PR #51821 完整报告

vllm-project/vllm

[Bugfix][ROCm][CI] Restore the DeepSeek-V4 input GEMM override point

合并时间：2026-08-13 09:33

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/51821

执行摘要

- 一句话：恢复 _fused_wqa_wkv_gemm 覆写点，修复 ROCm 上 DeepSeek-V4 输出乱码
- 推荐动作：值得精读。这是一个小而典型的“性能优化删除公共覆写点导致平台特定数值回归”案例：8 行修复背后是 ROCm 权重 preshuffle 与 AITER GEMM 的深层契约。建议关注三点：一是平台覆写方法应被视为稳定接口，性能内联前需检查子类覆写；二是跨平台回归难以被 CUDA CI 发现，ROCm 专项测试需覆盖此类模型精度；三是 PR body 中“与 #51750/#51768 独立但需协调”的问题拆解方式，适合作为多平台回归处理模板。

功能与动机

PR body 指出 gfx950 上 GSM8K 任务得分为 0.0000，阈值是 0.92，且“服务器正常启动但每个问题都回答成一串不可解析的重复 token，说明是数值问题而非测试框架问题”。二分定位到 commit 79c865b（#51430）删除 _fused_wqa_wkv_gemm 并内联调用；ROCm 上 fused_wqa_wkv 权重在加载时原地 preshuffle 且 block scale 被单独保留，之后只能由 AITER 的 preshuffled-B GEMM 相乘，内联后普通 linear 在乱序权重上计算产生垃圾输出；同时删除使 ROCm override 的 fallback 调用的 super() 方法不复存在。

实现拆解

1. 根因定位：通过二分窗口定位到 #51430 的 commit 79c865b。该 commit 在收窄 DeepSeek-V4 eager CUDA graph 区域时，删除了 DeepseekV4Attention._fused_wqa_wkv_gemm 并在唯一调用点 _run_parallel_input_projections 中内联为 self.fused_wqa_wkv(hidden_states)[0]。
2. 恢复覆写点：在 vllm/models/deepseek_v4/attention.py 基类重新添加 _fused_wqa_wkv_gemm 方法，方法体是对 fused_wqa_wkv 的简单包装：取 MergedColumnParallelLinear 返回的 (output, bias) 中的 qr_kv（bias 为 None）。该方法正是 ROCm 注意力层覆写的入口，删除会同时导致 ROCm 侧 fallback 调用不存在的 super() 方法而抛 AttributeError。
3. 改回调用点：把 execute_in_parallel 中的 lambda 从 self.fused_wqa_wkv(hidden_states)[0] 改为 self._fused_wqa_wkv_gemm(hidden_states)，多流 GEMM 并行结构、ln_events 事件编排与 token 阈值开关保持不变，CUDA 上计算结果与内联版本逐位一致。
4. 验证与配套：在 gfx950（MI355X）上按 tests/evals/gsm8k/configs/DeepSeek-V4-Flash-NVFP4.yaml 配置跑 GSM8K 400 题，准确率由 0.0000 恢复至 0.953、invalid rate 由

1.000 降到 0.000；同时用 TP1 与 compilation mode 0 复现失败，证明与张量并行和 CUDA graph 捕获无关。本次未新增自动化测试，CI 通过 /ci run 触发多次全量检查。

关键文件：

- vllm/models/deepseek_v4/attention.py（模块 注意力层；类别 source；类型 core-logic；符号 _fused_wqa_wkv_gemm）：唯一变更文件，恢复被 #51430 删除的 _fused_wqa_wkv_gemm 覆写点，修复 ROCm 上 DeepSeek-V4 输入投影对 preshuffle 权重跑普通 GEMM 导致的完全乱码。

