PR #37228 完整报告

vllm-project/vllm

[ROCM][Bugfix] Use correct stride in cp_mha_gather_cache_kernel for hybrid model (#37228)

合并时间：2026-03-27 01:33

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37228

执行摘要

- 一句话：修复 ROCM 后端在混合模型下 KV 缓存非连续内存访问错误，避免注意力输出 NaN。
- 推荐动作：该 PR 值得精读，尤其关注 Triton 内核中处理非连续内存的通用模式。设计决策亮点：采用传递 stride 而非仅第一维 stride，以预防未来其他维度非连续导致的静默错误。建议团队审查其他类似内核是否存在相同假设，并优先修复 shuffle 路径问题。

功能与动机

混合模型（如 Qwen3.5）的 KV 缓存布局为交错模式 [K_0][V_0][K_1][V_1]...，而非原始假设的连续布局 [K_all][V_all]。update_hybrid_attention_mamba_layout 使用 as_strided_() 重排 KV 块导致内存非连续，但内核使用硬编码指针算术，从而读取错误内存位置，产生垃圾值和 NaN。PR body 明确指出此问题，并引用测试命令 #35925 验证修复后无损坏响应。

实现拆解

修改仅涉及一个文件 vllm/v1/attention/backends/rocm_aiter_fa.py。关键改动点：1. 在 cp_mha_gather_cache_kernel 函数签名中添加 6 个 stride 参数（k/v_cache_stride0/1/2）。2. 在内核指针计算中，用 stride 参数替换硬编码的 num_heads * head_size * PAGE_SIZE 等计算。3. 在调用函数 cp_mha_gather_cache 中，通过 key_cache.stride() 和 value_cache.stride() 获取实际步长并传递给内核。4. 在 do_kv_cache_update 函数中添加 TODO 注释，指出 shuffle 路径同样存在此问题需后续修复。

关键文件：

- vllm/v1/attention/backends/rocm_aiter_fa.py（模块 attention/backends）：唯一修改文件，包含 ROCM 后端注意力内核的关键修复，直接影响混合模型下的 KV 缓存内存访问正确性。

关键符号：cp_mha_gather_cache_kernel, cp_mha_gather_cache, do_kv_cache_update

评论区精华

review 中主要讨论点：1. gemini-code-assist[bot] 指出修复不完整，仅适用于 'NHD' 缓存格式，而 'SHUFFLE' 格式路径仍使用硬编码步长，可能导致非连续缓存时错误内存访问。2. yuankaichen-amd 询问测试模型和 stride 传递细节，作者回复测试了 qwen3.5，并解释传递所有 stride 是更标准的 Triton 内核写法，可预防未来其他维度非连续导致的静默错误。结论

：修复被接受，但 shuffle 路径问题被标记为 TODO 待后续处理。

- 修复不完整：shuffle 路径未处理 (correctness): 问题被确认，在 do_kv_cache_update 中添加 TODO 注释，待后续修复。
- stride 传递设计决策 (design): 采用传递所有 stride 的方案，增强代码健壮性。

风险与影响

- 风险：技术风险：1. 回归风险：修改涉及核心注意力内核的指针计算，若 stride 传递或使用错误，可能导致内存访问越界或数据损坏。2. 兼容性风险：仅修复了 'NHD' 格式路径，'SHUFFLE' 格式路径未修复，使用混合模型时若启用 shuffle 可能仍出错。3. 测试覆盖不足：PR body 提到测试了 #35925 命令，但未提及是否有自动化测试覆盖此场景，可能依赖手动验证。风险具体位置：rocm_aiter_fa.py 中的指针计算逻辑变更。
- 影响：影响范围：1. 用户：使用 ROCM 后端运行混合模型（如 Qwen3.5）的用户将修复注意力输出 NaN 问题，提升模型推理稳定性。2. 系统：仅影响 ROCM 后端的特定内核，对 CUDA 或其他后端无影响。3. 团队：揭示了 Triton 内核中硬编码内存假设的通用问题，可能促使其他类似内核的审查和修复。影响程度：中等，修复特定但关键的内存访问错误，避免混合模型推理失败。
- 风险标记：核心路径变更，部分路径未修复，缺少自动化测试

关联脉络

- PR #38766 未知标题：yuankaichen-amd 在评论中提及此 PR 用于修复 shuffle 路径，与本 PR 解决相同问题但针对不同路径，关联紧密。
- PR #35925 未知标题：PR body 中引用测试命令 #35925，可能为测试混合模型的相关 PR 或 issue，用于验证修复效果。
- PR #37940 [NIXL][BUG] Fix Triton heterogeneous TP: 同属注意力后端 bugfix，涉及 Triton 和 KV 缓存布局，技术领域相似，可参考其修复模式。