






[← Back to Author Console](#)

# TPP-SD: Accelerating Transformer Point Process Sampling with Speculative Decoding

Shukai Gong, 

 05 May 2025 (modified: 17 May 2025)  NeurIPS 2025 Conference Submission  Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors 

[Revisions](#)  CC BY 4.0


**Keywords:** Transformer point process, sampling, speculative decoding

**TL;DR:** TPP-SD is a fast and distributionally consistent sampling framework for Transformer-based temporal point processes, achieving 2–6× speedup over autoregressive methods by adapting speculative decoding with a lightweight draft model.

**Abstract:**

We propose TPP-SD, a novel approach that accelerates Transformer temporal point process (TPP) sampling by adapting speculative decoding (SD) techniques from language models. By identifying the structural similarities between thinning algorithms for TPPs and speculative decoding for language models, we develop an efficient sampling framework that leverages a smaller draft model to generate multiple candidate events, which are then verified by the larger target model. TPP-SD maintains the same output distribution as autoregressive sampling while achieving significant acceleration. Experiments on both synthetic and real datasets demonstrate that our approach produces samples from identical distributions as standard methods, but with 2-6× speedup. Our ablation studies analyze the impact of hyperparameters such as draft length and draft model size on sampling efficiency. TPP-SD bridges the gap between powerful Transformer TPP models and the practical need for rapid sequence generation.


**PDF:**   pdf

**Checklist Confirmation:**  I confirm that I have included a paper checklist in the paper PDF.


**Supplementary Material:**   zip


**Financial Support:**  Shukai Gong

**Reviewer Nomination:**  Shukai Gong, 

**Responsible Reviewing:**  We acknowledge the responsible reviewing obligations as authors.

**Primary Area:** General machine learning (supervised, unsupervised, online, active, etc.)

**LLM Usage:**  Editing (e.g., grammar, spelling, word choice)

**Declaration:**  I confirm that the above information is accurate.

**Submission Number:** 