

PR #49770 完整报告

vllm-project/vllm

[CI] fix compile test | refactor VLLM_DISABLE_COMPILE_CACHE for tests

合并时间: 2026-07-25 12:15

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49770>

执行摘要

- 一句话: 重构测试编译缓存禁用方式, 修复 CI 不稳定
- 推荐动作: 值得快速合并。该 PR 解决了实际 CI 问题, 且重构后代码更清晰、更 DRY。建议团队在后续类似测试中优先使用 `disable_vllm_compile_cache` fixture。

功能与动机

修复 `tests/compile/test_compile_ranges.py` 在持久化 CI Agent 上的间歇性失败。根本原因是 warm vLLM compile cache 导致已编译图从磁盘加载, 从而跳过了自定义 post-grad passes, 使得测试无法正确计数。

实现拆解

1. 新增 fixture `disable_vllm_compile_cache` (`tests/conftest.py`): 组合 `use_fresh_inductor_cache` 和 `monkeypatch.setenv('VLLM_DISABLE_COMPILE_CACHE', '1')`, 一次性确保每次测试都使用全新 Inductor 缓存并禁止 vLLM 级联缓存, 强制编译过程实际执行。
2. 迁移 `test_compile_ranges.py` 的三个测试: 将 `test_compile_ranges`、`test_compile_sizes_produce_static_shapes`、`test_inductor_cache_compile_ranges` 的 fixture 从 `use_fresh_inductor_cache` 改为 `disable_vllm_compile_cache`, 并移除 `test_inductor_cache_compile_ranges` 中手动的 `monkeypatch.setenv('VLLM_DISABLE_COMPILE_CACHE', '1')`。
3. 迁移 `test_fusion_attn.py`: 将 `test_attention_quant_pattern` 的参数从 `monkeypatch, use_fresh_inductor_cache` 替换为 `disable_vllm_compile_cache`, 并删除函数体内的手动 `monkeypatch.setenv('VLLM_DISABLE_COMPILE_CACHE', '1')`。
4. 迁移 `test_mla_attn_quant_fusion.py`: 类似地, 将 `test_mla_attention_quant_pattern` 的参数从 `monkeypatch, use_fresh_inductor_cache` 替换为 `disable_vllm_compile_cache`, 并删除手动环境变量设置。

关键文件:

- `tests/conftest.py` (模块 测试配置; 类别 test; 类型 test-coverage; 符号 `disable_vllm_compile_cache`): 新增核心 fixture `disable_vllm_compile_cache`, 作为本次重构的基础设施, 供其他测试复用。

- tests/compile/test_compile_ranges.py (模块 编译范围; 类别 test; 类型 test-coverage; 符号 test_compile_ranges, test_compile_sizes_produce_static_shapes, test_inductor_cache_compile_ranges) : 主要修复对象: 三个测试因 warm compile cache 间歇性失败, 通过 fixture 替换修复。
- tests/compile/passes/test_fusion_attn.py (模块 融合注意力; 类别 test; 类型 test-coverage) : 迁移注意力融合测试, 移除重复的环境变量设置, 使用新 fixture 简化代码。
- tests/compile/passes/test_mla_attn_quant_fusion.py (模块 MLA 量化融合; 类别 test; 类型 test-coverage) : 迁移 MLA 注意力融合测试, 与 test_fusion_attn.py 类似的简化。

关键符号: disable_vllm_compile_cache, test_compile_ranges, test_compile_sizes_produce_static_shapes, test_inductor_cache_compile_ranges, test_attention_quant_pattern, test_mla_attention_quant_pattern

关键源码片段

tests/compile/test_compile_ranges.py

主要修复对象: 三个测试因 warm compile cache 间歇性失败, 通过 fixture 替换修复。

```
# Before:
# def test_compile_ranges(use_fresh_inductor_cache):
# ...
# def test_inductor_cache_compile_ranges(monkeypatch, use_fresh_inductor_cache):
# monkeypatch.setenv("VLLM_DISABLE_COMPILE_CACHE", "1")

# After: 统一使用 disable_vllm_compile_cache fixture
def test_compile_ranges(disable_vllm_compile_cache):
    # ... 测试使用 post_grad_range_checker 验证编译范围
    pass

def test_compile_sizes_produce_static_shapes(disable_vllm_compile_cache):
    """Verify that compile_sizes entries are compiled with fully concrete
    shapes (no SymInts), while compile_ranges entries retain dynamic shapes."""
    # ...
    pass

def test_inductor_cache_compile_ranges(disable_vllm_compile_cache):
    # disable_vllm_compile_cache sets VLLM_DISABLE_COMPILE_CACHE=1 to force
    # multiple compilations by disabling vLLM's on-disk compile cache.
    # ...
    pass
```

评论区精华

AndreasKaratzas 审核时只回复了 "LGTM", 没有实质讨论。Claude Bot 自动评论提醒这是 fork 来的 PR, 未执行自动审查。因此本 PR 没有显著的争论或设计权衡讨论。

- 审核流程 (other): 无实质技术讨论, 直接合并。

风险与影响

- 风险：风险极低。所有变更仅限于测试代码，且只是对已有逻辑的封装和 de-dup。新 fixture 的语义与之前的手动设置完全一致，不影响任何非测试流程。唯一需要注意：若未来有人修改 `VLLM_DISABLE_COMPILE_CACHE` 的行为，需同步更新 fixture 注释。
- 影响：直接影响四个测试文件中的编译测试，消除 CI 上的间歇性失败，提高 CI 稳定性。对其他系统模块无影响。团队在后续编写需要强制重新编译的测试时应直接使用 `disable_vllm_compile_cache` fixture，而不是手动设置环境变量。
- 风险标记：测试变更，CI 稳定性修复

关联脉络

- PR #49739 [ROCm][CI] Wait for ROCm VRAM to settle between compiled and eager LL…: 同为修复 ROCm 编译测试 CI 稳定性的 PR，与该 PR 关注点类似。
- PR #49749 [CI] Stabilize memory-sensitive compile and structured output tests: 同为稳定 CI 编译测试的 PR，改了同一测试目录下的文件。