PR #25938 完整报告

sgl-project/sglang

[Revert] nvidia-cutlass-dsl[cu13] 4.5.1 -> 4.5.0

合并时间：2026-05-21 14:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25938

执行摘要

- 一句话：回退 cutlass-dsl 版本至 4.5.0
- 推荐动作：该 PR 是紧急回退，用于解阻塞 CI 和用户部署，值得快速合并。但需要立即跟进根本修复（如 PR body 中提出的 fix_cutlass_dsl_libs() 函数方案），在 main() 中根据 GPU 家族执行不同的 libs 清理逻辑。建议精读 PR body 中的问题分析和后续修复方向。

功能与动机

nvidia-cutlass-dsl[cu13] 的 cu13 *extra* 在 PyPI 上会同时安装 -libs-base 和 -libs-cu13，两者写入相同路径但内容不同，导致 Blackwell 上缺少 sm_110 架构别名（GPUModuleOp TypeError）以及非 Blackwell 的 H100 上 LoRA CUDA 图录制出现 CUDBG_EXCEPTION_WARP_ILLEGAL_ADDRESS 回归（关联 Issue #25743）。之前 #25576 版本升级到 4.5.1 未能解决此问题，因此本 PR 回退依赖以暂时解除阻塞。

实现拆解

1. 修改 python/pyproject.toml 中第 41 行依赖声明，将 "nvidia-cutlass-dsl[cu13]==4.5.1" 改回 "nvidia-cutlass-dsl[cu13]==4.5.0"。
2. 无其他文件、源码逻辑或测试配套变更。

关键文件：

- python/pyproject.toml（模块 依赖管理；类别 config；类型 configuration）：唯一的变更文件，将 nvidia-cutlass-dsl[cu13] 从 4.5.1 回退到 4.5.0。

关键符号：未识别

评论区精华

reviewer mmangkad 指出应强制重新安装 nvidia-cutlass-dsl-libs-cu13 到最后以保证顺序正确，否则仍可能出现问题。PR 作者 Kangyan-Zhou 表示先回退再尝试其他建议，并承认该错误令人困惑。

- 强制重新安装 libs-cu13 的建议 (question): PR 作者决定先回退再尝试该建议。

风险与影响

- 风险：回退到 4.5.0 后，Blackwell 上因 -libs-cu13 缺失 sm_110 别名仍可能触发 GPUModuleOp TypeError（与 #25690 的原始原因相同）；H100 LoRA 回归暂时解除，但根本的文件冲突问题未解决。CI 中已有 GPU 测试（*base-b-test--gpu-large*，*base-b-test--gpu-b200*）可以验证这两类风险。
- 影响：影响范围有限：仅涉及 pyproject.toml 中一个依赖版本号变更。对用户无直接功能影响，但 Blackwell 用户如果遇到 cutlass-dsl 相关问题，需要等待后续的完整修复。
- 风险标记：依赖冲突，已知回归

关联脉络

- PR #25576 [Deps] Use cu13 extra for nvidia cutlass dsl: 本 PR 回退了 #25576 引入的版本升级（4.5.0 → 4.5.1）。
- PR #25690 [Fix] Try to fix error caused by latest cutedsl packages: 本 PR 涉及的问题与 #25690 的动机相同（处理 cutlass-dsl libs 冲突），且 #25743 回退了 #25690。
- PR #25743 Revert #25690 to unblock LoRA Qwen3-8B CUDA graph capture on main: 本 PR 和 #25743 类似，都是因为 cutlass-dsl 库冲突导致回归而做的回退。