PR #25958 完整报告

sgl-project/sglang

[CI] Force-reinstall nvidia-cutlass-dsl-libs-cu13 last to avoid wheel-mix TypeError

合并时间：2026-05-21 22:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25958

执行摘要

- 一句话：修复 cutlass-dsl wheel 混合安装导致 CI 类型错误
- 推荐动作：本 PR 精细地定位了 wheel 混合安装的根因，并用最小改动修复，值得优先合入并关注后续 cutlass-dsl 依赖管理策略。建议回归测试 LoRA 和 eagle 测试集，确保无其他异常。作者的分析与验证过程（devbox 上测试 4 种组合、md5 校验、LoRA 回归检查）值得借鉴。

功能与动机

修复 CI 中因 nvidia-cutlass-dsl[cu13] 的 wheel 混合安装导致的 TypeError: __init__(): incompatible function arguments 错误。该错误出现在 kernel 编译时，阻塞了 CU13 CI 上的 eagle / lora 测试。PR #25743 曾回退 cutlass-dsl 版本但未根本解决，本 PR 通过控制安装顺序彻底修复。

实现拆解

1. 升级 cutlass-dsl 版本：在 python/pyproject.toml 中将 nvidia-cutlass-dsl[cu13]==4.5.0 改为 ==4.5.1。
2. 新增强制重装函数：在 scripts/ci/cuda/ci_install_dependency.sh 中添加 force_reinstall_cutlass_dsl_libs_cu13 函数，解析 pyproject.toml 中 cutlass-dsl 版本，对 CU13 运行器执行 pip install --force-reinstall --no-deps nvidia-cutlass-dsl-libs-cu13==<version>。
3. 集成到主流程：在 main() 中的 install_sglang 和 download_flashinfer_cache 之后、stabilize_flashinfer_jit_paths 之前调用该函数，确保安装顺序在后，避免 wheel 混合。非 CU13 运行器跳过。

关键文件：

- scripts/ci/cuda/ci_install_dependency.sh（模块 CI 脚本；类别 infra；类型 infrastructure；符号 force_reinstall_cutlass_dsl_libs_cu13）：核心修复：新增 force_reinstall_cutlass_dsl_libs_cu13 函数，在 install_sglang 后强制重装 -libs-cu13 以消除 wheel 混合问题。
- python/pyproject.toml（模块 项目配置；类别 config；类型 configuration）：升级 nvidia-cutlass-dsl 版本从 4.5.0 到 4.5.1，与安装顺序修复配合。

关键符号：force_reinstall_cutlass_dsl_libs_cu13

关键源码片段

scripts/ci/cuda/ci_install_dependency.sh

核心修复：新增 force_reinstall_cutlass_dsl_libs_cu13 函数，在 install_sglang 后强制重装 -libs-cu13 以消除 wheel 混合问题。

```
# 在 install_sglang 完成后调用，保证 .py 和 .so 来自同一 wheel (BOTH-cu13 状态 )
force_reinstall_cutlass_dsl_libs_cu13() {
    # 非 CU13 运行器跳过（仅 -libs-base 安装，无冲突）
    if [ "$CU_MAJOR" != "13" ]; then
        return
    fi

    # 从 pyproject.toml 解析 cutlass-dsl 版本号，保持与项目配置同步
    CUTLASS_DSL_VERSION=$(grep -Po -m1 \
        'nvidia-cutlass-dsl(\\[[^]]+\\])?==\\K[0-9A-Za-z\\.\\-]+' \
        "${REPO_ROOT}/python/pyproject.toml" || echo "")
    if [ -z "$CUTLASS_DSL_VERSION" ]; then
        echo "WARNING: could not detect nvidia-cutlass-dsl version from pyproject.toml; skipping
        libs-cu13 force-reinstall"
        return
    fi

    # 强制重装：只重装 -libs-cu13，不处理依赖，避免破坏已安装的包
    $PIP_CMD install --force-reinstall --no-deps \
        "nvidia-cutlass-dsl-libs-cu13==${CUTLASS_DSL_VERSION}" \
        $PIP_INSTALL_SUFFIX

    mark_step_done "${FUNCNAME[0]}"
}
```

评论区精华

Review 中 gemini-code-assist[bot] 建议将 pyproject.toml 版本探测路径改为使用 ${REPO_ROOT} 变量，而非相对路径，以避免脚本工作目录依赖。作者采纳该建议并在后续 commit 中修复。讨论中 mmangkad 确认安装顺序是关键，而非版本号本身。

- 使用 ${REPO_ROOT} 路径而非相对路径 (style): 作者接受建议并修改后续 commit。

风险与影响

- 风险：
  1. 回归风险低：仅影响 CU13 CI 运行环境，强制重装不修改其他依赖，且已在 devbox 上验证 LoRA 测试通过（KL 散度 2.8e-4，阈值 5e-3）。
  2. 版本同步风险：函数从 pyproject.toml 解析版本，若版本格式变更（非 == 固定版本），则可能解析失败并跳过（有警告日志）。

3. 非 CU13 环境无影响：通过 CU_MAJOR 判断跳过，不改变现有行为。 - 影响：影响范围：仅限 CU13 CI 运行器，修复了 TypeError 导致的 kernel 编译失败，使 *eagle* / *lora* 测试恢复通过。影响程度：中高，因为 CI 阻塞直接影响开发效率，但变更范围小（2 个文件）。对其他运行器无影响。 - 风险标记：仅影响 CU13 CI 运行器，版本解析可能失败（有 fallback 日志），最小变更，回归风险低

关联脉络

- PR #25938 nvidia-cutlass-dsl[cu13] 4.5.1 -> 4.5.0: 回退 cutlass-dsl 版本的尝试，但被本 PR 取代（版本本身并非根因）。
- PR #25743 Revert #25690 to unblock LoRA Qwen3-8B CUDA graph capture on main: 关联 issue：曾因 wheel 混合问题回退 #25690，本 PR 为其根本修复。