PR #41019 完整报告

vllm-project/vllm

[xpu] bump up vllm-xpu-kernel v0.1.7

合并时间：2026-04-28 08:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41019

执行摘要

- 一句话：升级 XPU 内核至 v0.1.7 并移除挂起测试
- 推荐动作：变更简单、安全，可直接合并。建议关注后续 CI 中 lora 相关测试的状态，并跟进 test_qwenvl.py hang 问题的根因修复。

功能与动机

升级内核库以获取 bug 修复和新特性。同时，作者在评论中指出 test_qwenvl.py 在 XPU 上 hang 了，需要移除以保持 CI 稳定。

实现拆解

1. 升级依赖版本：在 requirements/xpu.txt 中将 vllm_xpu_kernels 的版本从 v0.1.5 更新到 v0.1.7。
2. 移除挂起测试：在 .buildkite/intel_jobs/lora_intel.yaml 中删除 LoRA 多模态测试步骤中的 (pytest -v -s lora/test_qwenvl.py || true) 一行。

关键文件：

- requirements/xpu.txt（模块 依赖配置；类别 docs；类型 documentation）：升级 vllm-xpu-kernels 从 0.1.5 到 0.1.7，是本次 PR 的主要变更。
- .buildkite/intel_jobs/lora_intel.yaml（模块 CI 配置；类别 config；类型 configuration）：移除导致 hang 的 LoRA 测试 test_qwenvl.py，保持 CI 稳定。

关键符号：未识别

关键源码片段

.buildkite/intel_jobs/lora_intel.yaml

移除导致 hang 的 LoRA 测试 test_qwenvl.py，保持 CI 稳定。

```
# .buildkite/intel_jobs/lora_intel.yaml 中的 commands 片段（已简化）
commands:
  - >-
    bash .buildkite/scripts/hardware_ci/run-intel-test.sh
    'cd tests &&
    pytest -v -s lora/test_default_mm_loras.py &&
```

```
# 该测试在 XPU 上 hang，已临时移除
# (pytest -v -s lora/test_qwenvl.py || true) &&
pytest -v -s lora/test_whisper.py'
```

评论区精华

作者 @jikunshang 在 PR 评论中明确指出已移除 CI 中挂起的 lora 测试（test_qwenvl.py），并请同事后续复现排查。没有其他 reviewer 讨论。

- 移除挂起的 LoRA 测试 (other): 临时移除 test_qwenvl.py，待后续排查修复。

风险与影响

- 风险：风险极低。变更仅涉及依赖版本号和 CI 配置删除一行测试调用。内核版本升级可能引入不兼容，但该库是 vllm 的 XPU 专用内核包，通常向后兼容。移除的测试是可选执行（带 || true），不影响其他测试。
- 影响：影响范围限于 Intel GPU 平台。升级内核库可能修复 XPU 上的底层 bug 或提升性能，但需关注回归。CI 中 lora 多模态测试覆盖略有减少。
- 风险标记：测试覆盖减少

关联脉络

- PR #39801 [ROCm][CI] Add missing quantization methods and fix online quant test failures: 同样属于平台相关的 CI 修复，通过移除 / 调整测试来保持 CI 稳定。