PR #39411 完整报告

vllm-project/vllm

[CI/Build] Fix memory cleanup in MM test

合并时间：2026-04-09 23:50

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39411

执行摘要

该 PR 通过调整 CI 测试配置和重命名测试函数，修复了多模态内存泄漏测试导致的 CI 失败，但 review 中揭示的进程清理缺陷未被解决，可能影响未来测试稳定性。

功能与动机

PR 旨在解决 CI 构建失败问题，引用失败链接（https://buildkite.com/vllm/ci/builds/60525/steps/canvas?jid=019d71f1-593e-4e3c-bbb7-9a9c25f2e273）。从 review 评论推断，根本原因是内存泄漏测试使用的 @create_new_process_for_each_test 装饰器存在进程清理缺陷，导致测试环境不稳定。

实现拆解

实现包括两个关键改动：

1. CI 配置文件调整：在 .buildkite/test_areas/models_multimodal.yaml 中，修改 pytest 命令，增加 --ignore models/multimodal/generation/test_memory_leak.py，将内存泄漏测试从常规测试套件中排除。
2. 测试文件修改：在 tests/models/multimodal/generation/test_memory_leak.py 中，将测试函数名从 test_qwen3_vl_no_memory_leak 重命名为 test_no_memory_leak，提升通用性。

评论区精华

review 中仅 gemini-code-assist[bot] 提出关键评论：

> "The @create_new_process_for_each_test decorator introduces a critical issue... The decorator currently calls os.setpgrp() in the parent process (the pytest worker). This makes the worker process itself a process group leader... This is a fragile design that can lead to silent test failures or unexpected..."

评论指出装饰器实现缺陷可能导致 pytest worker 被意外终止，但 PR 作者未回复，问题未被解决。

风险与影响

- 测试覆盖风险：通过 --ignore 排除内存泄漏测试，可能削弱多模态模型的内存泄漏检测能力。

- 进程清理缺陷：review 中提到的装饰器问题未被修复，未来可能重现 CI 失败或导致测试不稳定。
- 影响范围：主要影响 CI 系统和开发团队，修复当前失败提升构建成功率，但潜在缺陷需后续关注。

关联脉络

- 与 PR #39268（新增 Qwen3-VL 多模态内存泄漏检测测试）直接相关，本 PR 可能针对该测试引入的 CI 问题进行调整。
- 近期历史 PR 中多涉及多模态（如 #39409、#39268）和 CI 修复（如 #39421、#39390），反映团队在加强多模态功能测试的同时，持续优化 CI 稳定性。