PR #22882 完整报告

sgl-project/sglang

[HiSparse][BugFix]: Fix the memory leak issue during health checks.

合并时间：2026-04-15 19:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22882

执行摘要

- 一句话：修复 HiSparse 解码模式下健康检查时的内存泄漏问题。
- 推荐动作：该 PR 值得快速浏览，重点关注 process_batch_result_prebuilt 方法中新增的 HiSparse 协调器通知逻辑。设计决策是仅修复直接导致泄漏的问题，而未采纳 review 中关于补充多模态和 MoE 清理的建议，这可能是一个权衡点，需关注后续是否会出现相关内存问题。

功能与动机

PR body 中的错误日志显示，当启动 HiSparse 解码并发送健康检查请求时，调度器在 check_memory 过程中检测到 token_to_kv_pool_allocator 内存泄漏，导致 ValueError 异常。这表明在请求完成路径中，HiSparse 相关的资源未被正确清理。

实现拆解

1. 定位泄漏点：问题出现在 process_batch_result_prebuilt 方法中，当请求完成时，会释放 KV 缓存，但未通知 HiSparse 协调器进行其内部的资源清理。
2. 添加清理调用：在 python/sglang/srt/managers/scheduler_output_processor_mixin.py 文件的 process_batch_result_prebuilt 方法内，于 release_kv_cache 调用之前，新增条件判断 if self.enable_hisparse:，并调用 self.hisparse_coordinator.request_finished(req)。这确保了在 HiSparse 启用时，请求完成能触发协调器的清理逻辑。
3. 配套改动：本次变更仅涉及核心逻辑文件，未包含测试、配置或文档的配套改动。

关键文件：

- python/sglang/srt/managers/scheduler_output_processor_mixin.py（模块 调度器；类别 source；类型 core-logic；符号 process_batch_result_prebuilt）：这是修复内存泄漏的核心文件，在请求完成处理流程中添加了 HiSparse 协调器通知。

关键符号：process_batch_result_prebuilt

关键源码片段

python/sglang/srt/managers/scheduler_output_processor_mixin.py

这是修复内存泄漏的核心文件，在请求完成处理流程中添加了 HiSparse 协调器通知。

```
def process_batch_result_prebuilt(self: Scheduler, batch: ScheduleBatch):
```

```
    assert self.disaggregation_mode == DisaggregationMode.DECODE
    for req in batch.reqs:
        req.time_stats.set_decode_prebuilt_finish_time()
        req.check_finished()
        if req.finished():
            req.time_stats.set_quick_finish_time()
            # 新增：当 HiSparse 启用时，通知协调器请求已完成，以便清理相关资源
            if self.enable_hisparse:
                self.hisparse_coordinator.request_finished(req)
            release_kv_cache(req, self.tree_cache) # 原有的 KV 缓存释放

    # 注意：Logprobs 应在预填充引擎上处理
    self.stream_output(batch.reqs, batch.return_logprob)
```

评论区精华

reviewer gemini-code-assist[bot] 指出，此修复块缺少其他请求完成路径（如 _handle_finished_req）中存在的多模态输入清理和 MoE 专家收集步骤，建议补充以防止多模态内存泄漏并确保 MoE 专家被收集，同时建议调用 set_completion_time() 以保证跨执行路径的计时指标一致性。但作者未采纳该建议，最终仅添加了 HiSparse 协调器通知。

- 缺失的清理步骤 (correctness): 作者未采纳建议，仅添加了 HiSparse 协调器通知。

风险与影响

- 风险：风险较低：变更范围极小（仅 2 行代码），且位于明确的请求完成条件分支内。
- 回归风险：如果 self.hisparse_coordinator.request_finished 方法本身存在缺陷，可能引入新的问题。
- 兼容性：仅当 self.enable_hisparse 为 True 时执行，不影响非 HiSparse 模式。
- 测试覆盖：未添加新测试，依赖现有测试验证 HiSparse 功能。
- 影响：影响范围：主要影响使用 HiSparse 解码模式的用户，特别是进行健康检查或请求完成的场景。
- 用户影响：修复了内存泄漏，提升了系统稳定性和资源利用率，用户不再遇到健康检查时的异常崩溃。
- 系统影响：避免了因内存泄漏导致的潜在调度错误和性能下降。
- 团队影响：代码变更简单，易于理解和维护，但 review 中提到的缺失清理步骤可能在未来引发其他内存问题。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #22767 [HiCache] Fix memory host free logic when share_indices_with_anchor enabled: 同属内存泄漏修复，涉及 HiCache 模块，与本 PR 的 HiSparse 内存泄漏修复在主题上相关。

- PR #22753 Fix streaming session busy-check double-counting via active_pool_idxs: 同属内存和调度相关 bugfix，涉及流式会话和 KV 缓存，与本 PR 的内存泄漏修复在技术领域上相关。