PR #39555 完整报告

vllm-project/vllm

[ROCm][CI/Build] Fix memory cleanup in MM test

合并时间：2026-04-12 11:13

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39555

执行摘要

本 PR 修复了 ROCm CI 配置中多模态内存泄漏测试的执行问题，通过将 test_memory_leak.py 从忽略列表移出并作为独立命令运行，确保测试能够正确执行。这是对 PR #39411 的后续修复，主要影响 AMD 平台的 CI 测试稳定性，对生产代码无直接影响。

功能与动机

此变更旨在解决多模态测试中的内存清理问题。PR body 中明确说明这是 PR #39411 的后续修复（Parity PR），review 评论进一步指出原配置中 test_memory_leak.py 被错误忽略，且新增的测试命令缺少必要配置，可能导致 ROCm 平台测试不稳定。

实现拆解

仅修改了 .buildkite/test-amd.yaml 文件，在三个测试步骤中进行了相同调整：

1. 在原有 pytest 命令的 --ignore 列表中新增 test_memory_leak.py
2. 新增独立命令运行该测试文件

关键变更示例（以第一个步骤为例）：

```
- pytest -v -s models/multimodal -m core_model \
  --ignore models/multimodal/generation/test_common.py \
  ... \
  --ignore models/multimodal/generation/test_memory_leak.py \
  --ignore models/multimodal/processing
- pytest -v -s models/multimodal/generation/test_memory_leak.py -m core_model
```

评论区精华

review 中 gemini-code-assist[bot] 指出了三个关键问题：

> "The new pytest command for test_memory_leak.py is missing the -v and -s flags... multi-modal tests should use VLLM_WORKER_MULTIPROC_METHOD=spawn to ensure stability and avoid deadlocks."

这揭示了 ROCm 平台上多模态测试的特殊要求：必须使用 spawn 多进程方法避免死锁。提交者在后续提交中修复了这些问题。

风险与影响

风险：

- 若未正确添加 VLLM_WORKER_MULTIPROC_METHOD=spawn 环境变量，可能导致 ROCm 测试死锁
- 配置变更可能影响其他测试步骤，但变更范围小，风险可控

影响：

- 仅影响 ROCm CI 测试执行，提升多模态测试覆盖率和稳定性
- 对用户功能和系统性能无直接影响
- 有助于及早发现内存泄漏问题

关联脉络

这是 PR #39411 的后续修复，两者共同解决多模态测试的内存清理问题。从近期历史 PR 看，仓库持续关注多模态测试的稳定性（如 PR #39344、#39526 修复多模态模型问题），本 PR 是这一趋势在 CI 层面的体现。ROCm 平台的测试配置特殊要求（spawn 方法）在其他多模态相关 PR 中也有体现，形成了跨平台测试的一致模式。