PR #20751 完整报告

sgl-project/sglang

[NPU]Add a full test pipeline on NPU, resolve issues in the NPU test architecture

合并时间：2026-04-01 19:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20751

执行摘要

本 PR 新增了 NPU 完整测试流水线，通过创建手动触发的工作流、优化测试配置和修复环境问题，提升版本发布前的测试覆盖率，是 NPU 功能质量保证的关键基础设施改进。

功能与动机

根据 roadmap（issue 20079），需在版本发布前执行所有 NPU 测试用例，支持测试最新社区代码和每日构建包。PR body 明确指出：'add an NPU release pipeline to execute all NPU test cases before version release'，以解决测试架构中的不足，确保 NPU 功能的可靠性。

实现拆解

CI/CD 模块

- 新增完整测试流水线：在 .github/workflows/full-test-npu.yml 中定义手动触发的工作流，支持输入参数 image_a3（镜像路径）和 skip_install_flag（跳过安装标志），示例配置包括测试最新代码或每日构建包。
- 优化现有工作流：修改 .github/workflows/nightly-test-npu.yml 和 .github/workflows/pr-test-npu.yml，调整测试套件命名（如 per-commit-* 改为 stage-b-test-*）、更新定时任务（夜间测试时间从凌晨 1 点改为 2 点）并集成镜像参数，保持与社区配置一致。

测试框架模块

- 测试套件扩展：在 test/run_suite.py 中添加 'full-*' 套件列表（如 'full-1-npu-a3'），支持完整测试流水线的执行。
- 测试用例重命名与注册：将多个测试文件从 'test/srt/ascend/' 移动并重命名为 'test/registered/ascend/' 下的 'npu' 前缀文件（如 test_npu_hicache_mha.py），并在文件中添加 register_npu_ci 调用，注册到新测试套件。例如：
- 数据集预下载：在工作流中预下载 GSM8K 数据集（cp ~/.cache/modelscope/hub/datasets/tmp/test.jsonl /tmp），避免网络超时影响测试。

依赖管理模块

- 在 python/pyproject_npu.toml 中添加 'hf_transfer' 和 'huggingface_hub' 依赖，增强模型下载能力，支持测试环境搭建。

评论区精华

由于 Review 评论为空，未发生技术讨论。Issue 评论中仅有自动化 bot 指令（如 /tag-and-rerun-ci），用于触发 CI 测试，无实质性技术交锋。这表明变更可能在内部已达成共识，或通过提交历史迭代解决。

风险与影响

技术风险：

1. 工作流配置风险：新流水线文件中的输入参数处理逻辑（如默认值设置）若错误，可导致 CI 运行失败，需仔细验证语法。
2. 测试套件兼容性风险：重命名测试文件可能破坏现有测试运行器识别，需确保 test/run_suite.py 更新完整。
3. 环境依赖风险：预下载数据集和镜像路径配置增加维护负担，若缓存服务失效，可能引发测试超时。

影响分析：

- 对系统：提升 NPU 测试的全面性，减少版本发布时的回归风险，但新增工作流可能增加 CI 资源消耗（通过手动触发缓解）。
- 对团队：优化测试流程，提高效率，但开发者需适应新配置模式。

关联脉络

从历史 PR 看，NPU 相关改进持续进行：

- PR 21807 更新 Ascend 文档，与本 PR 的测试流水线协同，确保文档与测试对齐。
- PR 21347 修复 Qwen3.5 模型在 NPU 上的权重加载问题，本 PR 的测试流水线可用于验证此类修复，体现测试基础设施对功能稳定性的支撑。 整体上，本 PR 是 NPU 功能成熟度提升的一部分，通过强化测试架构为后续 NPU 特性（如量化、多模态）提供质量保障。