

PR #28221 完整报告

sgl-project/sglang

Fix EagleDraftExtendInput missing kv_indptr crash with triton/DP attention

合并时间: 2026-06-18 05:54

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28221>

执行摘要

- 一句话: 修复 EagleDraftExtendInput 缺少 kv_indptr 崩溃
- 推荐动作: 值得合并的小修复, 解决了因接口不兼容导致的崩溃问题。建议后续考虑添加类型注解 `Optional[torch.Tensor]` 以通过静态检查。

功能与动机

PR #24860 将 draft-extend 的 spec_info 从 EagleDraftInput 切换为 EagleDraftExtendInput, 但后者缺少 kv_indptr 属性。多个 attention 后端 (triton、flashinfer、flashinfer_mla、aiter、wave) 在 init_forward_metadata 中直接读取 spec_info.kv_indptr, 导致 AttributeError。该问题在 AMD CI 中暴露 (因使用 `--attention-backend triton`), 而 CUDA lane 因使用 fa3 后端 (不访问 kv_indptr) 未触发。

实现拆解

1. 在 `python/sglang/srt/speculative/eagle_info.py` 的 EagleDraftExtendInput 类中新增字段 `kv_indptr: torch.Tensor = None`, 位于 `__post_init__` 方法之前。
2. 当 kv_indptr 为 None 时 (draft-extend idle batch 情况), attention 后端回退到从 `seq_lens` 重建普通 metadata, 恢复 PR #24860 之前的行为。
3. 仅修改一个文件, 新增 4 行代码, 无其他配套变更。

关键文件:

- `python/sglang/srt/speculative/eagle_info.py` (模块 推测解码; 类别 source; 类型 core-logic; 符号 EagleDraftExtendInput): 核心修复文件, 在 EagleDraftExtendInput 类中新增 kv_indptr 字段以恢复接口兼容性。

关键符号: 未识别

关键源码片段

`python/sglang/srt/speculative/eagle_info.py`

核心修复文件, 在 EagleDraftExtendInput 类中新增 kv_indptr 字段以恢复接口兼容性。

```
# File: python/sglang/srt/speculative/eagle_info.py
# 在 EagleDraftExtendInput 类中新增 kv_indptr 字段, 修复与 attention 后端的接口兼容性。
# 当 kv_indptr 为 None 时 (draft-extend idle batch), attention 后端从 seq_lens 重建
```

metadata。

```
class EagleDraftExtendInput(SpecInput):
    # ... 前面字段省略 ...
    num_tokens_per_req: int = -1
    num_tokens_for_logprob_per_req: int = 1

    # None 对应 draft-extend 的 idle 批次;
    # attention 后端在其为 None 时回退到从 seq_lens 重建普通 metadata。
    kv_indptr: torch.Tensor = None # 新增字段

    def __post_init__(self):
        super().__init__(SpecInputType.EAGLE_DRAFT_EXTEND)
```

评论区精华

1. 类型注解建议: gemini-code-assist[bot] 建议将 kv_indptr 的类型注解从 torch.Tensor 改为 Optional[torch.Tensor], 因为默认值为 None, 使用 Optional 可防止静态类型分析错误。该建议未被采纳或回应。
 2. 维护者确认: HaiShaw 对 gemini-code-assist 的建议回复了 "?", 暗示其不确定或需要讨论, 但 PR 最终以 torch.Tensor = None 的形式合并。
- 类型注解应为 Optional[torch.Tensor] (style): HaiShaw 回复 '?' 表示不确定, 最终 PR 以 torch.Tensor = None 的形式合并, 未采纳建议。

风险与影响

- 风险:
 1. 回归风险低: 变更仅添加一个默认 None 的字段, 无行为改动, 现有逻辑在 kv_indptr 为 None 时会回退到 seq_lens 路径, 与旧版本兼容。
 2. 静态类型警告: 未使用 Optional 类型注解可能导致 mypy/pyright 警告, 但不影响运行时。
 3. 测试覆盖: PR 未新增测试, 但依赖的 CI 测试 test_eagle_dp_attention.py 已通过。
- 影响:
 1. 用户影响: 修复了 EAGLE3 + DP attention 在 triton/AMD 等后端的崩溃, 用户无需额外配置即可正常使用。
 2. 系统影响: 影响所有使用非 fa3 attention 后端 (triton、flashinfer、flashinfer_mla、aiter、wave) 的 EAGLE 推测解码场景。
 3. 团队影响: 最小化变更, 易于 review 和合并, 无需跨模块协调。 - 风险标记: 缺少测试覆盖

关联脉络

- PR #24860 [Spec] Install EagleDraftExtendInput as the V2 draft-extend spec_info: 本 PR 修复的 bug 由 PR #24860 引入, 后者将 draft-extend 的 spec_info 从

EagleDraftInput 切换为 EagleDraftExtendInput, 但未包含 kv_indptr 属性。