

PR #34254 完整报告

sgl-project/sglang

[NPU] Modified kernel tag version to 8.10

合并时间: 2026-08-10 19:40

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34254>

执行摘要

- 一句话: NPU 内核包 tag 更新至 2026.8.10
- 推荐动作: 值得快速浏览而非精读。可作为 NPU 发布流程中“CI 与镜像依赖版本同步”的典型案例: 三处 tag 必须保持一致。阅读时可留意版本号表述不一致 (body 7.24、评论 7.27、实际 2026.7.27) 对后续追溯的影响, 以及是否应把发行版本集中到单一配置源或增加自动化校验。

功能与动机

PR body 原文: “Update the tag version from 7.24 to 8.10 to align with the latest release version.” 即让 NPU 场景使用的 sgl-kernel-npu 预编译内核包与官方最新 Release 对齐, 避免 CI 与发布镜像继续使用旧版内核。需要指出: body 中写的 7.24 与实际补丁不一致, 真实改动是 2026.7.27 → 2026.8.10。

实现拆解

1. 更新正式发行镜像构建参数: 在 .github/workflows/release-docker-npu.yml 的 build-args 中, 将 SGLANG_KERNEL_NPU_TAG 从 2026.7.27 改为 2026.8.10, 同时保留 CANN_VERSION、DEVICE_TYPE、SGLANG_TAG 等参数。
2. 同步更新 nightly 发行镜像: 在 .github/workflows/release-docker-npu-nightly.yml 中做相同修改, 保证每日构建与正式发布使用同一内核版本。
3. 同步更新 CI 依赖安装: 在 scripts/ci/npu/npu_ci_install_dependency.sh 中把 SGLANG_KERNEL_NPU_TAG 变量改为 "2026.8.10", 该变量用于拼接 GitHub Release 下载 URL 与 zip 包名。
4. 配套测试与验证: 无新增测试。PR body 声明不涉及精度与性能测试; 实际有效性依赖新 tag 的 Release 存在且与当前环境 (py311、CANN 9.0.0) 匹配。

关键文件:

- .github/workflows/release-docker-npu.yml (模块 构建流水线; 类别 infra; 类型 configuration; 符号 SGLANG_KERNEL_NPU_TAG): 正式 NPU 发行镜像的构建配置, build-args 中 SGLANG_KERNEL_NPU_TAG 直接决定镜像内安装的 sgl-kernel-npu 内核包版本, 是本次升级的核心入口。
- .github/workflows/release-docker-npu-nightly.yml (模块 构建流水线; 类别 infra; 类型 configuration; 符号 SGLANG_KERNEL_NPU_TAG): nightly NPU 镜像构建与正式发布

共用同一内核版本约束，若不同步更新会导致每日构建与正式发布使用不同内核包。

- `scripts/ci/npu/npu_ci_install_dependency.sh`（模块 依赖安装；类别 `infra`；类型 `configuration`；符号 `SGLANG_KERNEL_NPU_TAG`）：NPU CI 依赖安装脚本通过该 tag 拼接下载 URL 并解压预编译内核包，直接影响 CI 环境能否获得新版本内核，是 CI 侧的关键修改点。

关键符号：未识别

关键源码片段

[.github/workflows/release-docker-npu.yml](#)

正式 NPU 发行镜像的构建配置，`build-args` 中 `SGLANG_KERNEL_NPU_TAG` 直接决定镜像内安装的 `sgl-kernel-npu` 内核包版本，是本次升级的核心入口。

```
# .github/workflows/release-docker-npu.yml
# NPU 发行镜像构建：通过 build-args 注入 sgl-kernel-npu 的版本 tag。
# 该值与 scripts/ci/npu/npu_ci_install_dependency.sh 中的 SGLANG_KERNEL_NPU_TAG
# 必须保持一致，
# 否则 CI 与发布镜像会使用不同版本的内核包。
build-args: |
  SGLANG_KERNEL_NPU_TAG=2026.8.10
  CANN_VERSION=${{ matrix.cann_version }}
  DEVICE_TYPE=${{ matrix.device_type }}
  SGLANG_TAG=${{ steps.version.outputs.version }}
```

