PR #23497 完整报告

sgl-project/sglang

ci: build sgl-kernel wheels for both cu129 and cu130

合并时间：2026-04-23 09:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23497

执行摘要

- 一句话：修复 CI 中 sgl-kernel wheel 构建矩阵，同时支持 cu129 和 cu130 以避免测试静默失败。
- 推荐动作：建议团队在修改 sgl-kernel 时关注此 PR，确保 CI 能正确测试内核变更。对于 CI 维护者，这是一个重要的配置修复，值得了解 wheel 选择逻辑和兼容性处理。

功能与动机

PR body 指出，H 系列测试运行器使用 NVIDIA 驱动 535.x，仅支持 CUDA ≤ 12.x，无法加载 cu130-built wheels。当 PR 修改 sgl-kernel 时，测试会静默安装公共 main 分支 wheel，导致内核变更被丢弃，引发 TypeError 等错误。例如 #21985 添加了 out= 参数，但由于 wheel 不匹配而失败，凸显了修复必要性。

实现拆解

1. 扩展构建矩阵：在 .github/workflows/pr-test.yml、pr-test-multimodal-gen.yml 和 pr-test-sgl-kernel.yml 中，将 sgl-kernel-build-wheels 作业的矩阵从仅 cuda-version: "13.0" 扩展为同时包含 "13.0" 和 "12.9"，确保 PR 构建 cu129 和 cu130 两种 wheel。
2. 修改下载模式：在上述 workflow 文件中，将下载 artifact 的 pattern 从 wheel-python3.10-cuda13.0 改为 wheel-python3.10-cuda*，以匹配所有 CUDA 版本 wheel。
3. 优化安装脚本：在 scripts/ci/cuda/ci_install_dependency.sh 中，调整 wheel 选择逻辑，优先选择文件名匹配 $CU_VERSION（如 +cu129 或 +cu130）的 wheel，并添加回退逻辑以兼容旧分支。
4. 无测试或部署配套改动：这是纯 CI 基础设施变更，不影响运行时行为，无需额外测试。

关键文件：

- scripts/ci/cuda/ci_install_dependency.sh（模块 安装脚本；类别 infra；类型 infrastructure）：核心安装脚本，负责选择匹配 CUDA 版本的 sgl-kernel wheel，避免静默回退到公共 wheel，是修复问题的关键。
- .github/workflows/pr-test.yml（模块 CI 流水线；类别 infra；类型 infrastructure）：主测试 workflow，修改构建矩阵和下载模式以支持双 CUDA 版本 wheel，影响所有相关测试作业。

关键符号：未识别

关键源码片段

scripts/ci/cuda/ci_install_dependency.sh

核心安装脚本，负责选择匹配 CUDA 版本的 sgl-kernel wheel，避免静默回退到公共 wheel，是修复问题的关键。

```
# Wheel filenames carry a +cuXYZ local version tag (e.g. sglang_kernel-0.4.0+cu130-...).
# 当构建矩阵产生多个 CUDA 版本 wheel 时，选择匹配测试运行器 $CU_VERSION 的 wheel，
# 以避免触发 PyPI 回退重安装（这会用公共 main 分支 wheel 替换 PR 构建的 wheel）。
KERNEL_WHL=$(ls sgl-kernel/dist/sglang_kernel-${SGL_KERNEL_VERSION_FROM_KERNEL}+
${CU_VERSION}-cp310-abi3-manylinux2014_${WHEEL_ARCH}.whl 2>/dev/null | head -1)
if [ -z "$KERNEL_WHL" ]; then
    # 回退逻辑：对于旧分支仅构建单版本 wheel 的情况，选择无 +cuXYZ 标签的 wheel。
    # 限制相同架构，避免选择不匹配 CUDA 版本的 wheel，防止静默替换。
    SINGLE_CUDA_WHL=$(ls sgl-kernel/dist/sglang_kernel-${SGL_KERNEL_VERSION_FROM_
    KERNEL}-cp310-abi3-manylinux2014_${WHEEL_ARCH}.whl 2>/dev/null | head -1)
    if [ -n "$SINGLE_CUDA_WHL" ]; then
        KERNEL_WHL="$SINGLE_CUDA_WHL"
    fi
fi
if [ -z "$KERNEL_WHL" ]; then
    echo "ERROR: No matching sgl-kernel wheel found in sgl-kernel/dist/ for version ${SGL_
    KERNEL_VERSION_FROM_KERNEL} arch ${WHEEL_ARCH} cuda ${CU_VERSION}"
    ls -alh sgl-kernel/dist/
    exit 1
fi
```

评论区精华

无 review 评论，但 PR body 中作者详细解释了问题根源和解决方案，强调静默失败的风险："H-series CI boxes still ship NVIDIA driver 535.x... which supports CUDA ≤ 12.x — it cannot load cu130-built wheels at all." 并指出修复后测试将正确运行 PR 的内核变更。

- 静默失败问题 (correctness): 通过构建双版本 wheel 和优化安装脚本来修复，确保测试正确运行 PR 的内核变更。

风险与影响

- 风险：风险较低：主要风险是 CI 配置错误可能导致 wheel 选择失败，但脚本添加了回退逻辑以兼容旧分支。兼容性风险：确保旧分支（仅构建单版本 wheel）仍能工作。性能风险：增加一个矩阵作业，延长 CI 时间约 10 分钟，但这是必要的成本以修复测试静默失败。
- 影响：对用户无直接影响，仅影响 CI 流程。对团队：修复了测试静默失败问题，确保 sgl-kernel 变更在 H 系列运行器上正确测试，避免假阳性结果。系统：提升 CI 可靠性，确保内核修改在混合驱动环境中得到验证。
- 风险标记：CI 配置变更，兼容性风险，测试静默失败修复

关联脉络

- PR #23119 未知（未在提供的历史 PR 列表中）：PR body 提及 #23119 将 sgl-kernel-build-wheels 矩阵改为仅 CUDA 13.0，是本 PR 修复的根源。
- PR #21985 未知（未在提供的历史 PR 列表中）：PR body 引用 #21985 作为示例，其中内核变更因 wheel 不匹配而失败，凸显了本 PR 修复的必要性。