

PR #38288 完整报告

sgl-project/sglang

[AMD][CI] Remove obsolete split-dim check from Kimi-K3 prefill test

合并时间: 2026-09-08 05:42

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/38288>

执行摘要

- 一句话: 修复 AMD CI 因引用已删除 split-dim 内核测试模块导致的收集失败。
- 推荐动作: 此 PR 是一个简单直接的 bugfix, 值得快速审阅和合并。它修复了一个阻塞 CI 的问题, 且改动范围明确、风险可控。建议关注其是否成功解决了报告的 CI 失败问题。

功能与动机

PR body 指出: “MI35x stage-b 测试在收集期间失败, 因为 `test_triton_mla_prefill_gfx950.py` 仍然导入已删除的 `extend_attention_split_dim` 模块。该专用 split-dim 内核及其分发谓词已在 `4e2fd9f1e9` 中移除; 通用的 `extend_attention_fwd` 路径现在处理这些 Kimi-K3 形状。遗留的导入阻止了该文件中所有实时数值测试的运行。” 此变更旨在修复此 CI 失败。

实现拆解

1. 入口与目标: 变更仅涉及测试文件 `test/registered/unit/layers/attention/test_triton_mla_prefill_gfx950.py`, 旨在修复其收集失败, 使其能继续在 AMD MI35x CI 上运行。
2. 核心变更: 移除了对已删除模块 `sglang.kernels.ops.attention.extend_attention_split_dim` 的导入, 以及依赖该模块的整个测试方法 `test_split_dim_dispatch_gates`。此测试方法验证了 `can_use_split_dim_absorbed_extend` 谓词逻辑, 但该逻辑所属的实现模块已被移除。
3. 保留内容: 保留了 `TestKimiK3TritonPrefill` 类中另外三个测试方法: `test_split_dim_ragged_bf16`、`test_zero_prefix_fp8_flag` 和 `test_absorbed_fp8_cached_prefix`。这些测试均调用当前存在的 `extend_attention_fwd` 函数, 验证 Kimi-K3 形状下的不同 prefill 场景 (BF16 ragged、零前缀 FP8、吸收式 FP8 缓存前缀), 确保核心功能测试覆盖不受影响。
4. 配套影响: 此为纯测试变更, 不影响任何运行时内核、调度逻辑或模型行为。通过修复测试收集失败, 确保了 AMD CI 上该测试套件的稳定运行。

关键文件:

- `test/registered/unit/layers/attention/test_triton_mla_prefill_gfx950.py` (模块 MLA 注意力测试; 类别 test; 类型 test-coverage; 符号 `test_split_dim_dispatch_gates`): 这是本次唯一变更的文件, 包含导致 CI 收集失败的过时导入和测试。

关键符号: `test_split_dim_dispatch_gates`, `can_use_split_dim_absorbed_extend`

关键源码片段

test/registered/unit/layers/attention/test_triton_mla_prefill_gfx950.py

这是本次唯一变更的文件，包含导致 CI 收集失败的过时导入和测试。

变更前：测试文件头部包含对已删除模块的导入

```
from sglang.kernels.ops.attention.extend_attention_split_dim import (
    can_use_split_dim_absorbed_extend,
)
```

变更后：测试文件头部移除了对已删除模块的导入

```
from sglang.kernels.ops.attention.extend_attention import extend_attention_fwd
from sglang.srt.envron import envs
from sglang.srt.utils import get_device, is_gfx95_supported, is_hip
from sglang.test.ci.ci_register import register_amd_ci
```

变更前：测试类中包含一个测试已删除逻辑的完整方法

```
def test_split_dim_dispatch_gates(self):
    # ... (约 60 行代码，测试 can_use_split_dim_absorbed_extend 函数的各种条件)
    self.assertTrue(
        can_use_split_dim_absorbed_extend(q, k, v, o, k_buffer, v_buffer, **kwargs)
    )
    # ... 测试 FP8 和其他覆盖条件
```

变更后：测试类中删除了上述整个 test_split_dim_dispatch_gates 方法。

只保留了三个针对 extend_attention_fwd 的测试方法：

test_split_dim_ragged_bf16, test_zero_prefix_fp8_flag, test_absorbed_fp8_cached_prefix

评论区精华

PR 中没有 review 讨论。唯一的评论是 reviewer [HaiShaw](#) 的批准（APPROVED），表明变更符合预期且无需进一步讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。此变更仅移除了对已删除代码的测试引用，属于测试清理。它不修改任何生产代码路径、内核或配置。唯一的风险是如果未来需要重新引入 split-dim 逻辑，需要手动恢复此测试，但鉴于原模块已被故意删除，这种可能性很低。
- 影响：影响范围小且直接。
- 对用户：无直接影响。此变更为内部测试修复。
- 对系统：无功能影响。确保了 AMD CI 上 Kimi-K3 prefill 测试套件的稳定性，防止因测试收集失败而掩盖其他潜在问题。
- 对团队：维护了 CI 的健康状态，确保相关测试能持续运行。移除了过时代码，减少了维护负担和潜在混淆。
- 风险标记：测试覆盖

关联脉络

- PR #38293 [CP V1 Deprecation 3.5/5] Deprecate HIP/NPU/MUSA prefill CP and remove legacy implementation: 此 PR 移除了旧的 prefill CP 实现，而 38288 移除了对已删除 split-dim 内核的测试引用。两者都是清理已废弃代码的维护工作，但功能线不同。38293 涉及更广泛的调度器重构，38288 是针对特定测试文件的清理。