem-solving, and software development.

The research also identifies several promising directions for future work, including formal analysis of acceptance rules, cross-family drafting approaches, and hardware-specific optimizations for diffusion-based inference. These directions suggest that SPECDIFF-2 represents not just an incremental improvement but a foundational contribution that could influence the broader trajectory of LLM inference optimization research.

6. Relevant Citations

Fast inference from transformers via speculative decoding

This paper is a foundational work that introduced the speculative decoding framework, a draft-then-verify procedure that the presented paper builds upon. It establishes the core principles of using a smaller drafter model to accelerate a larger verifier model, which is central to SpecDiff-2.

Leviathan, Y., Kalman, M., and Matias, Y. Fast inference from transformers via speculative decoding. In International Conference on Machine Learning, pp. 19274–19286. PMLR, 2023.

Speculative diffusion decoding: Accelerating language generation through diffusion

This is the direct predecessor to the main paper, introducing the original SpecDiff system which first proposed using non-autoregressive diffusion models as drafters. SpecDiff-2 directly addresses the key limitation of this work, which is the misalignment between the diffusion drafter and the autoregressive verifier.

Christopher, J. K., Bartoldson, B. R., Ben-Nun, T., Cardei, M., Kailkhura, B., and Fioretto, F. Speculative diffusion decoding: Accelerating language generation through diffusion. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 12042–12059, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long. 601. URL https://aclanthology.org/2025. naacl-long.601/.

EAGLE-2: Faster inference of language models with dynamic draft trees

This paper presents EAGLE-2, a state-of-the-art autoregressive speculative decoding method that serves as a primary baseline for SpecDiff-2's performance evaluation. It represents the main alternative approach that SpecDiff-2 aims to outperform, providing context for its claims of achieving a new state-of-the-art.

Li, Y., Wei, F., Zhang, C., and Zhang, H. Eagle-2: Faster inference of language models with dynamic draft trees. arXiv preprint arXiv:2406.16858, 2024b.

Distillspec: Improving speculative decoding via knowledge distillation

This work introduced a key method for aligning an autoregressive drafter with a verifier using knowledge distillation. The 'streak-distillation' objective proposed in SpecDiff-2 is a direct response and novel adaptation of this concept, tailored for the unique properties of non-autoregressive diffusion drafters.

Zhou, Y., Lyu, K., Rawat, A. S., Menon, A. K., Rostamizadeh, A., Kumar, S., Kagy, J.-F., and Agarwal, R. Distillspec: Improving speculative decoding via knowledge distillation. arXiv preprint arXiv:2310.08461, 2023.