PR #22284 完整报告

sgl-project/sglang

Add fast-fail to multimodal-gen CI

合并时间：2026-04-08 06:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22284

执行摘要

本 PR 为多模态生成 CI 添加快速失败机制，通过补全环境健康检查步骤和启用 pytest -x 参数，优化 PR 测试反馈速度，减少无效测试执行，属于基础设施改进。

功能与动机

PR 旨在解决多模态生成 CI 测试效率问题：

- 补全检查步骤：multimodal-gen-test-1-b200 任务此前缺少 check-stage-health 步骤，导致环境检查不一致。
- 启用快速失败：PR 运行中测试失败时，希望立即停止而非运行所有测试，以加速反馈并节省资源。PR body 明确说明："Add pytest -x (exitfirst) to run_suite.py when --continue-on-error is not set, so PR runs stop on first test failure instead of running all tests"。

实现拆解

变更涉及两个文件：

1. CI 工作流文件（.github/workflows/pr-test-multimodal-gen.yml）：
   - 在第 175 行后添加 - uses: ./.github/actions/check-stage-health，使该任务与其他多模态任务保持一致。
2. 测试运行脚本（python/sglang/multimodal_gen/test/run_suite.py）：
   - 修改 run_pytest 函数，新增 exitfirst 参数：python def run_pytest(files, filter_expr=None, exitfirst=False): base_cmd = [sys.executable, "-m", "pytest", "-s", "-v"] if exitfirst: base_cmd.append("-x")
   - 在 main 函数中，根据 --continue-on-error 参数决定是否启用快速失败：python exit_code = run_pytest(my_items, exitfirst=not args.continue_on_error)
   - 设计上区分了 PR 运行（默认启用 -x）和定时运行（使用 --continue-on-error 禁用 -x）。

评论区精华

无 review 评论，仅 PR author 在 body 中说明了变更内容和测试计划。

风险与影响

- 风险：check-stage-health 步骤可能引入额外检查失败，但这是有益的环境健康检查；pytest -x 可能掩盖后续可通过的测试失败，但 PR body 已通过场景区分（PR 运行启用，定时运行禁用）缓解此问题。
- 影响：
  - 对开发者：PR 测试失败时反馈更快，提升开发体验。
- 对系统：减少 CI 资源消耗，提高测试流水线效率。
- 对团队：标准化多模态 CI 检查步骤，增强一致性。

关联脉络

- 与 PR #22251（“[diffusion] CI: fix consistency check”）相关，同属多模态 / 扩散模型 CI 优化，关注测试稳定性和配置调整。
- 与 PR #22229（“fix(pcg,mm): fix zeroing of input_embeds when replay PCG”）相关，均涉及多模态模型测试修复，反映团队在持续完善该领域测试基础设施。