PR #23587 完整报告

sgl-project/sglang

ci: fix cu129 wheel tagging + pipefail-abort in install script (follow-up to #23497)

合并时间：2026-04-24 05:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23587

执行摘要

- 一句话：修复 #23497 引入的 cu129 轮子标签缺失和安装脚本 pipefail 问题
- 推荐动作：作为 #23497 的跟进修复，本 PR 改动小但关键，建议审阅者关注 pipefail 调优模式的使用。修改内容合理且 PR 描述清晰，可直接合并。

功能与动机

PR #23497 在 sgl-kernel 构建矩阵中恢复 cu129 后，H 系列 (cu129) 测试机无法加载正确轮子。具体表现为 #21985 的 CI 中 Install dependencies 步骤因 pipefail 退出，且 cu129 构建的轮子缺少 +cu129 标签，导致脚本找不到匹配轮子而静默回退到主分支轮子。

实现拆解

1. 修复轮子标签识别（sgl-kernel/rename_wheels.sh）：在 detect_cuda_suffix() 函数中添加 12.9 → +cu129 分支，确保 cu129 构建的轮子文件名包含 +cu129 后缀。
2. 修复 pipefail 导致脚本提前退出（scripts/ci/cuda/ci_install_dependency.sh）：在轮子查找管道命令后添加 || true，使 ls 无匹配时返回空字符串而非退出码 2，从而让后续的 [-z "$KERNEL_WHL"] 错误检查正常触发并输出清晰错误。
3. 移除无标签轮子回退逻辑（同上脚本）：删除 #23497 中添加的 fallback 逻辑，该逻辑允许安装不带 CUDA 标签的轮子，但这可能引入 ABI 不兼容或轮子被 PyPI 分支覆盖的隐患。现在仅允许版本号完全匹配 +cuXYZ 的轮子，否则显式报错。

关键文件：

- scripts/ci/cuda/ci_install_dependency.sh（模块 CI 脚本；类别 infra；类型 infrastructure）：修复安装脚本中 pipefail 导致的提前退出及移除潜在危险的无标签轮子回退逻辑
- sgl-kernel/rename_wheels.sh（模块 内核构建；类别 other；类型 core-logic）：添加 cu129 分支以保证 CUDA 12.9 构建的轮子获得正确的 +cu129 标签

关键符号：detect_cuda_suffix

关键源码片段

scripts/ci/cuda/ci_install_dependency.sh

修复安装脚本中 pipefail 导致的提前退出及移除潜在危险的无标签轮子回退逻辑

```
# KERNEL_WHL=$(ls ... 2>/dev/null | head -1) # 原代码，无匹配时 ls 返回 2 导致 set -e 退出
# `|| true` swallows `ls`'s exit-2-on-no-match so `set -o pipefail` doesn't abort the
# script before we reach the explicit error check.
KERNEL_WHL=$(ls sgl-kernel/dist/sglang_kernel-${SGL_KERNEL_VERSION_FROM_KERNEL}+
${CU_VERSION}-cp310-abi3-manylinux2014_${WHEEL_ARCH}.whl 2>/dev/null | head -1 || true)
if [ -z "$KERNEL_WHL" ]; then
  echo "ERROR: No matching sgl-kernel wheel found ..."
  ls -alh sgl-kernel/dist/
  exit 1
fi
```

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：本 PR 仅涉及 CI 脚本，风险较低。主要风险：修改 rename_wheels.sh 可能影响其他 CUDA 版本（12.4/12.8/13.0）的轮子命名，但逻辑与原有分支一致；移除 fallback 后，如果仍有老旧分支构建无标签轮子，安装会失败，但这种情况应极少且已有明确错误提示。
- 影响：影响范围：仅影响 cu129（H20）测试机的 CI 流程，其他 CUDA 版本无行为变化。影响程度：修复后 cu129 测试机能够正确识别和使用 PR 构建的 sgl-kernel 轮子，避免静默回退到主分支轮子导致的测试假阴性。
- 风险标记：CI 脚本变更，非核心逻辑

关联脉络

- PR #23497 ci: build sgl-kernel wheels for both cu129 and cu130: 本 PR 的动机和修改直接源于 #23497 引入的两个回归问题，是对 #23497 的跟进修复
- PR #21985 perf: eliminate attention DtoD copy by passing pre-allocated output to FA: PR body 中引用了 #21985 的 CI 失败作为问题案例，修复后该 PR 的 CI 有望通过
- PR #22392 perf: eliminate nvjet memset bubbles via CUTLASS FP8 GEMM: 同为涉及 sgl-kernel 的 PR，本修复确保这类 PR 的 CI 不会因为轮子问题失败