

PR #48942 完整报告

vllm-project/vllm

[XPU] Bump vllm_xpu_kernels to v0.1.11.1

合并时间: 2026-07-17 20:43

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48942>

执行摘要

- 一句话: 升级 XPU 内核版本修复 MoE 崩溃
- 推荐动作: 变更简单明确, 直接合入即可。建议后续关注 XPU 内核版本与 CUDA 端的 API 同步流程, 避免类似参数不匹配问题。

功能与动机

XPU 环境下 MoE 前向传播因参数不匹配崩溃: `_moe_C::moe_sum() expected at most 2 argument(s) but received 4 argument(s)`。根本原因是 CUDA 端 commit f7aadae5e5 为 `moe_sum` 添加了 `topk_ids` 和 `expert_map` 两个参数, 但 XPU 端内核注册未跟进。

实现拆解

1. 修改依赖版本号: 在 `requirements/xpu.txt` 中将 `vllm_xpu_kernels` 的下载链接从 `v0.1.11` 改为 `v0.1.11.1`。
2. 验证修复: 开发者在 B70 硬件上重现了崩溃后, 安装 `v0.1.11.1` 确认 4 参数调用成功且结果与 `torch.sum` 参考值比特精确。

关键文件:

- `requirements/xpu.txt` (模块 依赖配置; 类别 docs; 类型 dependency-update): 唯一变更文件, 将 XPU 内核依赖从 `v0.1.11` 升级到 `v0.1.11.1`, 修复关键崩溃。

关键符号: 未识别

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 回归风险低: 仅更新版本号, 未改动其他代码。XPU 内核版本兼容性良好, 新版本包含明确的 bugfix (PR#462)。
- 影响: 影响范围小: 仅影响使用 XPU + MoE 模型的用户, 修复了此类场景下的完全崩溃问题, 对其他平台无影响。

- 风险标记：依赖同步风险

关联脉络

- PR #47516 [XPU][UT]fix _POSSIBLE_KERNELS error on XPU: 同为 XPU 平台内核相关 bugfix，说明 XPU 后端持续修复中。