PR #38927 完整报告

vllm-project/vllm

[Bugfix][LoRA] Fix missing in_proj_z in Qwen3_5ForConditionalGenerati···

合并时间：2026-04-04 07:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38927

执行摘要

- 一句话：修复 Qwen3.5 模型在 LoRA 启用且 TP>1 时加载适配器报 IndexError 的 bug。
- 推荐动作：该 PR 值得快速浏览以理解 Qwen3.5 模型在 LoRA 下的模块映射机制。关注点：1) 了解 GDN 层在 LoRA 启用时从合并投影到分离投影的转换逻辑；2) 注意 update_packed_mapping 方法在模型初始化中的作用；3) 可结合 PR #36069 和 #36603 了解问题的演进和 workaround 方案。

功能与动机

修复 Qwen3.5 模型在启用 LoRA 且张量并行 (TP>1) 时加载适配器会抛出 IndexError: list index out of range 的问题。PR body 明确指出：当 LoRA 启用时，GDN 层使用分离的 in_proj_qkv 和 in_proj_z 投影而不是合并的 in_proj_qkvz，但 update_packed_mapping 方法只正确移除了 in_proj_qkvz 并添加了 in_proj_qkv，却遗漏了 in_proj_z 条目。这导致 slice_lora_b 在 TP>1 时引发索引错误。

实现拆解

在 vllm/model_executor/models/qwen3_5.py 文件的 update_packed_mapping 方法中添加一行代码：self.packed_modules_mapping["in_proj_z"] = ["in_proj_z"]。该方法在 LoRA 启用时负责更新模型模块映射，原本已正确处理了 in_proj_qkvz 到 in_proj_qkv 的转换，但遗漏了对应的 in_proj_z 映射。

关键文件：

- vllm/model_executor/models/qwen3_5.py（模块 model_executor/models）：唯一修改的文件，包含 Qwen3.5 模型实现，修复了 update_packed_mapping 方法中 LoRA 启用时的模块映射缺失问题。

关键符号：update_packed_mapping

评论区精华

review 讨论较少，主要关注点：

1) vadiklyutiy 要求修复 DCO 签名问题，作者 elenalil-aws 随后执行了修复；2) gemini-code-assist[bot] 的自动评论指出没有 review 评论可评估。没有出现技术争议或设计权衡讨论，因为这是一个明确的单行 bug 修复。

- DCO 签名修复 (other): DCO 问题已解决，PR 被合并。

风险与影响

- 风险：风险极低： 1) 变更仅添加一行映射代码，不涉及核心逻辑修改； 2) 修复针对特定条件（LoRA 启用且 TP>1）下的错误，不会影响其他场景； 3) 已有 PR #36069 通过边界检查缓解了症状，但本 PR 修复了根本原因，因此风险可控。潜在风险：如果其他模型有类似映射问题可能未被发现，但本变更范围明确限定于 Qwen3.5 模型。
- 影响：影响范围： 1) 用户：修复了 Qwen3.5 模型在 LoRA 启用且 TP>1 时无法加载适配器的问题，提升了模型兼容性和用户体验； 2) 系统：确保 GDN 层在 LoRA 下的正确投影映射，避免运行时异常； 3) 团队：解决了之前通过 workaround（PR #36069）缓解但未根治的问题，减少了技术债务。影响程度：中等，针对特定模型和配置，但解决了关键功能阻塞。
  - 风险标记：特定配置触发，映射一致性风险

关联脉络

- PR #36069 [Bugfix][LoRA] Fix slice_lora_b out of range error: 通过边界检查缓解了同一问题的症状，但未修复根本原因，与本 PR 直接相关。
- PR #36603 [LoRA] Add support for Qwen3.5 MoE: 评论中识别了此根本原因但未修复，与本 PR 问题相同。
- PR #37114 [LoRA] Fix expert base_layer loading: 处理不同的 LoRA 问题（专家 base_layer 加载），但同属 LoRA 相关修复。