PR #38615 完整报告

vllm-project/vllm

[ROCm] Fix aiter persistent mode mla with q/o nhead<16 for kimi-k2.5 tp8

合并时间：2026-04-03 18:54

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38615

执行摘要

- 一句话：修复 ROCm Aiter MLA 后端在注意力头数小于 16 时元数据分配与内核输入形状不匹配的问题。
- 推荐动作：该 PR 值得精读，尤其关注头填充机制与元数据分配的一致性设计。对于 ROCm 平台开发者和多模态模型用户，可学习如何调试形状不匹配问题及利用 max 函数简化边界条件处理。

功能与动机

PR body 指出，AiterMLAImpl 已包含头填充机制：当 num_heads 为 4 或 8 时，q 张量会通过 repeat_interleave 填充到 16 个头再传递给内核。但 persistent mode MLA 实现为原始头数（如 8）分配缓冲区，而内核接收的是 16 个头的 q 张量，导致预分配的元数据缓冲区与实际内核输入之间的形状不匹配。这影响了 Kimi-K2.5 模型在 TP=8 下的正确运行。

实现拆解

仅修改了 vllm/v1/attention/backends/mla/rocm_aiter_mla.py 文件。关键改动是将 self._num_attention_heads 的计算从 vllm_config.model_config.get_num_attention_heads(vllm_config.parallel_config) 改为 max(16, self.num_heads)。这确保了元数据分配的头数至少为 16，与 AiterMLAImpl 中头填充逻辑（_needs_head_repeat 条件）保持一致，避免形状不匹配。

关键文件：

- vllm/v1/attention/backends/mla/rocm_aiter_mla.py（模块 attention/backends/mla）：唯一修改的文件，修复了 ROCm Aiter MLA 后端元数据分配逻辑，确保与内核输入形状一致。

关键符号：init

评论区精华

review 中 gemini-code-assist[bot] 建议简化逻辑，使用 max(16, self.num_heads) 来提高鲁棒性并与 AiterMLAImpl 实现保持一致。tjtanaa 指出修复与 _needs_head_repeat 条件相关，并验证了 gemini 反馈的有效性。wufann 接受了反馈并更新了代码。讨论焦点在于确保元数据分配与内核输入形状的一致性，无重大争议，结论明确采纳简化方案。

- 元数据分配头数简化 (correctness): 采纳建议，更新代码使用 max(16, self.num_heads)。

- 修复与 _needs_head_repeat 条件关联 (design): 确认修复正确性，关联代码逻辑。

风险与影响

- 风险：风险较低。变更仅影响 ROCm Aiter MLA 后端在注意力头数小于 16 的场景，修复了形状不匹配的 bug。潜在风险包括： 1) 对头数 >=16 的现有场景无影响，但需确保 max(16, self.num_heads) 不会意外改变其他配置； 2) 依赖 self.num_heads 的正确初始化，但该变量已在基类 MLACommonMetadataBuilder 中定义，风险可控。
  - 影响：影响范围有限但关键：修复了 Kimi-K2.5 等模型在 ROCm 平台 TP=8 下使用 Aiter MLA 后端时的运行错误，使 GSM8K 评估准确率从故障状态恢复至 93.4%。仅影响使用该特定后端且头数小于 16 的用户，对大多数其他配置无影响。有助于提升 ROCm 平台多模态模型支持的稳定性。
  - 风险标记：形状不匹配修复，平台特定逻辑

关联脉络

- PR #36205 [mla] Support fused FP8/NVFP4 output quantization in MLA attention (#35792): 同属 MLA 注意力相关优化，涉及 ROCm 平台和量化支持，技术上下文相关。
- PR #33657 [XPU] Initial support for GDN attention on Qwen3-next/Qwen3.5: 类似平台特定注意力后端支持，可对比不同硬件（XPU vs ROCm）的实现模式。