PR #49423 完整报告

vllm-project/vllm

[CI] Fix stale/fragile untethered kernels-root tests

合并时间：2026-07-23 04:43

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49423

执行摘要

- 一句话：修复长期未在 CI 运行的内核测试代码
- 推荐动作：建议批准合并。此 PR 是 CI 基建的重要一环，使之前被排除的测试重新得到验证。值得关注的设计决策包括：使用 1-based 索引与 NULL_BLOCK_ID 设计的匹配、fp8 比较的 ulp 容忍策略、以及 autouse fixture 解决测试隔离问题。

功能与动机

根据 Issue #49340 的审计，许多测试文件从未在 CI 中运行，导致它们陈旧和损坏。此 PR 修复了其中一批被标记为 broken 的文件，以便可以启用 kernel-root 的全局 CI job。

实现拆解

1. 修复 FLA 内核测试 (test_fused_recurrent_packed_decode.py, test_fused_sigmoid_gating_delta_rule.py)：由于 CUDA graph padding 保留索引 0 为 NULL_BLOCK_ID，状态索引需从 1 开始，ssm_state 张量多分配一行，输出比较时过滤无效行。
2. 修复 fp8 cache 精度比较 (test_fused_minimax_m3_qknorm_rope_kv_insert.py)：新增 assert_fp8_cache_close 函数，允许 1 ulp 误差，因为 fused kernel 从 fp32 直接量化到 e4m3，而参考路径经 bf16 再量化会有四舍五入差异。
3. 修复 deep_gemm 导入问题 (test_fused_inv_rope_fp8_quant.py)：移除从 deep_gemm 顶层的显式导入，改为本地实现 ceil_div 和 calc_diff，并添加 is_deep_gemm_supported 检查以便在不支持的平台跳过对应测试。
4. 解决测试设备污染 (tests/kernels/conftest.py)：新增 autouse fixture reset_default_torch_device，在每个测试后重置 torch.set_default_device(None)，防止后续测试意外在 CUDA 上创建张量。
5. 更新 flex_attention 测试模型 (test_flex_attention.py)：使用无需认证的 Qwen2.5-1.5B-Instruct 替代需要 token 的 meta-llama 模型。

关键文件：

- tests/kernels/conftest.py（模块 测试夹具；类别 test；类型 test-coverage；符号 reset_default_torch_device）：新增 autouse fixture，解决多个 kernel 测试调用 set_default_device 后未恢复导致的测试隔离问题，是所有 kernel 测试的基础设施。

- tests/kernels/test_fused_inv_rope_fp8_quant.py（模块 逆 RoPE 量化；类别 test；类型 test-coverage；符号 ceil_div, calc_diff）：修复 deep_gemm 导入问题：本地实现 ceil_div 和 calc_diff，添加 is_deep_gemm_supported 检查，使测试在不支持 DeepGEMM 的平台可跳过。
- tests/kernels/test_fused_minimax_m3_qknorm_rope_kv_insert.py（模块 Minimax KV 插入；类别 test；类型 test-coverage；符号 assert_fp8_cache_close）：新增 assert_fp8_cache_close 函数，允许 1 ulp 误差，解决 fused kernel 与 reshape_and_cache_flash 参考路径因量化顺序不同导致的精度差异。
- tests/kernels/test_fused_recurrent_packed_decode.py（模块 递归打包解码；类别 test；类型 test-coverage）：将状态索引从 0-based 改为 1-based 以跳过 NULL_BLOCK_ID，避免输出行包含未初始化的值（NaN）。
- tests/kernels/test_fused_sigmoid_gating_delta_rule.py（模块 Sigmoid 门控；类别 test；类型 test-coverage）：同样修复 NULL_BLOCK_ID 索引问题，将状态索引从 0 改为 1 开始，并增加状态张量尺寸。
- tests/kernels/test_flex_attention.py（模块 Flex 注意力；类别 test；类型 test-coverage）：替换需要认证的模型为开源模型，避免 CI 因 token 问题失败。

关键符号：reset_default_torch_device, assert_fp8_cache_close, ceil_div, calc_diff, test_einsum_end_to_end, test_fused_recurrent_packed_decode_matches_reference, test_fused_sigmoid_gating_delta_rule_update_non_spec, test_fused_sigmoid_gating_delta_rule_update_spec, test_sparse_full, test_sparse_skip_index_branch, test_block_mask_direct_vs_slow_path

关键源码片段

tests/kernels/conftest.py

新增 autouse fixture，解决多个 kernel 测试调用 set_default_device 后未恢复导致的测试隔离问题，是所有 kernel 测试的基础设施。

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
import pytest
import torch


@pytest.fixture(autouse=True)
def reset_default_torch_device():
    """Several kernel tests call torch.set_default_device without restoring
    it, which poisons subsequent tests in the same pytest run (e.g. CPU
    tensors silently created on CUDA). Restore the factory default after
    every test.
    """
    yield
    torch.set_default_device(None)
```

tests/kernels/test_fused_inv_rope_fp8_quant.py

修复 deep_gemm 导入问题：本地实现 ceil_div 和 calc_diff，添加 is_deep_gemm_supported 检查，使测试在不支持 DeepGEMM 的平台可跳过。

```
# Base: from deep_gemm.utils.math import ceil_div
# Head: 内联定义

    def ceil_div(a: int, b: int) -> int:
        return (a + b - 1) // b

    # Before: from deep_gemm.testing import calc_diff
    # Now: local implementation
    def calc_diff(x, y):
        x, y = x.double(), y.double()
        denominator = (x * x + y * y).sum()
        sim = 2 * (x * y).sum() / denominator
        return 1 - sim

    # 并且添加了 is_deep_gemm_supported 检查以跳过不支持的平台
    if not is_deep_gemm_supported():
        pytest.skip("DeepGEMM not supported on this platform")
```

评论区精华

该 PR 来自 fork，自动化审查被跳过；维护者 mgoin 直接批准。无实质讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，主要涉及测试代码。但需要注意：assert_fp8_cache_close 允许 1 ulp 可能暂时掩盖真正的量化差异；conftest 中的 reset_default_torch_device 可能与其他期望默认设备为 None 的测试冲突（但目前看是必要的）。总体回归风险小。
- 影响：对用户无直接影响；对 CI 带来正面影响，确保 tests/kernels/ 下的所有测试被运行，提高内核代码质量监控。对团队：需要关注后续新添加的测试是否也会因设备污染等问题而需要类似的 fixture。
- 风险标记：测试精度放宽可能掩盖偶发误差，conftest fixture 可能与其他 fixture 交互，DeepGEMM 导入路径变更可能影响未来升级

关联脉络

- PR #49340 [CI] Wire untethered test files into CI jobs: 此 PR 是 Issue #49340 的后续，修复该 Issue 中标记为 broken 的测试文件，以便将全部 kernel-root 测试接入 CI。
- PR #39064 [FLA] Use NULL_BLOCK_ID for CUDA graph padding in FLA kernels: 该 PR 引入了 NULL_BLOCK_ID 的设计，导致需要将测试索引从 0 改为 1。