PR #23455 完整报告

sgl-project/sglang

[AMD] Restore test_zimage_turbo.py and test_int4fp8_moe.py with __main__ entry

合并时间：2026-04-23 13:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23455

执行摘要

- 一句话：恢复 AMD 测试文件，添加 __main__入口以修复 CI 静默跳过。
- 推荐动作：建议测试维护者关注入口点添加模式，以确保测试文件能正确执行；对于性能敏感测试，可考虑实现模型缓存以优化 CI 运行时间。

功能与动机

PR body 指出，文件在 #23305 中被移除，原因是缺少 __main__ 入口块，导致在 CI 中静默跳过。添加入口点以确保测试在 python3 file.py -f 下能实际执行，提升测试覆盖率。

实现拆解

1. 恢复 Z-Image-Turbo 测试文件：在 test/registered/amd/test_zimage_turbo.py 中添加 if __name__== "__main__": sys.exit(pytest.main([__file__, "-v"]))，使 pytest 测试能作为脚本运行。
2. 恢复 int4fp8_moe 量化测试文件：在 test/registered/quant/test_int4fp8_moe.py 中添加 if __name__== "__main__": unittest.main()，确保 unittest 测试能独立执行。
3. CI suite 注册调整：在提交中调整了 register_amd_ci 的 suite 参数，确保测试被正确调度到 AMD CI 套件（如从 stage-b-test-1-gpu-small-amd 改为 stage-b-test-2-gpu-large-amd）。

关键文件：

- test/registered/amd/test_zimage_turbo.py（模块 扩散测试；类别 test；类型 test-coverage；符号 _save_image_and_write_summary, _compute_clip_score, TestZImageTurboAMD, teardown_class）：恢复 Z-Image-Turbo 扩散模型测试，添加 pytest 入口点以确保 CI 执行
- test/registered/quant/test_int4fp8_moe.py（模块 量化测试；类别 test；类型 test-coverage；符号 TestMixtralAccuracy, setUpClass, tearDownClass, test_gsm8k）：恢复 Mixtral-8x7B int4fp8_moe 量化准确性测试，添加 unittest 入口点以确保 CI 执行

关键符号：_save_image_and_write_summary, _compute_clip_score, TestZImageTurboAMD.test_diffusion_generation, TestMixtralAccuracy.test_gsm8k

关键源码片段

test/registered/amd/test_zimage_turbo.py

恢复 Z-Image-Turbo 扩散模型测试，添加 pytest 入口点以确保 CI 执行

```
import sys
import pytest

# ... 其他导入和测试代码 ...

if __name__ == "__main__":
    # 添加 pytest 入口点，使文件能作为脚本运行，确保 CI 测试执行
    sys.exit(pytest.main([__file__, "-v"]))
```

test/registered/quant/test_int4fp8_moe.py

恢复 Mixtral-8x7B int4fp8_moe 量化准确性测试，添加 unittest 入口点以确保 CI 执行

```
import unittest

# ... 其他导入和测试代码 ...

class TestMixtralAccuracy(CustomTestCase):
    # 测试类定义，包括 setUpClass、tearDownClass 和 test_gsm8k 方法
    @classmethod
    def setUpClass(cls):
        cls.model = "mistralai/Mixtral-8x7B-Instruct-v0.1"
        # 启动服务器等初始化逻辑
        cls.process = popen_launch_server(cls.model, cls.base_url, timeout=45 * 60, other_args=
        other_args)

    def test_gsm8k(self):
        # 运行 GSM8K 评估并断言准确性
        metrics = run_eval(args)
        self.assertGreater(metrics["score"], 0.56)

if __name__ == "__main__":
    # 添加 unittest 入口点，使测试能独立执行
    unittest.main()
```

评论区精华

review 中，gemini-code-assist[bot] 指出两个问题：在 test_zimage_turbo.py 中，CLIP 模型和处理器每次调用都加载，建议缓存以提升性能；在 test_int4fp8_moe.py 中，est_time=313 可能低估了测试时间，建议更新。PR 被批准合并，但这些问题可能未在本次 PR 中解决。

- CLIP 模型缓存建议 (performance): 建议未在 PR 中实施，PR 已合并，但性能优化可能作为后续改进。
- 测试时间估计更新 (testing): 建议未在 PR 中实施，PR 已合并，但时间估计可能需后续调整。

风险与影响

- 风险：风险较低：主要风险是测试性能瓶颈（CLIP 模型加载）可能导致 CI 时间增加；est_time 低估可能引起 CI 调度问题或超时。但变更本身是测试修复，不影响生产代码。
- 影响：影响范围：确保 AMD 相关的扩散模型和量化测试在 CI 中运行，提高测试覆盖率，有助于早期发现问题。对最终用户无直接影响，但提升了系统维护的可靠性。
- 风险标记：测试性能瓶颈，CI 时间估计不足

关联脉络

- PR #23305 从上下文推断，可能是一个移除测试文件的 PR: PR body 提到文件在 #23305 中被移除，因此本 PR 是恢复这些文件，直接关联。