

PR #31199 完整报告

sgl-project/sglang

[CI] Fix Flash MLA SM120 test import path

合并时间: 2026-07-15 00:01

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31199>

执行摘要

- 一句话: 修复 Flash MLA SM120 测试导入路径
- 推荐动作: 该 PR 是一个标准、低风险的后续修复, 无需精读。但可作为注意力内核迁移 (PR#30789) 的配套变更, 作为重构时需同步更新测试导入的例子。

功能与动机

PR#30789 将通用注意力内核迁移到 `sglang.kernels` 统一命名空间, 导致旧导入路径 `sglang.srt.layers.attention.flash_mla_sm120` 失效, SM120 测试因 `ImportError` 而失败。本次修复的目的是使测试能够正确导入模块, 从而恢复 CI 的正常运行。

实现拆解

1. 在测试文件 `test/registered/kernels/test_flash_mla_backends.py` 中, 将 `from sglang.srt.layers.attention import flash_mla_sm120 as fmod` 替换为 `from sglang.kernels.ops.attention import flash_mla_sm120 as fmod`, 同时保留其他显式导入不变。
2. 调整导入顺序以使分组更清晰: 将新的 `fmod` 导入放在 `from sglang.kernels.ops.attention.flash_mla_sm120 import` 之前。
3. 删除旧的导入行。
4. 无其他配置或测试逻辑变更。

关键文件:

- `test/registered/kernels/test_flash_mla_backends.py` (模块 测试; 类别 `test`; 类型 `test-coverage`): 唯一修改的文件; 修复 SM120 FlashMLA 测试的导入路径以匹配内核迁移后的新命名空间。

关键符号: 未识别

关键源码片段

`test/registered/kernels/test_flash_mla_backends.py`

唯一修改的文件; 修复 SM120 FlashMLA 测试的导入路径以匹配内核迁移后的新命名空间。

```
"""SM120 FlashMLA sparse decode unit tests.
```

Validates the SM120-specific FlashMLA implementation...

"""

```
from __future__ import annotations
```

```
import unittest
```

```
from unittest import mock
```

```
import torch
```

```
# 修复: 导入路径从 sglang.srt.layers... 改为 sglang.kernels.ops...
```

```
from sglang.kernels.ops.attention import flash_mla_sm120 as fmod
```

```
from sglang.kernels.ops.attention.flash_mla_sm120 import (
```

```
    _D,
```

```
    _NOPE_DIM,
```

```
    _NOPE_ROPE_STRIDE,
```

```
    _NUM_TILES,
```

```
    _ROPE_DIM,
```

```
    _SCALE_STRIDE,
```

```
    _TILE_SIZE,
```

```
    _gather_and_dequant,
```

```
    _sm120_sparse_decode_fwd,
```

```
    flash_mla_with_kvcache_sm120,
```

```
)
```

```
from sglang.kernels.ops.attention.flash_mla_sm120_triton import (
```

```
    _apply_attn_sink,
```

```
    _merge_partial_attn,
```

```
    flash_mla_sparse_decode_triton,
```

```
)
```

```
# 已删除: from sglang.srt.layers.attention import flash_mla_sm120 as fmod
```

```
from sglang.test.ci.ci_register import register_cuda_ci
```

```
from sglang.test.test_utils import CustomTestCase
```

```
register_cuda_ci(est_time=45, stage="base-b", runner_config="1-gpu-large")
```

```
# ... 后续测试代码不变
```

评论区精华

该 PR 没有引发实质性讨论。提交者 mmangkad 在评论中请求重新运行测试，CI 通过后，合并者 alexnails 直接批准并合并。

- 暂无高价值评论线程

风险与影响

- 风险: 变更本身风险极低: 仅涉及单行导入路径的修改, 无运行时逻辑变更。但若未来再次重构导入路径, 此类硬编码的导入需要同步更新。当前修复已通过 CI 验证。

- 影响：直接影响：修复 SM120 FlashMLA 测试在 CI 中的导入错误，使测试恢复正常执行。
间接影响：确保该测试持续验证 SM120 架构的 FlashMLA 实现正确性，避免回归。影响范围仅限于单个测试文件，不涉及生产代码。
- 风险标记：暂无

关联脉络

- PR #30789 [Kernel] Migrate generic attention kernels to sglang.kernels (RFC #29630, Phase 2.5, 4/7): 本 PR 修复了因 PR#30789 中注意力内核迁移导致的导入路径失效问题。