PR #49609 完整报告

vllm-project/vllm

[CI][Bugfix] Fix test isolation in block_int8/ptpc_fp8 MoE kernel tests

合并时间：2026-07-24 01:26

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49609

执行摘要

本次 PR 修复了 MoE kernel 测试中 test_block_int8.py 和 test_triton_moe_ptpc_fp8.py 的测试隔离问题。将 setup_cuda fixture 的 scope 从 "module" 改为默认的 "function"，确保每个测试用例前都重新设置默认 CUDA 设备，避免前序测试（如 test_batched_moe）污染默认设备状态导致后续测试在 CPU 上执行而崩溃。

功能与动机

test_block_int8.py 和 test_triton_moe_ptpc_fp8.py 使用了一个 autouse fixture setup_cuda 来设置默认 CUDA 设备，但 fixture 的 scope 被设为 "module"，导致整个测试模块只执行一次。当同一测试套件中前序的 test_batched_moe 等模块改变了 torch 默认设备状态后，setup_cuda 不会重新运行，使得后续测试在错误的默认设备上运行，触发 NotImplementedError: Could not run '_moe_C::topk_softmax' with arguments from the 'CPU' backend。

实现拆解

1. 修改 tests/kernels/moe/test_block_int8.py 中 setup_cuda fixture 的 scope="module" 参数移除，使 scope 降为默认的 "function"，即每个测试函数执行前都会重新设置默认 CUDA 设备。
2. 同步修改 tests/kernels/moe/test_triton_moe_ptpc_fp8.py 中对应的 setup_cuda fixture，同样移除 scope="module"。
3. 更新两个文件中的 docstring，将 "Sets the default CUDA device for all tests in this module" 改为 "Sets the default CUDA device before each test in this module"，以准确反映新行为。

tests/kernels/moe/test_block_int8.py: 变更后 fixture 定义

```
@pytest.fixture(autouse=True)
def setup_cuda():
    """Sets the default CUDA device before each test in this module."""
    torch.set_default_device("cuda")
```

tests/kernels/moe/test_triton_moe_ptpc_fp8.py: 变更后 fixture 定义
@pytest.fixture(autouse=True) defsetup_cuda(): """Sets the default CUDA device before each test in this module.""" torch.set_default_device("cuda")

评论区精华

无 review 评论。AndreasKaratzas 已批准。

风险与影响

- 风险：极低。仅涉及测试 fixture scope 调整，不涉及生产代码。每个测试前多一次 torch.set_default_device 调用，微乎其微。
- 影响：修复了 CI 中因测试执行顺序导致的不可重现崩溃，提高测试可靠性。无外部用户影响。

关联脉络

暂无直接关联的其他 PR。但该问题与其他使用 scope="module" 的 autouse fixture 测试文件中可能存在类似隐患，值得审视全部 MoE kernel 测试。