

PR #37073 完整报告

sgl-project/sglang

Use Flashinfer 0.6.18 release for CUDA 13.4 package

合并时间: 2026-08-30 12:06

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37073>

执行摘要

- 一句话: cu134 镜像改用 FlashInfer 0.6.18 正式版
- 推荐动作: 不值得精读, 变更机械简单, 读完 diff 即可。可作为依赖跟进型基础设施 PR 的参考样本: 在预发布产物被上游删除后, 如何快速把版本 ARG 切到正式版并顺带对齐其他 Dockerfile 的能力; 值得注意的是 ai-dynamo 使用 --pre 不定版本安装带来的可复现性问题, 后续可考虑固定版本号。

功能与动机

PR body 明确说明: Flash 0.6.18 was officially released and the rc10 wheels got deleted, so the last build failed, 因此需要直接使用 0.6.18 release 替代 rc10; 同时 bring in dynamo install from main Dockerfile, 使 cu134 镜像与主 Dockerfile 的运行时能力对齐, 避免后续因 AI Dynamo 缺失而出现新的构建或运行问题。

实现拆解

变更入口是 docker/Dockerfile.cu134, 共两处修改:

1. FlashInfer 版本 ARG 固化: 将 FLASHINFER_VERSION 从 0.6.18rc10 改为 0.6.18。该 ARG 在镜像构建阶段决定安装的 FlashInfer 版本, 切到正式版后构建不再依赖已被上游删除的 rc10 产物, 直接修复上次构建失败。
2. 新增 AI Dynamo 安装步骤: 在 CUDA 13 分支的 Python 依赖安装段之后追加一个 RUN 步骤, 使用 --pre --extra-index-url <https://pypi.nvidia.com/> 从 NVIDIA 官方索引安装 ai-dynamo。该步骤与 main Dockerfile 保持一致, 补齐 cu134 镜像的 Dynamo 运行时能力, 属于能力对齐而非新功能。

配套情况: 无测试、无文档、无配置迁移改动。此类镜像构建变更主要依赖镜像构建 CI 验证, 本 PR 的 PR Test (Base) 已通过。

关键文件:

- docker/Dockerfile.cu134 (模块 镜像构建; 类别 infra; 类型 dependency-pin; 符号 FLASHINFER_VERSION): 唯一变更文件, 同时承载 FlashInfer 版本固化与 AI Dynamo 安装步骤补充, 直接决定 CUDA 13.4 镜像能否构建成功。

关键符号: 未识别

关键源码片段

docker/Dockerfile.cu134

唯一变更文件，同时承载 FlashInfer 版本固化与 AI Dynamo 安装步骤补充，直接决定 CUDA 13.4 镜像能否构建成功。

```
# 变更点 1: FlashInfer 版本 ARG 从 0.6.18rc10 固化为 0.6.18 正式版。
# 0.6.18 正式发布后，rc10 预发布 wheel 已从源中被删除，
# 继续引用 rc10 会导致 CUDA 13.4 镜像构建失败。
ARG FLASHINFER_VERSION=0.6.18

# 变更点 2: 与 main Dockerfile 对齐，为 CUDA 13 分支补装 AI Dynamo。
# 使用 NVIDIA 官方索引，并接受预发布版本。
# 注意: --pre 不固定版本号，镜像构建结果会随上游发布漂移。
RUN --mount=type=cache,target=/root/.cache/pip \
    python3 -m pip install --pre --extra-index-url https://pypi.nvidia.com/ ai-dynamo
```

评论区精华

本 PR 没有任何 Review 评论或讨论线程，Fridge003 在 CI 通过后直接批准合并，说明该变更意图清晰、风险认知一致，未产生设计争议或未解决疑虑。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 可复现性风险：ai-dynamo 使用 --pre 不定版本安装，镜像每次构建可能拉到不同的预发布版本，构建结果随时间漂移；同时 pypi.nvidia.com 的可用性与网络可达性成为构建链路的对外依赖。
 - 行为差异风险：FlashInfer 0.6.18 正式版与 0.6.18rc10 之间若存在行为差异，cu134 镜像需要重新验证；不过主 Dockerfile 已在 PR #36954 中先行升级到正式版，相关风险已在前置环节消化，本 PR 属于同步跟进。
 - 测试缺口：无配套测试，cu134 镜像的实际可用性只能靠后续镜像构建 CI 和用户使用反馈暴露。
 - 影响：影响范围集中在 CUDA 13.4 镜像的构建与发布链路，不涉及运行期推理逻辑。对用户而言，修复了 cu134 镜像构建失败的问题，并让该镜像具备与主镜像一致的 AI Dynamo 能力；对团队而言，消除了 rc10 产物删除造成的构建阻塞，是一条低风险的基础设施修复。
- 风险标记：预发布依赖漂移，无测试覆盖，外部索引依赖

关联脉络

- PR #36954 [Deps] Bump FlashInfer to 0.6.18: 同一 FlashInfer 0.6.18 依赖升级主线：该 PR 完成主 Dockerfile、pyproject 与 Kimi-K3 定制镜像的升级，本 PR 是其 cu134 专属 Dockerfile 上的同步收尾。

- PR #36985 test: re-enable FlashInfer per-token NVFP4 coverage: FlashInfer 0.6.18 升级后的回归验证环节，恢复 Blackwell NVFP4 场景下的精度测试，共同构成 0.6.18 的完整落地。