

PR #37083 完整报告

sgl-project/sglang

[AMD] Cherry-pick dsv4 fp4 kv-cache fix aiter commit

合并时间: 2026-08-30 12:17

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37083>

执行摘要

- 一句话: ROCm 镜像 cherry-pick aiter 提交, 修复 v4 fp4 kv-cache 问题
- 推荐动作: 该 PR 改动极小 (+2/-0、单文件), 技术含量不高, 但体现了一个值得注意的维护模式: 在依赖 aiter 的 Dockerfile 构建链中以 cherry-pick 方式提前下放上游补丁。适合快速浏览确认构建链语义, 无需精读; 建议合并后登记跟踪项, 确保 aiter 升级时清理。

功能与动机

PR body 明确说明此变更目的为 Cherry-pick aiter#5034 for v4 fp4 indexer, 即把 aiter 上游的 kv cache 正确性修复提前带入 ROCm 镜像, 使 AMD 用户在启用 `--enable-deepseek-v4-fp4-indexer` 时获得修复后的行为。作者同时提交了 GSM8K 精度与吞吐数据作为依据, 证明修复有效且无回归。

实现拆解

实现可拆解为以下几步:

1. 变更入口: 唯一改动在 `docker/rocm.Dockerfile` 的 aiter 构建段。原流程在 `git checkout -f ${AITER_COMMIT}` 之后直接执行 `git submodule update --init --recursive`, 本 PR 在其间插入 `git cherry-pick --no-commit 8578af153f4fa1e007fede7e3c1e1b373f07af4c`, 把上游修复应用到当前提交之上。
2. 采用 `--no-commit` 的原因: 镜像构建每次从全新 clone 开始 (前面已有 `git checkout -f` 强制切到指定提交), 工作树没有真实用户改动, 因此不需要在 aiter 仓库内额外生成提交记录, 只需将补丁内容落到工作树即可; 后续 `git submodule update` 与 `pip install -r requirements.txt` 均基于补丁后的代码执行。
3. 可维护性标注: 新增注释说明这是临时 cherry-pick, 下一次 aiter 升级时可移除, 避免后续维护者面对一个来源不明的固定提交而困惑。
4. 验证方式: 作者用 `rocm/sgl-dev:v0.5.18-rocm720-mi35x-20260829` 镜像, cherry-pick 后手动删除旧的 JIT 缓存产物 (`aiter/jit/module_dsv4_rotate_quant.so` 及 `build` 目录), 再用 PR#36581 的 fp4 indexer 分支跑 GSM8K, 得到 Accuracy 0.954、Latency 108.931 s、Output throughput 1070.266 token/s; 无 indexer 基线为 0.951、115.289 s、1018.012 token/s。
5. 测试与 CI 配套: 本 PR 没有新增单元测试; CI 中 Base 一条为绿色, Extra 与 AMD ROCm 7.2 两条为失败状态, 需要人工确认是否为 ROCm 镜像构建或既知环境问题。

关键文件：

- `docker/rocm.Dockerfile`（模块 镜像构建；类别 `infra`；类型 `infrastructure`）：唯一变更文件：在 ROCm 镜像的 aiter 构建链中插入 `git cherry-pick --no-commit 8578af1`，将 v4 fp4 indexer 的 kv cache 修复下放到当前 `AITER_COMMIT` 工作树，并加注释说明该步骤可随 aiter 升级移除。

关键符号：未识别

关键源码片段

`docker/rocm.Dockerfile`

唯一变更文件：在 ROCm 镜像的 aiter 构建链中插入 `git cherry-pick --no-commit 8578af1`，将 v4 fp4 indexer 的 kv cache 修复下放到当前 `AITER_COMMIT` 工作树，并加注释说明该步骤可随 aiter 升级移除。

```
# 先卸载 aiter，避免早于注释规则的 AITER_COMMIT 覆盖失效；
# 当前工作树来自刚执行的 git clone，没有真实用户改动需要保留。
# 临时 cherry-pick aiter 上游 8578af1 提交（对应 aiter#5034），
# 修复 v4 fp4 indexer 的 kv cache 问题，下次 aiter 升级后可移除。
RUN pip uninstall -y aiter \
    && git clone ${AITER_REPO} \
    && cd aiter \
    && git checkout -f ${AITER_COMMIT} \
    && git cherry-pick --no-commit 8578af153f4fa1e007fede7e3c1e1b373f07af4c \
    && git submodule update --init --recursive \
    && pip install -r requirements.txt
```

评论区精华

该 PR 没有产生任何 review 评论或讨论线程，审核人 HaiShaw 直接给予 APPROVED。唯一实质信息来自 PR body 中的验证数据与注释表述：注释用 `may be removed in next aiter upgrade` 明确说明这是过渡性补丁，需要跟踪上游合入状态。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 临时补丁冲突风险：8578af1 是针对当前 `AITER_COMMIT` 的补丁；一旦后续更新 `AITER_COMMIT`，该 cherry-pick 可能产生冲突，导致 ROCm 镜像构建直接失败。
 - 重复应用风险：若 aiter 上游新版本已合入等价修复，构建时仍会再次 cherry-pick，可能出现冲突或重复改动；需要维护者跟踪 aiter#5034 的合入进度并及时删掉这一行。
 - 影响范围：改动只作用于 `docker/rocm.Dockerfile` 构建产物，不涉及已发布 wheel 或运行时源码；风险集中在 ROCm 镜像构建链路。
 - CI 状态：PR Check Extra 与 AMD ROCm 7.2 两条流水线为失败状态，合并前需确认失败原因与本补丁无关。

- 影响：

- 用户侧：AMD ROCm 上启用 v4 fp4 indexer 的用户将获得修复后的 kv cache 行为；按 PR 数据，GSM8K 准确率 $0.951 \rightarrow 0.954$ ，端到端延迟 $115.289\text{ s} \rightarrow 108.931\text{ s}$ （约 -5.5%），输出吞吐 $1018.012 \rightarrow 1070.266\text{ token/s}$ （约 +5.1%）。
- 系统侧：改动只进入镜像构建链，已运行的线上服务不受影响；新拉取的 ROCm 镜像才包含该修复。
- 团队侧：需要建立跟踪项，在下次 aiter 升级时核对 aiter#5034 是否已包含，并及时移除本行，否则后续构建可能出现冲突或重复修复。
- 风险标记：临时 cherry-pick 依赖上游提交，aiter 升级后需移除，仅影响 ROCm 镜像构建

关联脉络

- PR #36581 [AMD] v4 fp4 indexer 功能相关 PR（PR body 中引用）：PR body 明确用它作为验证分支：git fetch origin pull/36581/head:pr-36581 后启用 --enable-deepseek-v4-fp4-indexer，本补丁的价值建立在该功能之上的 kv cache 修复。
- PR #37132 [AMD] Fix the QuickReduce bf16 cast failing to build for CDNA: 同为 AMD 侧 kernel/ 构建链路修复，说明近期 AMD + 构建链是活跃修复线，与本 PR 同属 ROCm 镜像质量保障。
- PR #36985 test: re-enable FlashInfer per-token NVFP4 coverage: 同属 FP4 量化精度保障脉络，但平台在 Blackwell；与本 PR 关联较间接，可佐证 FP4 精度与 kv cache 路径是当前重点验证对象。