Shift Parallelism vs Pipeline Parallelism for Multimodal LLM Token Throughput Efficiency

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Can Shift Parallelism maintain high token throughput efficiency when scaled to multimodal LLMs processing variable-length image-text sequences compared to pipeline parallelism. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Shift Parallelism: Low-Latency, High-Throughput LLM Inference for Dynamic Workloads. Research question: Can Shift Parallelism maintain high token throughput efficiency when scaled to multimodal LLMs processing variable-length image-text sequences compared to pipeline parallelism?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

15 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 4.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Llama-3.3-70B serves diverse use cases including sentiment analysis, retrieval-augmented generation (RAG), and coding ag	×	0.03
Interactive or latency-sensitive requests typically arrive one or a few at a time, with response latencies (TTFT and TPO	×	0.04
Batch or throughput-sensitive requests usually arrive in large volumes (thousands to millions at once), where aggregate	×	0.03
Time-to-first-token (TTFT) is the time after a client submits a prompt until the first characters of response text (toke	×	0.07
Time-per-output-token (TPOT) is the time between each subsequent token until the response is completed after the first r	×	0.04
Combined throughput is the total number of tokens (both prompt and response) processed by the inference system per unit	×	0.06
TTFT and TPOT shape the quality of service for interactive applications.	×	0.04
Combined throughput shapes the quality of service for batch use cases and impacts the cost of running the service for th	×	0.03
A vanilla LLM involves a series of transformer layers, each consisting of an attention mechanism and a multi-layer perce	×	0.02
The weights in the transformer layer correspond to the QKV (query, key, and value).	×	0.01
In the benchmark table, Shift Parallelism achieves a TTFT of 102 ms for Llama-70B.	×	0.06
In the benchmark table, Shift Parallelism achieves a TPOT of 10.1 ms for Llama-70B.	×	0.06
In the benchmark table, Shift Parallelism achieves a combined throughput of 37.4k tokens/s for Llama-70B.	×	0.08
In the benchmark table, Shift Parallelism achieves a TTFT of 86.41 ms for Qwen-32B.	×	0.05
In the benchmark table, Shift Parallelism achieves a TPOT of 9.48 ms for Qwen-32B.	×	0.06
In the benchmark table, Shift Parallelism achieves a combined throughput of 53.8k tokens/s for Qwen-32B.	×	0.08

References

- http://arxiv.org/abs/2509.16495v2
- http://arxiv.org/abs/2508.19294v2
- http://arxiv.org/abs/2510.05186v1