

PR #49911 完整报告

vllm-project/vllm

[CI][ROCm] Keep global GPU memory cleanup opt-in

合并时间: 2026-07-27 11:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49911>

执行摘要

- 一句话: AMD CI 全局 GPU 内存清理改为 opt-in
- 推荐动作: 建议 AMD CI 维护人员确认相关测试是否已正确设置 `VLLM_TEST_CLEAN_GPU_MEMORY=1`。此 PR 提供了合理的设计决策, 值得阅读。

功能与动机

PR body 指出: 函数作用域的 fixture 无法释放存活时间更长的 fixture 持有的内存, 自动清理可能导致误判或 OOM。同时引用了 Buildkite CI 构建的失败日志作为参考。

实现拆解

1. 修改 `_should_clean_gpu_memory_between_tests()` 函数(`tests/conftest.py`): 移除原有的三路分支, 直接返回 `os.getenv("VLLM_TEST_CLEAN_GPU_MEMORY", "0") == "1"`。
2. 删除默认 ROCm 平台自动清理逻辑: 原代码中 `current_platform.is_rocm()` 的检查被移除, 不再为 AMD CI 自动启用清理。
3. 保留有针对性的 teardown: 保留从 PR #49242 引入的 Qwen-VL 和 Nixl 相关特设清理, 不受到影响。

关键文件:

- `tests/conftest.py` (模块 测试配置; 类别 test; 类型 test-coverage; 符号 `_should_clean_gpu_memory_between_tests`): 核心变更文件, 修改了 GPU 内存清理的决策逻辑, 将 ROCm 平台默认清理改为 opt-in。

关键符号: `_should_clean_gpu_memory_between_tests`

评论区精华

无 review 讨论。Claude bot 自动评论提示 review 未运行。dllehr-amd 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险: 风险很低。变更仅限于测试基础设施的配置逻辑, 不涉及核心推理、调度或模型代码。如果 AMD CI 中某些测试依赖自动清理来避免 OOM, 则需要显式设置环境变量, 否则可能

引入测试失败，但 PR 明确为 opt-in 且已有针对性的 teardown。

- 影响：直接影响 AMD CI 上的测试环境：原本自动执行的 GPU 内存清理现在需要手动设置环境变量。影响范围小，仅限测试配置。非 ROCm 平台不受影响。
- 风险标记：测试基础设施变更，需确认 CI 设置

关联脉络

- PR #49242 [ROCm] Some PR that introduced targeted teardown for Qwen-VL and Nixl: PR body 提到保留来自 #49242 的针对性 teardown，确保不会重复或冲突。