# PR #22573 完整报告

sgl-project/sglang

fix: restore CPU flash_attn test to use sgl_kernel directly

合并时间：2026-04-11 12:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22573

## 执行摘要

- 一句话：修复 CPU 测试因导入路径变更导致的 NotImplementedError，恢复使用原生 CPU 实现。
- 推荐动作：这是一个简单的修复 PR，无需深入精读。值得关注的是它揭示了测试环境对硬件依赖的敏感性，以及导入路径选择对跨平台兼容性的影响。

## 功能与动机

修复由 PR #20796 引入的 CPU CI 失败问题。PR #20796 将测试导入路径改为使用社区版 fa3 内核，但该路径在 CPU 环境下会检查 CUDA sm80+ 支持，导致抛出 NotImplementedError。需要恢复使用 CPU 原生实现以确保测试通过。

## 实现拆解

仅修改一个测试文件 test/srt/cpu/test_flash_attn.py：1. 移除对 sglang.jit_kernel.flash_attention 的导入；2. 添加 sgl_kernel 导入（仅用于类型提示）；3. 将 flash_attn_varlen_func 直接赋值为 torch.ops.sgl_kernel.flash_attn_varlen_func，这是 CPU 原生实现。

关键文件：

- test/srt/cpu/test_flash_attn.py（模块 test）：唯一被修改的文件，修复了导入路径导致的 CPU 测试失败问题。

关键符号：flash_attn_varlen_func

## 评论区精华

本 PR 没有 review 评论，但从 PR body 和关联 Issue 可以看出，这是一个直接的修复，旨在解决由 PR #20796 引入的回归问题。

- 导入路径导致的 CPU 测试失败 (correctness): 恢复使用 torch.ops.sgl_kernel.flash_attn_varlen_func 作为 CPU 原生实现。

## 风险与影响

- 风险：风险极低：1. 仅修改测试文件，不影响生产代码；2. 恢复为原有工作路径，已被历史验证；3. 变更范围极小（4 行改动），回归风险可控。

- 影响：影响范围有限：1. 确保 CPU CI 测试套件（per-commit-cpu）能够正常执行 test_flash_attn.py；2. 对用户和系统无直接影响；3. 对团队而言，修复了 CI 稳定性问题，避免因测试失败干扰开发流程。
- 风险标记：测试环境依赖

## 关联脉络

- PR #20796 Kernels community fa3: 本 PR 修复了由 #20796 引入的回归问题，#20796 将导入路径改为使用社区版 fa3 内核，导致 CPU 测试失败。