关键符号：_fused_wqa_wkv_gemm, _run_parallel_input_projections

关键源码片段

vllm/models/deepseek_v4/attention.py

唯一变更文件，恢复被 #51430 删除的 _fused_wqa_wkv_gemm 覆写点，修复 ROCm 上 DeepSeek-V4 输入投影对 preshuffle 权重跑普通 GEMM 导致的完全乱码。

```
    def _fused_wqa_wkv_gemm(self, hidden_states: torch.Tensor) -> torch.Tensor:
        # 覆写点：ROCm 注意力层在模型加载时会原地 preshuffle 该权重，
        # 之后只能由 AITER 的 preshuffled-B GEMM 正确相乘，普通 linear 会算错。
        # MergedColumnParallelLinear 返回 (output, bias)，bias 为 None。
        qr_kv, _ = self.fused_wqa_wkv(hidden_states)
        return qr_kv

    def _run_parallel_input_projections(
        self, hidden_states: torch.Tensor
    ) -> tuple[
        torch.Tensor,
        torch.Tensor | None,
        torch.Tensor | None,
        torch.Tensor | None,
    ]:
        # compressor 与 indexer 的辅助投影仍按原逻辑挂到 aux_fns 上，
        # 这里只展示本次改动涉及的调用点。
        qr_kv, (kv_score, indexer_weights, indexer_kv_score) = execute_in_parallel(
            # 关键改动：调用点重新经过 _fused_wqa_wkv_gemm 方法，
            # 让 CUDA 与 ROCm 各自绑定到正确的 GEMM 实现。
            lambda: self._fused_wqa_wkv_gemm(hidden_states),
            aux_fns,
            self.ln_events[0],
            self.ln_events[1:4],
            aux_streams,
            enable=hidden_states.shape[0]
            <= envs.VLLM_MULTI_STREAM_GEMM_TOKEN_THRESHOLD,
        )

        return qr_kv, kv_score, indexer_kv_score, indexer_weights
```

评论区精华

该 PR 来自 fork，claude[bot] 自动 review 被禁用，审阅者 zyongye 直接批准，没有留下逐行评论，因此没有实质性 review 交锋。最有价值的是 PR body 中对三个修复边界的论证，作者以此回应可能存在的"是否与 #51750/#51768 重复"的疑问：

> "This is not an alternative to either of them. On CUDA the restored method computes exactly what the inlined call computed, so this change is a no-op there and does nothing for the B200 failure — #51768 is still needed for that." "In the other direction, if #51750 lands after all, this one becomes redundant and I will close it." 这段论述明确了：ROCm 侧故障与 NVIDIA 侧 MRV1 PIECEWISE graph 故障是两个独立回归面，各自的修复互不替代但需要协调合并顺序。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险：改动仅 8 行且 CUDA 上为纯等价重构，但恢复的基类方法会被所有平台共用；若其他平台（如 CPU、XPU）未来在 fused_wqa_wkv 上做类似 preshuffle 或覆写，需要同步关注。
  2. 缺少自动化测试：本次无新增测试文件，修复依赖本地 gfx950 GSM8K 对比验证；若后续有重构再次删除或内联该方法，可能再次静默破坏 ROCm 路径。
  3. 与 #51750 冲突风险：若全量 revert #51430 的 #51750 先合并，本 PR 将成为冗余改动，合并顺序需要维护者协调。
  4. 数值路径隐蔽性：该故障只在 ROCm 权重 preshuffle 语义下出现，CUDA CI 全绿无法暴露，属于平台特有数值回归，依赖 ROCm 专项 CI 覆盖。 - 影响：影响范围集中在 ROCm 平台 + DeepSeek-V4（含 Flash-NVFP4 量化路径）推理正确性：修复前 gfx950 上 GSM8K 准确率为 0，修复后恢复至 0.953，达到 0.92 阈值；用户侧影响为 ROCm 集群上 DeepSeek-V4 生产可用性恢复。CUDA、CPU、XPU 等其他平台无行为变化，性能路径（多流并行、CUDA graph 收窄）保持 #51430 的原样。团队侧需要与 #51750/#51768 的合并顺序做一次协调，避免 revert 冲突或重复劳动。 - 风险标记：ROCm 平台特有回归，CUDA CI 无法覆盖，缺少自动化测试，与 #51750 全量 revert 存在合并互斥

关联脉络

- PR #51430 [Perf] Narrow DeepSeek V4 eager CUDA graph region: 本 PR 的根因：该 commit 删除 _fused_wqa_wkv_gemm 并内联调用，导致 ROCm 上 DeepSeek-V4 输入投影对 preshuffle 权重跑普通 GEMM，输出完全乱码。
- PR #51768 [Bugfix] Guard DeepSeek V4 MRV1 piecewise CUDA graphs: 同一回归在 NVIDIA 侧的独立修复，只针对 MRV1 + PIECEWISE CUDA graph 组合，与本 PR 互不替代；关联 Issue 表明其合并后 ROCm 仍需要本 PR。