PR #38152 完整报告

vllm-project/vllm

Disable dual stream execution of input projection for Qwen3

合并时间：2026-03-26 09:20

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38152

执行摘要

- 一句话：为 Qwen3 模型禁用输入投影的双流执行，修复冷编译时间约 4 倍的回归。
- 推荐动作：建议技术管理者关注此 PR，因为它展示了性能优化与编译时间的权衡，以及临时回退的策略。工程师可学习如何安全地移除自定义操作以避免编译回归。

功能与动机

根据 PR body，当前双流执行需要自定义操作传递层名作为字符串，导致冷编译时间回归约 4 倍，因此暂时禁用此优化以恢复编译性能。

实现拆解

修改了两个模型文件：qwen3_5.py 和 qwen3_next.py。移除了对 torch.ops.vllm.gdn_in_proj 的调用，代之以直接调用 self.in_proj_qkvz 和 self.in_proj_ba 方法。同时删除了辅助函数 _forward_in_proj、maybe_execute_in_parallel 的使用以及相关事件和流对象。

关键文件：

- vllm/model_executor/models/qwen3_5.py（模块 model_executor/models）：修改 Qwen3.5 模型的输入投影逻辑，移除双流执行代码，直接调用 in_proj_qkvz 和 in_proj_ba。
- vllm/model_executor/models/qwen3_next.py（模块 model_executor/models）：修改 Qwen3 Next 模型的输入投影逻辑，移除自定义操作 gdn_in_proj、并行执行辅助函数和相关事件对象。

关键符号：forward (in qwen3_next.py), gdn_in_proj (removed), _forward_in_proj (removed)

评论区精华

review 讨论有限，主要来自自动化工具；zou3519 批准并道歉，表示理解临时回退的麻烦。没有深入的技术争议或未解决疑虑。

- 临时回退双流优化 (design): PR 被批准合并，作为临时解决方案。

风险与影响

- 风险：风险包括：1) 性能下降：移除双流优化可能降低推理时的并行执行效率；2) 兼容性：代码修改可能影响其他依赖自定义操作的组件；3) 临时解决方案：回退是暂时的，未来需重新启用，可能引入维护复杂性。具体到文件，修改了模型核心逻辑，需确保正确性。
- 影响：影响范围局限于 Qwen3 和 Qwen3.5 模型。用户可能注意到冷编译时间改善，但推理性能可能略有下降。对系统来说，减少了自定义操作依赖，简化了代码路径。团队需关注后续 #38123 的进展以重新启用优化。
- 风险标记：性能回归风险，临时解决方案，核心路径变更

关联脉络

- PR #38046 [compile] Add some more startup tests for top models: 与此 PR 相关，因为都涉及 torch.compile 和编译时间优化测试。