PR #22727 完整报告

sgl-project/sglang

Revert "Upgrade CI default CUDA version from 12.9 to 13.0"

合并时间：2026-04-14 05:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22727

执行摘要

本次 PR 回滚了 CUDA 13.0 升级，将 CI 默认 CUDA 版本恢复为 12.9，以解决内核测试失败问题。改动涉及多个 CI 配置文件和脚本，确保构建环境稳定，但延迟了与 Torch 2.11 的 CUDA 版本匹配。基础设施团队应关注此变更，以规避未来升级的类似障碍。

功能与动机

回滚决策基于关联 Issue #21441 中作者 Fridge003 的评论："There are some kernel tests yet to be fixed"。这表明升级到 CUDA 13.0 后，内核测试出现未修复的问题，影响了 CI 流水线的稳定性。因此，通过回滚来临时解决测试障碍，维持开发效率。

实现拆解

实现通过修改 5 个文件，将 CUDA 版本从 13.0 降级回 12.9：

文件路径	关键变更	影响
.github/workflows/pr-test.yml	将构建矩阵中的 cuda-version 从 "13.0" 改为 "12.9"，并注释掉 13.0 配置	直接影响所有 CI 测试任务的 CUDA 环境
python/pyproject.toml	将 cuda-python 依赖从 >=13.0 改为 ==12.9，Torch 索引从 cu130 改回 cu129	调整 Python 包管理，确保依赖兼容性
scripts/ci/cuda/ci_install_dependency.sh	将 CU_VERSION 变量从 cu130 改回 cu129，简化 sgl-kernel wheel 安装逻辑	核心安装脚本，控制构建流程的 CUDA 版本
scripts/ci/cuda/ci_install_deepep.sh	修改 Blackwell 架构编译配置，因 CI 环境错误报告 CUDA 版本	避免因环境配置问题导致的编译错误
scripts/ci/cuda/ci_download_flashinfer_jit_cache.sh	将 CU_VERSION 从 cu130 改回 cu129	次要脚本更新，保持一致性

评论区精华

无 review 讨论。关键决策线索来自 Issue #21441 的评论：

> Fridge003: "There are some kernel tests yet to be fixed"

这直接导致了回滚操作，强调测试稳定性在 CI 升级中的优先级。

风险与影响

- 技术风险：回滚可能掩盖 CUDA 13.0 的兼容性问题，长期需重新处理；环境配置不一致（如 ci_install_deepep.sh 中所述）可能引发编译错误；依赖降级可能与新特性冲突。
- 影响分析：主要影响 CI 构建和测试环境，确保开发流水线稳定；对用户无直接影响，但延迟 CUDA 升级可能限制新硬件支持和性能优化。

关联脉络

- 本 PR 直接回滚了 PR #21441，后者旨在升级 CUDA 版本以匹配 Torch 2.11。
- 从近期历史 PR 看，类似基础设施变更（如 PR 22657、22653）常涉及依赖和 CI 调整，表明团队在管理多硬件环境时面临复杂性。
- 未来可能需要重新评估 CUDA 13.0 升级，并加强测试覆盖以避免类似回滚。