PR #24569 完整报告

sgl-project/sglang

[AMD] Register 8 CPU-bound unit tests for AMD 1-GPU PR CI

合并时间：2026-05-09 07:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24569

执行摘要

- 一句话：为 AMD CI 注册 8 个 CPU-bound 单元测试
- 推荐动作：建议其他平台的开发者在添加新的 CPU-bound 测试时，参考此模式同时注册 register_cuda_ci 和 register_amd_ci，以最大化跨平台覆盖。该 PR 本身逻辑简单，无需深度精读，但团队的 CI 基础设施设计（通过装饰器式注册分离套件定义）值得借鉴。

功能与动机

这些单元测试已经是 NVIDIA per-commit CI 的一部分，且仅使用纯 Python/CPU torch 操作，没有 FlashInfer、FA3、ModelOpt/Marlin 等 GPU 专属依赖。将它们注册到 AMD CI 中，可以在不增加额外维护负担的前提下，更好地在 AMD 平台上发现回归问题，确保代码跨平台兼容性。PR body 中明确列出了每个测试的预估时间和安全依据。

实现拆解

1. 识别可移植的 CPU-bound 测试：从现有 NVIDIA CI 的 stage-b-test-1-gpu-small 套件中，筛选出仅依赖纯 Python 和 CPU torch ops 的测试文件，共 8 个，涵盖 auto_benchmark 工具、conv 层、LoRA 内存池、模型加载模拟（Llava）以及 radix cache 等模块。
2. 导入并注册 AMD CI：在每个测试文件的导入区，将 register_cuda_ci 的导入替换为同时导入 register_amd_ci 和 register_cuda_ci，并在已有 register_cuda_ci 调用下方新增一行 register_amd_ci(est_time=..., suite="stage-b-test-1-gpu-small-amd")，预估时间与 NVIDIA CI 设置一致。
3. 验证注册结果：使用 AST 基于收集器重新扫描 test/registered/，确认 AMD CI 测试计数从 88 升至 96，且所有新文件都出现在 stage-b-test-1-gpu-small-amd 套件中。通过 AMD PR CI 实际运行，所有 14 个分区的测试全部通过。
4. 保持 NVIDIA CI 不变：原有 register_cuda_ci 调用和测试逻辑未做任何修改，确保对 NVIDIA CI 无影响。

关键文件：

- test/registered/unit/models/test_llava.py（模块 llava 测试；类别 test；类型 test-coverage）：核心变更文件之一，为 Llava 模型加载测试添加 AMD CI 注册，重要性评分最高。

- test/registered/unit/auto_benchmark/test_dataset_tools.py（模块 数据集；类别 test；类型 test-coverage）：为 auto_benchmark 数据集工具测试添加 AMD CI 注册，扩展基准测试覆盖。
- test/registered/unit/auto_benchmark/test_run_candidate.py（模块 候选实验；类别 test；类型 test-coverage）：为 auto_benchmark 候选实验测试添加 AMD CI 注册。
- test/registered/unit/auto_benchmark/test_search_tools.py（模块 搜索工具；类别 test；类型 test-coverage）：为 auto_benchmark 搜索工具测试添加 AMD CI 注册。
- test/registered/unit/layers/test_conv_layer.py（模块 卷积层；类别 test；类型 test-coverage）：为卷积层测试添加 AMD CI 注册，该测试使用 CPU 张量。
- test/registered/unit/lora/test_mem_pool_ep_unit.py（模块 LoRA 池；类别 test；类型 test-coverage）：为 LoRA 内存池单元测试添加 AMD CI 注册，该测试被标记为 CPU-only。
- test/registered/unit/mem_cache/test_decode_radix_lock_ref.py（模块 radix 锁；类别 test；类型 test-coverage）：为 radix cache 解码锁引用测试添加 AMD CI 注册。
- test/registered/unit/mem_cache/test_radix_cache_slru_accuracy.py（模块 radix 缓存；类别 test；类型 test-coverage）：为 radix cache SLRU 精度测试添加 AMD CI 注册。

关键符号：未识别

关键源码片段

test/registered/unit/models/test_llava.py

核心变更文件之一，为 Llava 模型加载测试添加 AMD CI 注册，重要性评分最高。

```
import unittest
from unittest.mock import patch

from sglang.srt.models.llava import AutoModel, LlavaForConditionalGeneration
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.test_utils import CustomTestCase

# 注册到 NVIDIA CI（保持不变）
register_cuda_ci(est_time=9, suite="stage-b-test-1-gpu-small")
# 新增注册到 AMD CI（测试不依赖 GPU 专属库，可安全运行）
register_amd_ci(est_time=9, suite="stage-b-test-1-gpu-small-amd")
```

评论区精华

该 PR 的 review 过程非常简洁，HaiShaw 直接 approve，没有提出任何修改意见或讨论。PR author 在 comment 中发布了 CI 状态确认，所有新增测试均通过。因此没有实质性的技术争论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及在每个测试文件中增加一行注册调用和一个导入语句，不修改任何测试逻辑或生产代码。PR author 已确认测试仅使用 CPU 和 plain torch ops，不依赖任何 AMD 不支持的 GPU 专属库，因此 AMD CI 上执行失败的可能性很小。实际 CI 运行结果也验证了这一点：所有 14 个分区的测试均稳定通过。
- 影响：对最终用户无直接功能影响。对开发团队：AMD PR CI 的单元测试数量从 88 增加到 96，覆盖了更多模块（auto_benchmark、conv 层、LoRA、模型加载模拟、radix cache 等），有助于在 AMD 平台上提前发现回归。新增测试总预计执行时间约 61 秒，分散在 14 个分区中，对 CI 整体耗时影响可忽略。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR