

PR #49373 完整报告

vllm-project/vllm

[Bugfix][ROCm] Fix ROCM_AITER_FA & ROCM_AITER_UNIFIED_ATTN QK-Norm+RoPE+KVCache fusion for the packed KV-cache [BLOCKS, HEADS, BLOCK_SIZE, 2*HEAD_DIM] layout

合并时间: 2026-08-08 00:54

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49373>

执行摘要

- 一句话: 修复 ROCm AITER QK-Norm+RoPE+KV 融合在 packed KV 布局下的崩溃
- 推荐动作: 值得精读, 尤其是 `_split_kv_cache` 的提取方式和后端 `gate` 的对称性设计。该 PR 展示了如何在多后端代码中通过共享辅助方法消除布局假设耦合, 并利用 `gate` 避免不支持的 kernel 路径, 是 ROCm AITER 后端维护中典型的布局兼容修复。建议关注后续 AITER 版本升级时的回归测试。

功能与动机

PR body 明确指出: FA fused `do_qk_norm_rope_kvcache_update` 使用了旧的 `unbind(1)`, 在 packed 布局 (nb, nkvh, bs, 2*hs) 且 `num_kv_heads != 2` 时会出错; 同时 AITER 融合内核不支持 shuffle 模式下的 packed 布局, 需要 `gate`。作者还注明本 PR 依赖已合并的 AITER PR (ROCm/aiter#4312), 将在 AITER ~0.19 版本中落地。

实现拆解

1. 统一 KV 拆分逻辑 (vllm/v1/attention/backends/rocm_aiter_fa.py): 新增 `_split_kv_cache` 方法, 内部执行 `kv_cache.transpose(1, 2).split(self.head_size, dim=-1)`, 将 packed 布局 (B, H, N, 2*hs) 拆成 ((B, N, H, hs), (B, N, H, hs))。forward、`do_kv_cache_update`、`do_qk_norm_rope_kvcache_update`、`do_rope_and_kv_cache_update` 四处全部改为调用该方法, 替换原先的 `unbind(1)` 和重复的 `transpose/split` 内联代码。
2. 增加融合 `gate` (rocm_aiter_fa.py): `fused_qk_norm_rope_kvcache_supported` 从仅判断 `rocm_aiter_ops.is_enabled()` 改为同时要求 `not rocm_aiter_ops.is_shuffle_kv_cache_enabled()`, 与 `fused_rope_kvcache_supported` 保持一致, 防止 shuffle 模式下走不支持的融合路径。
3. 清理冗余注释 (vllm/v1/attention/backends/rocm_aiter_unified_attn.py): 删除 `do_qk_norm_rope_kvcache_update` 中关于 `_split_kv_cache` 选维度的过时注释, 逻辑本身无变化, Unified Attention 路径继续使用 `_split_kv_cache`。
4. 测试修正 (tests/compile/passes/test_rocm_aiter_qk_norm_rope_kvcache_fusion.py): `kv_stride_order` 参数从 (0,1,2,3,4) / (1,0,2,3,4) 两个 5D 排列改为单个 (0,1,2,3) (packed_4d), 匹配 4D packed KV 布局; 新增对 ROCM_AITER_FA + `use_shuffle_kv_layout=1` 组合的 `pytest.skip`, 因为该组合现在不会触发融合。

关键文件：

- vllm/v1/attention/backends/rocm_aiter_fa.py（模块 注意力后端；类别 source；类型 core-logic；符号 `_split_kv_cache`, `fused_qk_norm_rope_kvcache_supported`, `do_qk_norm_rope_kvcache_update`, `do_kv_cache_update`）：核心修复文件：新增 `_split_kv_cache` 统一 KV 拆分，替换错误的 `unbind(1)`，并给 `fused_qk_norm_rope_kvcache_supported` 增加 shuffle layout gate。
- vllm/v1/attention/backends/rocm_aiter_unified_attn.py（模块 注意力后端；类别 source；类型 core-logic；符号 `do_qk_norm_rope_kvcache_update`）：同步清理 `do_qk_norm_rope_kvcache_update` 中的过时注释，确认 unified 路径与 FA 路径共用 `_split_kv_cache`，保持两个后端行为一致。
- tests/compile/passes/test_rocm_aiter_qk_norm_rope_kvcache_fusion.py（模块 融合测试；类别 test；类型 test-coverage；符号 `test_qk_norm_rope_kvcache_fusion`）：修正测试参数以匹配 packed 4D 布局，并新增 FA+shuffle 组合的 skip 逻辑，防止无效测试组合产生误导性失败。

关键符号：`_split_kv_cache`, `fused_qk_norm_rope_kvcache_supported`,
`do_qk_norm_rope_kvcache_update`, `do_kv_cache_update`,
`do_rope_and_kv_cache_update`, `test_qk_norm_rope_kvcache_fusion`

