PR #42671 完整报告

vllm-project/vllm

fix: use keyword arguments for shard_id and expert_id in weight_loade···

合并时间：2026-05-19 13:29

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42671

执行摘要

该 PR 修复了 Qwen3-VL-235B-A22B-Instruct 模型在 TPU 推理时因权重加载函数调用使用位置参数导致的崩溃问题，将 load_fused_expert_weights 的参数改为关键字参数。

功能与动机

PR 正文指出，当前 qwen3_moe.py 中 load_fused_expert_weights 调用使用了位置参数传递 shard_id 和 expert_id，但新接口要求它们作为关键字参数，这在 Qwen3-VL 大模型上导致 TPU 推理崩溃。修复旨在恢复该模型的正常推理。

实现拆解

1. 定位问题行：在 vllm/model_executor/models/qwen3_moe.py 第 645 行。
2. 修改调用方式：将 load_fused_expert_weights(..., shard_id, expert_id) 改为 load_fused_expert_weights(..., shard_id=..., expert_id=...)。
3. 提交与合并：由作者 junyanxu 提交单个 commit，经审核后合并。

由于材料中未提供具体源码片段，此处仅描述：变更位于 qwen3_moe.py 文件的 load_fused_expert_weights 调用行，将位置参数替换为关键字参数。

评论区精华

无实质讨论，自动机器人评论未提出具体建议。

风险与影响

- 风险：极低。仅修改一行代码，不改变逻辑。
- 影响：仅影响 Qwen3-VL 模型在 TPU 上的推理，其他模型或平台无影响。

关联脉络

此 PR 与近期模型重构 PR（如 DeepSeek V4 迁移 #43039）可能相关，均涉及权重加载接口的一致性维护。