PR #25756 完整报告

sgl-project/sglang

[Fix] Fix extra uninstall of cutlass packages

合并时间：2026-05-20 01:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25756

执行摘要

- 一句话：移除 CI 中 cutlass 包的额外卸载，修复 LoRA 回归
- 推荐动作：该 PR 是一个标准的回归回退操作，展示了当临时修复引入更严重问题时如何快速止损。值得关注的是，依赖冲突问题（cutlass 包 extras 机制）并未根本解决，未来可能需要更持久的方案。

功能与动机

PR 修复了 #25743 中报告的 LoRA Qwen3-8B CUDA 图捕获回归问题。该回归由 #25690 引入，其中添加了 purge_cutlass_libs_base 来清理 cutlass 包，但导致了非法地址异常。

实现拆解

1. 在 scripts/ci/cuda/ci_install_dependency.sh 中删除 purge_cutlass_libs_base() 函数定义（减少 17 行）。
2. 在 main() 函数中移除对该函数的调用。
3. 回归原因在于 nvidia-cutlass-dsl 的 extras 机制导致 -libs-base 和 -libs-cu13 包文件冲突，之前的清理逻辑会卸载 -libs-base 并强制重装 -libs-cu13，但这可能改变了 CUDA 图捕获时的依赖状态。

关键文件：

- scripts/ci/cuda/ci_install_dependency.sh（模块 CI 脚本；类别 infra；类型 infrastructure；符号 purge_cutlass_libs_base, main）：唯一的变更文件，删除了导致回归的 purge_cutlass_libs_base 函数及其调用。

关键符号：purge_cutlass_libs_base, main

关键源码片段

scripts/ci/cuda/ci_install_dependency.sh

唯一的变更文件，删除了导致回归的 purge_cutlass_libs_base 函数及其调用。

```
# 以下为被删除的整个函数（以前位于 download_flashinfer_cache 之后）
# purge_cutlass_libs_base() { ... }
# 该函数用于卸载 nvidia-cutlass-dsl-libs-base 并强制重装 -libs-cu13
# 但该方法导致了 CUDA 图捕获时的非法地址异常（issue #25743）
```

```
# 因此在 PR #25756 中被完全移除

# 在 main() 中，对应的调用行也已删除：
# main() {
# ...
# purge_cutlass_libs_base # 此行被删除
# stabilize_flashinfer_jit_paths
# ...
# }
```

评论区精华

审核机器人（gemini-code-assist）在早期提交中建议将 flash-attn-4 的版本约束改为带 extra 的固定版本以确保稳定性，但该建议涉及 pyproject.toml，而最终 PR 并未修改该文件，因此讨论未实际落地。

- flash-attn-4 版本约束建议 (other): 该建议针对的是中间提交，最终 PR 未修改 pyproject.toml，因此未采纳。

风险与影响

- 风险：风险极低：变更仅涉及 CI 安装脚本，回退了一个已被证明有问题的修复。若 cutlass 包冲突在未来再次出现，需采用不同方案解决。
- 影响：影响范围：CI 构建流程。LoRA Qwen3-8B 的测试重新通过，恢复了 main 分支的稳定性。对用户无直接影响。
- 风险标记：回归修复，依赖冲突未根本解决

关联脉络

- PR #25743 Revert #25690 to unblock LoRA Qwen3-8B CUDA graph capture on main: 该 issue 报告了回归并触发本 PR 的回退操作。
- PR #25690 [Fix] Try to fix error caused by latest cutedsl packages: 原始修复 PR，其引入的 purge_cutlass_libs_base 导致了回归，被本 PR 回退。