PR #25064 完整报告

sgl-project/sglang

[Bug Fix] Add priority property to DecodeRequest to fix AttributeError with --enable-priority-scheduling

合并时间：2026-05-14 11:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25064

执行摘要

- 一句话：修复 DecodeRequest 缺少 priority 属性导致的崩溃
- 推荐动作：值得精读，因为：
  1. 示例了数据类代理属性的标准做法；
  2. 展示了跨模块调用在分离模式下的属性缺失 bug 模式；
  3. 代码变更虽小但修复了关键路径的崩溃。后续可考虑为 DecodeRequest 添加单元测试，验证所有委托属性都与 Req 同步。

功能与动机

当同时使用 --disaggregation-mode decode 和 --enable-priority-scheduling 时，decode worker 在 metrics 上报时因 DecodeRequest 缺少 priority 属性而崩溃。关联 Issue #25057 详细描述了调用链：event_loop_normal_disagg_decode → process_batch_result_decode → report_decode_stats → QueueCount.from_reqs() → req.priority。该 bug 影响了 PD decode 模式下的 priority 调度功能。

实现拆解

1. 在 python/sglang/srt/disaggregation/decode.py 的 DecodeRequest 数据类中新增 priority 属性，类型标注为 Optional[int]，实现委托给 self.req.priority。
2. 与已有的 seqlen 属性采用完全相同的代理模式，保持代码风格一致。
3. 此变更仅添加 4 行代码，无测试文件变更，但 PR 作者在 2×RTX 3090 上使用 fake 传输后端验证了修复：8 个并发请求返回 200，/metrics 端点正常，decode 指标携带了 priority 标签。

关键文件：

- python/sglang/srt/disaggregation/decode.py（模块 调度器；类别 source；类型 core-logic；符号 priority）：在 DecodeRequest 数据类中新增 priority 属性，修复 AttributeError。

关键符号：DecodeRequest.priority

关键源码片段

python/sglang/srt/disaggregation/decode.py

在 DecodeRequest 数据类中新增 priority 属性，修复 AttributeError。

```
@dataclass
class DecodeRequest:
    req: Req
    kv_receiver: CommonKVReceiver
    waiting_for_input: bool = False
    metadata_buffer_index: int = -1

    @property
    def seqlen(self) -> int:
        return self.req.seqlen

    # 新增 priority 属性，代理到内部的 Req 对象，
    # 用于 QueueCount.from_reqs() 在 priority 调度启用时读取。
    @property
    def priority(self) -> Optional[int]:
        return self.req.priority
```

评论区精华

Review 中 gemini-code-assist[bot] 建议为 priority 属性添加类型标注 Optional[int]，以与 seqlen 属性保持一致性并提升类型安全。该建议被采纳，最终代码中包含了类型标注。审核者 ShangmingCai 批准了 PR，认为这是一个小而干净的修复，可以跳过 CI 以节省资源。

- 添加类型标注建议 (style): 类型标注被采纳并体现在最终代码中。

风险与影响

- 风险：风险极低：变更仅为在数据类上添加一个属性代理，不涉及任何逻辑修改。但缺少自动化测试覆盖，未验证真实传输后端场景，后续重构可能需要额外维护。
- 影响：直接影响：启用 PD decode 模式 + priority 调度的用户不再遇到 AttributeError 崩溃。影响范围限定在 decode 预分配 / 传输队列中使用的 DecodeRequest 对象。对 prefill 侧无影响。
- 风险标记：缺少测试覆盖，低影响变更

关联脉络

- PR #25057 [Bug] AttributeError: 'DecodeRequest' object has no attribute 'priority': 关联的 bug 报告，本 PR 直接修复此 issue。