PR #46760 完整报告

vllm-project/vllm

[ROCm][Bugfix] Pass num_kv_splits to aiter mla_reduce_v1

合并时间：2026-06-27 05:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46760

执行摘要

- 一句话：修复 ROCm MLA 预填充崩溃，传入 num_kv_splits 参数
- 推荐动作：该 PR 是应对上游 AITER 接口变更的适配修复，变更极小（+3 行注释 +1 行逻辑），但修复了一个阻断性 bug，值得立即合并和回传。建议回顾者重点关注参数位置是否正确，以及是否有其他未覆盖的 mla_reduce_v1 调用点。

功能与动机

AITER 更新到 v0.1.16.post2 后，mla_reduce_v1 的签名新增了必需的 num_kv_splits 参数（位置 7）。vLLM 的调用仍使用旧签名，导致 out_3d 张量被绑定到 num_kv_splits 整数槽位，引发 RuntimeError。每个使用 FP8 MLA 预填充的模型（如 DeepSeek-R1-0528、Kimi-K2.6）在 gfx950 上首次预填充请求时均崩溃。

实现拆解

1. 定位问题：在 vllm/v1/attention/backends/mla/rocm_aiter_mla.py 的 _mla_fp8_prefill_attn 方法中，第 828 行附近调用 self._mla_reduce_v1 时，参数列表缺少 num_kv_splits。
2. 修改调用签名：在 tile_q 参数后插入整型参数 0，作为 num_kv_splits 的传值。该值使内核内部 max(cu_num, 0) == cu_num，完全匹配 AITER 升级前的行为。
3. 验证：在 MI355X (gfx950) 上使用 DeepSeek-R1-0528-MXFP4 TP=8 配置测试，修复前首次预填充即崩溃，修复后端到端正常。A/B 对比显示精度（gsm8k）和延迟（TPOT/ITL/E2EL）无回归，吞吐量持平或略有提升。

关键文件：

- vllm/v1/attention/backends/mla/rocm_aiter_mla.py（模块 注意力；类别 source；类型 core-logic）：修复的核心文件，在 MLA FP8 预填充缩减调用中补传 num_kv_splits 参数。

关键符号：_mla_fp8_prefill_attn

关键源码片段

vllm/v1/attention/backends/mla/rocm_aiter_mla.py

修复的核心文件，在 MLA FP8 预填充缩减调用中补传 num_kv_splits 参数。

```
# vllm/v1/attention/backends/mla/rocm_aiter_mla.py
```

```
# Phase 2: reduction across KV splits.
self._mla_reduce_v1(
    logits,
    attn_lse,
    attn_metadata.fp8_prefill_reduce_indptr,
    attn_metadata.fp8_prefill_reduce_final_map,
    attn_metadata.fp8_prefill_reduce_partial_map,
    tile_q,
    # num_kv_splits added by ROCm/aiter#3391; 0 selects the kernel
    # default max(cu_num, 0) == cu_num, matching pre-#3391 behavior.
    0,
    out_3d,
    final_lse,
)
```

这段代码展示了修复后对 _mla_reduce_v1 的调用。关键变更是在 tile_q 之后、out_3d 之前插入整型参数 0。该参数对应 num_kv_splits，传入 0 使内核内部 params.max_splits = max(cu_num, 0) = cu_num，完全复现升级前的行为，同时避免使用 out_3d 张量错误地占据该槽位。

评论区精华

本 PR 无 review 讨论记录。维护者 dllehr-amd 在两次审核中均批准（LGTM），无异议。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 回归风险低：仅修改一个函数调用的参数列表，传入硬编码 0，逻辑等价于旧行为。
  - 覆盖范围窄：仅影响 ROCm 平台上使用 AITER MLA FP8 预填充后端的模型（如 DeepSeek-R1、Kimi-K2.6 及其 MXFP4 变体）。不影响其他后端或解码阶段。
  - 兼容性：依赖于升级后的 AITER（>= v0.1.16.post2），若回退 AITER 版本会导致不匹配，但这是 AITER 版本更新带来的预期需要。
- 影响：
  - 用户：修复了 ROCm (gfx950) 上 DeepSeek-R1、Kimi-K2.6 等 MLA 模型在预填充时的崩溃故障。受影响的用户可以正常使用 MLA FP8 预填充。
  - 系统：无外部接口变化，不影响 API 兼容性。
  - 团队：低风险、高价值的修复，维护成本极低。
  - 风险标记：第三方依赖接口变更，仅 ROCm gfx950，低风险

关联脉络

- PR #46692 [Frontend][Gpt-oss] Use process_eos() to flush Harmony Parser outputs.: 此 PR 将 AITER 升级到 v0.1.16.post2，引入了 mla_reduce_v1 签名变更，是导致本次 bug 的根源。

- PR #3689 aiter #3391 Caused SGLang MLA FP8/MXFP4 Prefill Paths Fail: SGLang 也遇到了完全相同的 bug，证明了此问题的广泛性，并验证了修复方案的正确性。