PR #21960 完整报告

sgl-project/sglang

[diffusion][CI]: route multimodal component accuracy through run_suite

合并时间：2026-04-10 23:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21960

执行摘要

- 一句话：统一多模态组件准确性测试至 run_suite.py 入口点，简化 CI 工作流。
- 推荐动作：建议 CI 工程师和测试团队精读此 PR，重点关注 run_suite.py 中组件准确性套件的设计决策（如文件级分区与项目级分区的权衡）和 CI 工作流的更新模式，以借鉴如何集成特殊测试需求到统一运行框架中。

功能与动机

根据 PR body，动机源于 Issue #18709，即多模态扩散组件准确性测试此前临时使用显式的 workflow-side pytest/torchrun 命令以通过 CI，但这使得测试脱离正常的多模态运行器路径。目标是恢复 python/sglang/multimodal_gen/test/run_suite.py 作为唯一入口点，同时保持组件准确性所需的执行行为（如文件级隔离和分布式启动）。

实现拆解

实现方案主要涉及三个层面：1) 在 run_suite.py 中添加 component-accuracy-1-gpu 和 component-accuracy-2-gpu 新套件，并引入窄分支逻辑，针对这些套件进行文件级分区和分布式执行（2-GPU 使用 torch.distributed.run）；2) 更新 CI 工作流文件 .github/workflows/pr-test-multimodal-gen.yml，将组件准确性作业路由到 run_suite.py 而非直接命令；3) 修复相关准确性测试文件，如 accuracy_utils.py 和 component_accuracy.py，以支持新模型（如 LTX-2.3）和改善加载逻辑。

关键文件：

- .github/workflows/pr-test-multimodal-gen.yml（模块 CI）：核心 CI 工作流文件，添加了组件准确性测试作业并修改调用方式，直接影响测试执行流程。
- python/sglang/multimodal_gen/test/run_suite.py（模块 测试运行器）：测试运行器主文件，新增组件准确性套件和执行逻辑，是统一入口点的关键实现。
- python/sglang/multimodal_gen/test/server/accuracy_utils.py（模块 准确性测试）：准确性测试工具文件，大幅修改了组件路径解析和模型加载逻辑，影响测试准确性和兼容性。
- python/sglang/multimodal_gen/test/server/component_accuracy.py（模块 准确性测试）：组件准确性引擎文件，增加了本地 safetensors 加载和清理逻辑，提升测试健壮性。

关键符号：run_component_accuracy_files, partition_test_files, _build_transformer_hook_inputs

评论区精华

主要讨论在 Issue 评论中，围绕测试稳定性和 CI 重跑。mickqian 评论：“could you also make sure this test is robust and not flaky”，Ratish1 回应已本地测试并修复问题，如显式设置 2-GPU 测试的 num_gpus=2 以防止 CI 挂起。结论是测试已稳定化，并通过多次重跑验证。

- 测试稳定性讨论 (testing): 测试已稳定化，通过多次 CI 重跑验证。

风险与影响

- 风险：技术风险包括：1) CI 配置变更（.github/workflows/pr-test-multimodal-gen.yml）可能引入错误，导致测试作业失败或不触发；2) 执行逻辑变更（run_suite.py 中的新分支）可能影响现有测试套件（如 unit/1-gpu/2-gpu）的兼容性；3) 准确性测试代码修改（如 accuracy_utils.py 中的音频输入支持）可能引入回归错误；4) 多模型兼容性风险，新增的 LTX-2.3 跳过逻辑可能掩盖潜在准确性差异。
- 影响：影响范围：对最终用户透明，但显著改善 CI 流程和测试团队的工作效率。影响程度：中等，统一了测试入口点，提升 CI 可维护性和稳定性；间接增强多模态扩散模型的测试覆盖和可靠性。
- 风险标记：CI 配置变更，测试执行逻辑变更，多模型兼容性

关联脉络

- PR #22483 [CI] Remove Slack notification from ci-auto-bisect workflow: 同为 CI 基础设施改进，展示了仓库对 CI 流程的持续优化趋势。
- PR #22305 [CI] Update est_time for 64 tests based on actual elapsed times: 涉及 CI 测试时间管理，与本 PR 的测试执行优化相关。