PR #45052 完整报告

vllm-project/vllm

[Bug] Fix test flashmla for DSv4

合并时间：2026-06-12 04:22

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45052

执行摘要

- 一句话：修复 DSv4 FlashMLA 测试用例
- 推荐动作：建议合并，这是一个及时的测试修复，确保上游 API 变更后测试仍然有效。可略读，逻辑简单。

功能与动机

PR body 指出 pytest tests/kernels/attention/test_flashmla_sparse.py 会因上游 API 接口变更而报错：AttributeError: 'FlashMLASchedMeta' object has no attribute 'dtype' 和 AttributeError: module 'vllm.v1.attention.ops.flashmla' has no attribute 'flash_mla_sparse_prefill'。修复后测试全部通过。

实现拆解

修改 tests/kernels/attention/test_flashmla_sparse.py 文件，共 5 行增加、3 行删除：

1. 更新 test_sparse_flashmla_metadata_smoke 断言：将 assert tile_md.dtype == torch.int32 和 assert num_splits.dtype == torch.int32 替换为 assert isinstance(tile_md, fm.FlashMLASchedMeta)、assert tile_md.tile_scheduler_metadata is None 和 assert tile_md.num_splits is None 和 assert num_splits is None，适配 FlashMLASchedMeta 的新结构。
2. 更新 test_sparse_flashmla_prefill_smoke 函数调用：将 fm.flash_mla_sparse_prefill(q, kv, indices, 1.0, d_v) 改为 fm.flash_mla_sparse_fwd(q, kv, indices, 1.0, d_v)，匹配重命名后的 API。
3. 其他行不变：test_sparse_flashmla_decode_smoke 用例未受影响，断言已正确使用新 API。

关键文件：

- tests/kernels/attention/test_flashmla_sparse.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，修复了两个因上游 API 变更导致的测试失败，确保 DSv4 稀疏 FlashMLA 注意力测试通过。

关键符号：test_sparse_flashmla_metadata_smoke, test_sparse_flashmla_prefill_smoke

关键源码片段

tests/kernels/attention/test_flashmla_sparse.py

唯一变更文件，修复了两个因上游 API 变更导致的测试失败，确保 DSv4 稀疏 FlashMLA 注意力测试通过。

```
# tests/kernels/attention/test_flashmla_sparse.py
# 修复：使用 FlashMLASchedMeta 新属性断言，而非废弃的 dtype

def test_sparse_flashmla_metadata_smoke():
    import vllm.v1.attention.ops.flashmla as fm
    ok, reason = fm.is_flashmla_sparse_supported()
    if not ok:
        pytest.skip(reason)
    device = torch.device("cuda")
    batch_size = 1
    seqlen_q = 1
    num_heads_q = 128
    num_heads_k = 1
    q_seq_per_hk = seqlen_q * num_heads_q // num_heads_k
    topk = 128
    cache_seqlens = torch.zeros(batch_size, dtype=torch.int32, device=device)
    tile_md, num_splits = fm.get_mla_metadata(
        cache_seqlens,
        q_seq_per_hk,
        num_heads_k,
        num_heads_q=num_heads_q,
        topk=topk,
        is_fp8_kvcache=True,
    )
    # 旧 assert: 检查 tile_md.dtype 和 num_splits.dtype
    # 新 assert: 检查 FlashMLASchedMeta 类型和内部属性
    assert isinstance(tile_md, fm.FlashMLASchedMeta)
    assert tile_md.tile_scheduler_metadata is None
    assert tile_md.num_splits is None
    assert num_splits is None


def test_sparse_flashmla_prefill_smoke():
    import vllm.v1.attention.ops.flashmla as fm
    ok, reason = fm.is_flashmla_sparse_supported()
    if not ok:
        pytest.skip(reason)
    device = torch.device("cuda")
    s_q = 1
    s_kv = 1
    h_q = 64
    h_kv = 1
    d_qk = 576
    d_v = 512
```

```
topk = 128
q = torch.zeros((s_q, h_q, d_qk), dtype=torch.bfloat16, device=device)
kv = torch.zeros((s_kv, h_kv, d_qk), dtype=torch.bfloat16, device=device)
indices = torch.zeros((s_q, h_kv, topk), dtype=torch.int32, device=device)
# 旧调用：fm.flash_mla_sparse_prefill
# 新调用：fm.flash_mla_sparse_fwd
out, max_logits, lse = fm.flash_mla_sparse_fwd(q, kv, indices, 1.0, d_v)
assert out.shape == (s_q, h_q, d_v)
assert max_logits.shape == (s_q, h_q)
assert lse.shape == (s_q, h_q)
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅修改了测试文件中的断言和函数名，不涉及生产代码。测试通过验证了与上游 API 的一致性，无回归风险。
- 影响：修复了 DSv4 相关测试的 CI 失败，确保 DeepSeek V4 模型的稀疏 FlashMLA 注意力功能持续可测。对用户无直接影响，但保障了后续开发质量。
- 风险标记：暂无

关联脉络

- PR #36902 [Kernel][Helion][1/N] Add Helion kernel for per_token_group_fp8_quant: 同为 Kernel/Attention 相关，显示该目录活跃度较高。
- PR #41797 [Attention] add triton diff-kv backend for mimo: 同为 Attention 后端变更，显示 Attention 模块正在演进。