# PR #24721 完整报告

sgl-project/sglang

ci: prune per-commit CUDA tests — move 25 files + 13 testcases to test/manual/

合并时间：2026-05-09 06:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24721

## 执行摘要

- 一句话：裁剪 per-commit CUDA 测试，39 个移入 manual
- 推荐动作：本 PR 展示了通过目录约定控制 CI 测试发现的简洁方法，值得学习。建议关注其对测试覆盖的长期影响，确保 manual 测试在关键发布前仍被有效执行。设计决策：使用文件系统路径而非配置列表来管理 CI 范围，降低了维护复杂度。

## 功能与动机

根据 2026-05-07 工作会话的修剪目录，将标记为 remove 的测试移出 per-commit CUDA CI，以缩短 CI 管道时间。Per-commit CUDA 管道减少了约 174 分钟的注册估计时间（整文件移动 8622 秒 + 子集移除 1831 秒）。

## 实现拆解

1. CI 注册解析器（python/sglang/test/ci/ci_register.py）仅扫描 test/registered/*/.py，因此将文件移至 test/manual/ 可自动跳过 CI 发现，注册调用变为运行时无操作。
2. 整文件移动：将 25 个测试文件（如 test_qwen35_fp4_triton、test_deepseek_v3_basic 等）直接从 registered 移至 manual 对应子目录，保持文件内所有测试类不变。
3. 测试类拆分：对于 10 个文件中的 13 个测试类，将指定类迁移到 manual 下的 _archived.py 文件中，同时保留原文件并调整 register_cuda_ci 的 est_time 参数（按剩余测试数量等比缩小）。
4. 同步处理 AMD 副本：test_deepseek_v3_basic.py 的 AMD 版本也一同移动，确保平台一致。
5. 测试计划：验证 CI 通过，并手动 spot-check 移动后的测试仍然可运行。

关键文件：

- test/registered/4-gpu-models/test_qwen35_models.py（模块 多卡模型；类别 test；类型 test-coverage；符号 TestQwen35FP4, TestQwen35FP4MTP, register_cuda_ci, TestQwen35FP4MTPV2）：原注册测试文件，移除了两个高耗时测试类（TestQwen35FP4、TestQwen35FP4MTP）并调整估计时间从 768s 到 260s，是测试裁剪的核心操作之一。
- test/manual/4-gpu-models/test_qwen35_models_archived.py（模块 多卡模型；类别 test；类型 test-coverage；符号 TestQwen35FP4, TestQwen35FP4MTP, setUpClass, tearDownClass）：新创建的归档文件，包含从原 registered 文件移出的 TestQwen35FP4 和 TestQwen35FP4MTP 两个测试类，确保可手动运行。

- test/manual/mla/test_mla_int8_deepseek_v3_archived.py（模块 注意力模块；类别 test；类型 test-coverage；符号 TestMLADeepseekV3ChannelInt8, TestMLADeepseekV3BlockInt8, setUpClass, tearDownClass）：归档了 DeepSeek-V3 INT8 量化测试的通道和块精度测试类，减少 CI 时间。
- test/manual/distributed/test_dp_attention_archived.py（模块 分布式；类别 test；类型 test-coverage；符号 TestDPAttentionDP2TP2DeepseekV3MTP, setUpClass, tearDownClass, test_gsm8k）：归档了 DP Attention + DeepSeek-V3 MTP 分布式测试类，进一步缩减 CI 耗时。
- test/manual/quant/test_nvfp4_gemm_archived.py（模块 量化；类别 test；类型 test-coverage；符号 FP4GemmBase, TestFP4GemmAuto, setUpClass, tearDownClass）：归档了 NVFP4 GEMM 后端测试（auto 模式），减少 CI 中量化相关测试时间。
- test/manual/mla/test_flashmla_archived.py（模块 注意力模块；类别 test；类型 test-coverage；符号 TestFlashMLAAttnBackend, setUpClass, tearDownClass, test_gsm8k）：归档了 FlashMLA 注意力后端测试类，包含 MTP 推测解码测试。

关键符号：TestQwen35FP4.test_gsm8k, TestQwen35FP4MTP.test_gsm8k, TestMLADeepseekV3ChannelInt8.test_gsm8k, TestMLADeepseekV3BlockInt8.test_gsm8k, TestDPAttentionDP2TP2DeepseekV3MTP.test_gsm8k, TestFP4GemmAuto.test_gsm8k, TestFlashMLAAttnBackend.test_gsm8k

## 评论区精华

审核者 Kangyan-Zhou 对 test/manual/amd/test_deepseek_v3_basic.py 的移动提出疑问："Is this change expected?"，确认该移动是否经工作会话同意。根据 PR 描述，这是整文件移除的一部分，AMD 副本也按决策移动，属于预期行为。该疑问未引发进一步争议，PR 随后合并。

- AMD 测试文件移动是否预期 (question): PR 描述中已说明工作会话决定对两个平台副本都执行整文件移除，因此移动是预期的。审核者未进一步追问，PR 随后合并。

## 风险与影响

- 风险：
  1. CI 覆盖缺失：移出的测试不再自动运行，可能遗漏回归，尤其是 DeepSeek、Qwen3.5 FP4、MLA 等关键模型的精度测试。
  2. 手动测试可维护性：manual 目录下的测试可能因缺乏持续运行而逐渐失效。
  3. 时间估计偏差：调整后的 est_time 基于剩余测试数量等比计算，可能与实际运行时间有偏差，但影响较小。
  4. 文件路径依赖：移动文件可能破坏相对导入，但 PR 已通过 ast.parse 验证语法正确。 - 影响：对用户：无直接功能影响。对系统：CI 管道时间减少约 174 分钟，加快开发迭代。对团队：开发者需要手动运行这些测试以验证相关变更；需维护两套测试位置（registered 和 manual）。影响范围中等，涉及多个模型模块（DeepSeek、Qwen、NVFP4、MLA 等）的测试覆盖。 - 风险标记：CI 覆盖减少，手动测试可维护性，时间估计偏差

# 关联脉络

- 暂无明显关联 PR