[scripts/ci/npu/npu_ci_install_dependency.sh](#)

NPU CI 依赖安装脚本通过该 tag 拼接下载 URL 并解压预编译内核包，直接影响 CI 环境能否获得新版本内核，是 CI 侧的关键修改点。

```
# scripts/ci/npu/npu_ci_install_dependency.sh
# 指定 sgl-kernel-npu 预编译包的 Release tag，升级到 2026.8.10 以对齐最新发布版本。
# 该变量会被拼接进下载 URL 与 zip 包名，因此必须与 GitHub Releases 中实际存在的 tag
# 完全一致，
# 同时需匹配当前 CI 环境（py311、CANN 9.0.0、DEVICE_TYPE 与架构）。
SGLANG_KERNEL_NPU_TAG="2026.8.10"

# 下载并解压对应 tag 的预编译内核包
mkdir sgl-kernel-npu
(cd sgl-kernel-npu && wget "${GITHUB_PROXY_URL:=https://github.com/sgl-project/sgl-kernel-npu/releases/download/${SGLANG_KERNEL_NPU_TAG}/sgl-kernel-npu-${SGLANG_KERNEL_NPU_TAG}-torch${PYTORCH_VERSION}-py311-cann9.0.0-${DEVICE_TYPE}-${arch}.zip" \
  && unzip ./sgl-kernel-npu-${SGLANG_KERNEL_NPU_TAG}-torch${PYTORCH_VERSION}-py311-cann9.0.0-${DEVICE_TYPE}-${arch}.zip)
```

评论区精华

唯一的 review 评论来自作者本人的 self-review，在 [.github/workflows/release-docker-npu.yml](#) 第 1 行注明“Update the tag version from 7.27 to 8.10”，与 PR body 中“7.24 →

8.10” 的表述不一致；结合补丁内容可确认实际版本为 2026.7.27 → 2026.8.10，body 中的 7.24 为笔误。该差异未被其他 reviewer 讨论或要求修正，合并时以代码为准。

- 版本号表述不一致：body 写 7.24，评论写 7.27 (question)：没有引发技术讨论；合并时以代码为准，即 2026.7.27 → 2026.8.10，body 中的 7.24 属于笔误。

风险与影响

- 风险：

1. 外部 Release 可用性风险：npu_ci_install_dependency.sh 会直接下载 sgl-kernel-npu-2026.8.10-...zip，若该 Release tag 尚未发布或包体损坏，NPU CI 将整体失败；两个 Docker workflow 的 build-args 同样会在镜像构建阶段因拉取失败而中断。
2. 版本同步维护风险：CI 脚本与两个 workflow 共 3 处 tag 需要人工同步修改，本次 PR 已覆盖全部 3 处，但后续升级仍依赖人工，缺少自动化校验或单一配置源。
3. 兼容性风险：新内核包若依赖更高的 CANN 版本或改变算子行为，可能影响 NPU 精度与性能；PR 未提供相关验证，属纯 tag 升级，风险较低。- 影响：影响范围集中在 NPU 通道：日常与 nightly 的 NPU Docker 镜像发布，以及 NPU CI 的依赖安装阶段。影响程度中等偏低——新 tag 可用时对用户无感知；若新 tag 不可用，将阻断 NPU 镜像发布与 CI 流水线。对团队而言，NPU 维护者需确认 2026.8.10 的 sgl-kernel-npu Release 与当前 CANN、设备组合的兼容性，并保持 CI 与镜像两处版本引用一致。- 风险标记：依赖外部 Release tag，版本号表述不一致，CI 与镜像版本人工同步

关联脉络

- PR #34253 [CI] Add output_tag input to the Patch Docker Image workflow: 同属 Docker 镜像发布 workflow 的配置演进，反映镜像发布通道近期集中调整。
- PR #34103 [NPU] Increase the retry count to 3 for the GSM8K.: 同属 NPU CI 稳定性维护线，与本 PR 的 NPU 版本升级相互补充。
- PR #34231 [CI] Keep the torch compilation cache instead of wiping it on install: 同属 CI 依赖安装脚本维护，本 PR 与其处于同一 scripts/ci 目录下的平行改动。