PR #49292 完整报告

vllm-project/vllm

Fix Qwen3-VL M-RoPE on the Transformers modeling backend (grids + compile)

合并时间：2026-07-22 02:28

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49292

执行摘要

- 一句话：修复 Transformers 后端 Qwen3-VL M-RoPE 的两个 bug
- 推荐动作：值得阅读。虽然改动量小，但揭示了 torch.compile 模式下一个容易被忽略的步幅陷阱，并提供了清晰的解决方案。

功能与动机

用户尝试使用 Qwen/Qwen3-VL-4B-Instruct-FP8 模型时，在 *eager* 和 *compile* 两种模式下均遇到失败。Eager 模式下，空 video_grid_thw 传为空张量触发了 transformers 的 None 检查；编译模式下，*unsqueeze* 操作保留了原步幅信息，导致 torch.compile 动态符号推导错误。

实现拆解

1. 修复空 video_grid_thw 传值问题（get_mrope_input_positions 方法）：将 torch.tensor([]) 改为 None，匹配 transformers 对空输入的预期。
2. 修复编译模式下的步幅问题（forward 方法）：在 positions[:, None] 后增加 .contiguous()，强制生成新的连续张量，消除陈旧步幅信息。

关键文件：

- vllm/model_executor/models/transformers/multimodal.py（模块 多模态后端；类别 source；类型 data-contract）：修复了两个关键数据契约问题：空 video_grid_thw 传递方式、M-RoPE positions 张量连续性，直接影响 Qwen3-VL 在 Transformers 后端上的 *eager* 和 *compile* 推理。

关键符号：forward, get_mrope_input_positions

评论区精华

PR 无 review 评论，仅由 hmellor 批准合并。提交历史显示作者曾尝试使用 .as_strided() 等方式避免内存拷贝，最终回归 .contiguous() 方案。

- 暂无高价值评论线程

风险与影响

- 风险：改动仅 3 行源码，风险极低。.contiguous() 可能引入微小内存拷贝开销，但对推理性能影响可忽略。空张量改 None 仅影响空视频输入分支，不影响正常流程。

- 影响：影响范围：仅 Transformers 建模后端 + Qwen3-VL 模型（使用 M-RoPE 且可能无视频输入）。解决了 Qwen3-VL 模型在该后端上的可用性问题，使 *eager* 和 *compile* 两种模式均能正常运行。
- 风险标记：暂无

关联脉络

- PR #12128 MRoPE fix: PR body 引用 #12128 中关于添加 buffer 到 MRoPE positions 以避免非连续张量的讨论，这是本 PR 编译修复的背景参考。