PR #24344 完整报告

sgl-project/sglang

[Fix] NGRAMWorker.update_weights_from_tensor — delegate to target worker

合并时间：2026-05-05 07:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24344

执行摘要

- 一句话：为 NGRAMWorker 添加 update_weights_from_tensor 委托方法
- 推荐动作：该 PR 是一个正确且简洁的修复，值得纳入。虽然变动很小，但修复了一个影响 NGRAM 可用性的关键 bug。开发者可以快速合并。

功能与动机

当 SGLang 以 --speculative-algorithm NGRAM 启动时，任何调用 update_weights_from_tensor 的代码（如 RL 框架的在线权重同步）都会崩溃，错误为 AttributeError: 'NGRAMWorker' object has no attribute 'update_weights_from_tensor'。调度器混入类 scheduler_update_weights_mixin.py 通过 self.draft_worker or self.tp_worker 分发此方法，而 EAGLEWorker 实现了它，但 NGRAMWorker 没有，导致首次调用即失败。关联 Issue #24343 记录了此问题。

实现拆解

1. 在 python/sglang/srt/speculative/ngram_worker.py 的 clear_cache_pool 方法之后、add_external_corpus 方法之前，新增 update_weights_from_tensor 方法。
2. 该方法仅一行：return self.target_worker.update_weights_from_tensor(recv_req)，将调用委托给目标 worker。
3. 语义分析：NGRAM 没有可学习的权重，其 NgramCorpus 是基于请求 token 流构建的 CPU 查找结构；NGRAMWorker.model_runner 就是 target_worker.model_runner，因此委托是正确且高效的。

关键文件：

- python/sglang/srt/speculative/ngram_worker.py（模块 投机解码；类别 source；类型 core-logic；符号 update_weights_from_tensor）：新增了 update_weights_from_tensor 方法，委托给 target_worker，修复了 NGRAM 模式下 RL 框架调用崩溃的 bug。

关键符号：update_weights_from_tensor

关键源码片段

python/sglang/srt/speculative/ngram_worker.py

新增了 update_weights_from_tensor 方法，委托给 target_worker，修复了 NGRAM 模式下 RL 框架调用崩溃的 bug。

```
# python/sglang/srt/speculative/ngram_worker.py
# 新增于 clear_cache_pool 方法之后，add_external_corpus 方法之前

def update_weights_from_tensor(self, recv_req):
    # NGRAM 没有自己的可学习权重，其 n-gram 语料库是基于请求 token 流
    # 构建的 CPU 查找结构。model_runner 实际上与 target worker 共享。
    # 调度器混入类通过 `self.draft_worker or self.tp_worker` 分发此方法，
    # 因此没有此方法时调用会抛出 AttributeError。
    return self.target_worker.update_weights_from_tensor(recv_req)
```

评论区精华

审阅者均快速批准了该 PR，无深入技术争议。PR 作者在 body 中详细解释了为什么委托是正确的修复方式，并讨论了是否需要对 n-gram corpus 进行失效处理，但认为 corpus 会自然更新，无需显式重置。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅添加了一行委托方法，不影响 NGRAMWorker 的其他逻辑（clear_cache_pool, add_external_corpus 等）。update_weights_from_tensor 不被热路径调用，因此对推理性能无影响。委托路径与 EAGLEWorker 的实现模式一致，经过了 TpModelWorker 的验证。
- 影响：修复了 NGRAM 模式下 RL 权重同步的阻塞 bug，使得 --speculative-algorithm NGRAM 可以与在线学习框架（如 slime、OpenRLHF、verl）配合使用。内部评测显示 NGRAM 在 Qwen2.5-32B SFT 上相比非推测解码基线带来了约 3.2 倍的吞吐量提升。
- 风险标记：暂无

关联脉络

- PR #24343 [Bug] NGRAMWorker missing update_weights_from_tensor — RL weight sync crashes with AttributeError: 关联 issue，报告并讨论了此 bug。