关键源码片段

vllm/v1/attention/backends/rocm_aiter_fa.py

核心修复文件：新增 `_split_kv_cache` 统一 KV 拆分，替换错误的 `unbind(1)`，并给 `fused_qk_norm_rope_kvcache_supported` 增加 shuffle layout gate。

```
# vllm/v1/attention/backends/rocm_aiter_fa.py
```

```
def _split_kv_cache(
    self, kv_cache: torch.Tensor
) -> tuple[torch.Tensor, torch.Tensor]:
    # packed KV 布局为 (num_blocks, num_kv_heads, block_size, 2*head_size),
    # 先交换 heads 与 block 维度得到 (B, H, N, 2*hs),
    # 再沿最后一维按 head_size 拆成 K 与 V 两个独立 cache。
    # 注意：不能用 unbind(1)，因为 num_kv_heads != 2 时该维度不是 K/V 通道。
    return kv_cache.transpose(1, 2).split(self.head_size, dim=-1)

def fused_qk_norm_rope_kvcache_supported(self):
    # 只有当 AITER 启用且未使用 shuffle KV 布局时才允许融合；
    # shuffle 模式下写入走 reshape_and_cache_shuffle_triton 专用路径，
    # 融合内核不支持 packed+shuffle 组合，因此必须 gate 掉，
    # 这与 fused_rope_kvcache_supported 的判定保持一致。
    return (
        rocm_aiter_ops.is_enabled()
        and not rocm_aiter_ops.is_shuffle_kv_cache_enabled()
    )

def do_qk_norm_rope_kvcache_update(
```

```

self,
layer: AttentionLayer,
qkv: torch.Tensor,
q_out: torch.Tensor,
k_out: torch.Tensor,
positions: torch.Tensor,
q_weight: torch.Tensor,
k_weight: torch.Tensor,
rms_norm_eps: float,
cos_sin_cache: torch.Tensor,
is_neox: bool,
kv_cache: torch.Tensor,
layer_slot_mapping: torch.Tensor,
):
    # 原实现使用 kv_cache.unbind(1), 在 packed 布局且 num_kv_heads != 2 时
    # 会把 block 维当作 K/V 通道, 导致读写错乱; 统一改用 _split_kv_cache。
    key_cache, value_cache = self._split_kv_cache(kv_cache)
    rocm_aiter_ops.do_qk_norm_rope_kvcache_update(
        qkv=qkv,
        q_weight=q_weight,
        k_weight=k_weight,
        cos_sin_cache=cos_sin_cache,
        positions=positions,
        num_heads_q=self.num_heads,
        num_heads_k=self.num_kv_heads,
        head_dim=self.head_size,
        is_neox=is_neox,
        rms_norm_eps=rms_norm_eps,
        q_out=q_out,
        k_out=k_out,
        key_cache=key_cache,
        value_cache=value_cache,
        slot_mapping=layer_slot_mapping,
        k_scale=layer._k_scale_cpu,
        v_scale=layer._v_scale_cpu,
        kv_cache_dtype=self.kv_cache_dtype,
        use_shuffle_layout=rocm_aiter_ops.is_shuffle_kv_cache_enabled(),
    )

```

评论区精华

Review 由 claude[bot] 自动评论 (fork PR 默认不自动审查), AndreasKaratzas 和 mgoin 均 APPROVED, AndreasKaratzas 给出 "LGTM"。PR 内主要讨论集中在作者对依赖的说明: jhu960213 在 Issue 评论中明确本 PR 依赖 ROCm/aiter#4312, 预计在 AITER ~0.19 版本发布后可用, 这是测试通过的前提。

- AITER 版本依赖说明 (other): 测试需在 AITER 0.19+ 环境运行, 属于外部依赖前置条件。

- FA+shuffle 组合跳过测试 (testing): 接受 skip, 避免无效测试组合; shuffle 路径的融合覆盖留待内核支持后补充。

风险与影响

- 风险:
 1. 依赖风险: 修复依赖 AITER 0.19 (ROCm/aiter#4312) 中的融合内核行为, 若用户使用旧版 AITER, 融合路径可能仍存在布局兼容问题, 需等待 pin 版本更新。
 2. 行为回归风险: fused_qk_norm_rope_kvcache_supported 在 shuffle 模式下从启用变为禁用, FA 路径会回退到 reshape_and_cache_shuffle_triton 等旧路径, 可能在部分场景引入轻微性能回退, 但功能不受影响。
 3. 测试覆盖缺口: 测试将 kv_stride_order 从 5D 两个排列收敛为 4D 单一排列, 并跳过 FA+shuffle 组合, 意味着 shuffle 布局下的融合路径和旧 K/V-first 排列不再有回归保护, 后续若这些路径被重新启用需补测。
 4. 一致性风险: Unified Attention 路径继续硬编码 use_shuffle_layout=False 且未加 gate, 若未来统一后端的 NHD 假设被打破, 该路径可能引入同类布局错误。- 影响: 影响范围限于 ROCm + AITER 后端的 v1 注意力路径, 具体涉及 ROCM_AITER_FA 与 ROCM_AITER_UNIFIED_ATT_N 两个注意力后端。修复解决了 GQA/MQA 模型 (num_kv_heads != 2) 在 packed KV 布局下的潜在错误 KV 读写, 属于正确性修复; 对使用 shuffle KV layout 的 FA 用户会有融合降级 (功能不变)。由于改动集中在单一后端且门控清晰, 对非 ROCm 平台和其他注意力后端无影响。团队维护者需确保 AITER 版本同步更新以获得完整修复效果。- 风险标记: 外部依赖版本, 行为回退, 测试覆盖缺口, 平台限定

关联脉络

- PR #50930 [Test] Add ROCm AITER MLA op registration and env gating tests: 同为 ROCm AITER 注意力算子的注册与门控测试, 共享 AITER 内核环境和 CI 配置, 后续需同步验证本次融合 gate 的测试覆盖。
- PR #51402 [ROCm][CI][Bugfix] Do not microbatch a step that splits a prefix from its writer: 同为 ROCm v1 注意力路径的 CI 偶发失败修复, 与本次 FA 后端 gate 调整都属于 ROCm 注意力稳定性的同类演进。
- PR #51357 Fix ROCm architecture import on non-ROCm platforms: 修复 ROCm 模块在非 ROCm 平台的导入问题, 与本次 ROCm 后端注意力路径修复同属 ROCm 代码健康度维护。