

PR #23443 完整报告

sgl-project/sglang

[diffusion] ci: allow using prebuilt sgl-kernel wheel for GT regeneration

合并时间: 2026-04-22 14:44

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/23443>

执行摘要

- 一句话: 为扩散模型 GT 生成 CI 工作流添加预构建 sgl-kernel wheel 支持, 解决 PyPI wheel ABI 不兼容问题。
- 推荐动作: 该 PR 主要涉及 CI 基础设施, 对核心功能无影响。建议 CI 维护者关注其实现模式, 可作为处理类似 ABI 不兼容问题的参考。普通开发者无需深入阅读。

功能与动机

PR body 中明确指出, 当分支升级 torch 时, PyPI 发布的 sglang-kernel wheel 会因 ABI 不兼容而加载失败 (错误: undefined symbol: _ZN3c104cuda29c10_cuda_check_implementationEiPKcS2_ib)。这导致在 torch 升级分支上无法重新生成扩散模型的 GT 基准。该 PR 通过允许使用预构建的 wheel 来解决此问题。

实现拆解

1. 添加输入参数: 在 .github/workflows/diffusion-ci-gt-gen.yml 的 workflow_dispatch 部分新增可选输入 kernel_artifact_run_id, 用于指定预构建 wheel 的来源运行 ID。
2. 条件化下载步骤: 在三个作业 (1-gpu-h100、2-gpu-h100、4-gpu-b200) 的 Install dependencies 步骤前, 新增两个条件步骤: Prepare sgl-kernel/dist for prebuilt wheel (清空目录) 和 Download prebuilt sgl-kernel wheel (使用 actions/download-artifact 下载指定运行中的 wheel-python3.10-cuda13.0 工件)。
3. 控制安装逻辑: 修改 Install dependencies 步骤, 当 kernel_artifact_run_id 设置时, 传递 CUSTOM_BUILD_SGL_KERNEL=true 环境变量, 触发 scripts/ci/cuda/ci_install_dependency.sh 中已有的预构建 wheel 安装路径。
4. 保持向后兼容: 输入为空时, 工作流行为不变, 仍从 PyPI 安装 sglang-kernel。

关键文件:

- .github/workflows/diffusion-ci-gt-gen.yml (模块 CI 工作流; 类别 infra; 类型 infrastructure): 这是唯一变更的文件, 实现了整个功能: 添加输入参数、条件化下载和安装步骤。

关键符号: 未识别

评论区精华

无 review 评论，PR 由作者直接合并。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 依赖过期风险：预构建 wheel 来自其他 CI 运行，工件默认 90 天后过期，若过期后使用会导致下载失败。
 2. ABI 匹配风险：要求 kernel_artifact_run_id 必须与当前分支提交相同，若提交不匹配可能导致 wheel 与当前环境 ABI 不兼容。
 3. 配置错误风险：用户可能错误设置 kernel_artifact_run_id 或未确保提交一致，导致工作流失败。
 4. 维护负担：新增的输入参数和条件步骤增加了工作流配置的复杂性。
- 影响：
 1. 对用户影响：为开发者在 torch 升级等场景下重新生成扩散 GT 基准提供了可行路径，提升了开发灵活性。
 2. 对系统影响：仅影响 CI 工作流，不改变运行时行为或核心代码。
 3. 对团队影响：简化了在 ABI 不兼容分支上的 GT 维护流程，减少了阻塞。 - 风险标记：依赖外部工件，ABI 兼容性要求，配置复杂性增加

关联脉络

- PR #23422 [diffusion] support custom output folder name in GT generation workflow: 同属扩散模型 GT 生成 CI 工作流改进，涉及同一文件 .diffusion-ci-gt-gen.yml 的修改。
- PR #23417 [ci] split stage-c-test-4-gpu-b200 to enable a low-disk runner pool: 同属 CI 基础设施优化，涉及 CI 工作流配置和资源管理。
- PR #22830 ci: enable /rerun-test for nightly test suites: 同属 CI 工具能力扩展，涉及 GitHub Actions 工作流和脚本修改。