

PR #47519 完整报告

vllm-project/vllm

[ROCm][CI] Fix Kernels and Kernels attention test failures

合并时间: 2026-07-05 04:59

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47519>

执行摘要

- 一句话: 修复 ROCm 内核测试因 `prefill_backend` 为 `None` 失败
- 推荐动作: 低风险 bugfix, 值得合并。测试修复确保了 CI 的可靠性, 但需留意 MoE 测试的剩余失败。

功能与动机

ROCm CI 上的 Kernels 和 Kernels Attention 测试失败, 原因为 `prefill_backend` 为 `None` 时尝试 clone 操作。同时 MoE 测试因分布式环境未初始化和 `rocm_aiter_ops.is_enabled()` 返回 `False` 而失败。PR 旨在修复这些失败以使 CI 通过。

实现拆解

1. `tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py`: 将 `prefill_backend` 从 `None` 改为 `torch.empty((1,))`, 避免 clone 操作失败。
2. `tests/kernels/moe/test_ocp_mx_moe.py`:
 - 新增 `import types` 和 `from vllm._aiter_ops import rocm_aiter_ops`。
 - 在测试函数中添加 `monkeypatch` 参数。
 - 使用 `monkeypatch.setattr` 模拟 `ps._TP` 为一个 `SimpleNamespace(world_size=1)`, 绕过分布式环境检查。
 - 使用 `monkeypatch.setattr` 强制设置 `rocm_aiter_ops._AITER_ENABLED = True`, 避免 `assertion` 失败。

关键文件:

- `tests/kernels/moe/test_ocp_mx_moe.py` (模块 测试模块; 类别 `test`; 类型 `test-coverage`) : 修复 MoE 测试因分布式环境未初始化和 AITER 禁用导致的失败。
- `tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py` (模块 测试模块; 类别 `test`; 类型 `test-coverage`) : 修复 `prefill_backend` 为 `None` 时 clone 操作失败的问题。

关键符号: `test_rocm_mxfp4_moe_oracle`

关键源码片段

`tests/kernels/moe/test_ocp_mx_moe.py`

修复 MoE 测试因分布式环境未初始化和 AITER 禁用导致的失败。

```
# tests/kernels/moe/test_ocp_mx_moe.py (partial)
# 在 test_rocm_mxfp4_moe_oracle 函数开头添加 monkeypatch 模拟

import vllm.distributed.parallel_state as ps
# ...

# 设置 TP 组防止 should_use_cdna4_mx_scale_swizzle 检查失败
monkeypatch.setattr(ps, "_TP", types.SimpleNamespace(world_size=1))

# AITER 必须启用，否则 aiter_mxfp4_w4a8_moe 在 dispatch 前断言失败
monkeypatch.setattr(rocm_aiter_ops, "_AITER_ENABLED", True)
```

tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py

修复 prefill_backend 为 None 时 clone 操作失败的问题。

```
# tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py (partial)
# 原先是 SimpleNamespace(prefill_backend=None)
# builder 在内部对 prefill_backend 调用了 clone，None 导致崩溃
# 改为一个有效的空张量以满足 clone 要求
vllm_config.compilation_config.static_forward_context[layer_name] = (
    types.SimpleNamespace(prefill_backend=torch.empty((1,)))
)
```

评论区精华

无 reviewer 讨论。审核者 AndreasKaratzas 已批准并确认测试通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅限于测试文件，不影响生产代码。模拟分布式环境和 AITER 状态是测试常见做法，但需注意模拟值与真实环境一致，否则可能隐藏真正的问题。
- 影响：影响范围仅限于 ROCm CI 的 Kernels 和 Kernels Attention 测试。修复后这些测试将能正常执行，提升 CI 稳定性。MoE 测试仍有其他无关的准确性失败，本次不解决。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR