PR #20972 完整报告

sgl-project/sglang

Remove sync when enabling return_logprob

合并时间：2026-03-28 07:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20972

执行摘要

本 PR 通过移除启用 return_logprob 时的 GPU-CPU 同步操作，采用异步传输模式，将 token 吞吐量从 196.23 tok/s 提升至 246.91 tok/s，优化了采样器和调度器输出处理的性能。

功能与动机

PR 旨在解决启用 return_logprob 时的性能瓶颈，benchmark 显示同步延迟导致吞吐量较低。作者引用模式：'compute logprobs without copying it to cpu -> do gpu to cpu transfer in copy_to_cpu(), which is async -> convert cpu tensor to list in scheduler'，以消除不必要的等待时间，提升推理效率。

实现拆解

- sampler.py 模块（采样层）：修改 _attach_logprobs_to_output 和 compute_logprobs_only 函数，为 get_top_logprobs 和 get_token_ids_logprobs 调用添加 no_copy_to_cpu=True 参数，避免 logprobs 计算时的同步复制。
- scheduler_output_processor_mixin.py 模块（调度管理器）：在 process_batch_result_prefill 和 process_batch_result_decode 函数中添加异步转换代码，将 GPU tensor 转换为 list，移除对 batch.is_spec_v2 的条件检查，统一处理 top logprobs 和 token ids logprobs。

评论区精华

review 中仅有 reviewer ispobock 批准 PR，未提出评论或争议，因此无讨论内容。

风险与影响

- 风险：异步 GPU-CPU 传输可能引入数据一致性风险，如 scheduler 处理时数据未就绪；修改核心路径可能影响其他依赖同步的代码；缺少全面回归测试覆盖。
- 影响：用户端直接提升 token 吞吐量约 25.8%，系统端优化 GPU 资源利用率，团队需关注异步模式的稳定性。

关联脉络

与近期 PR 21503（JIT 内核性能优化）和 21514（调度器 tensor 处理修复）相关，体现了仓库在性能优化和调度正确性方面的持续演进。