

PR #34818 完整报告

sgl-project/sglang

[CI] Install sgl-eval in xeon (CPU) Docker image

合并时间: 2026-08-17 14:00

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34818>

执行摘要

- 一句话: Xeon 镜像补装 sgl-eval, 修复 CPU CI 测试失败
- 推荐动作: 值得快速了解。该 PR 体现了跨 CI 平台共享依赖版本固定脚本的实践, 避免同一依赖在多个 Dockerfile 中各自维护版本。后续维护 CPU 或其他镜像的 CI 依赖时, 可参照此模式。

功能与动机

PR body 明确指出: "PR #34477 routed mmlu eval through the external `sgl-eval` CLI and added the install step to the cuda / amd / npu CI dependency scripts, but the CPU path builds its image from docker/xeon.Dockerfile, which was not updated. As a result test_mmlu_torch_compile_cpu (test/registered/cpu/test_cpu_graph.py) fails on the xeon-gnr runner with `FileNotFoundError: 'sgl-eval'.`" 即 CPU 镜像缺少 sgl-eval 导致相关测试失败, 需要对齐其他平台的安装方式。

实现拆解

1. 定位问题: PR #34477 引入 sgl-eval 依赖, 并更新了 CUDA/AMD/NPU 的 CI 依赖脚本, 但 docker/xeon.Dockerfile 未同步, 导致 CPU 镜像中缺少 sgl-eval 命令。
2. 修改镜像构建脚本: 在 docker/xeon.Dockerfile 的 RUN 链中, 追加 `&& source /sgl-workspace/sglang/scripts/ci/utils/sgl_eval_ref.sh && uv pip install "$SGL_EVAL_SPEC"`, 在镜像构建阶段完成 sgl-eval 安装。
3. 复用共享版本锁定: 通过 scripts/ci/utils/sgl_eval_ref.sh 获取统一的 \$SGL_EVAL_SPEC 版本号, 使 CPU 与 CUDA/AMD/NPU 镜像的 sgl-eval 版本保持一致, 后续升级只需修改该脚本一处。
4. 验证: PR 描述中的 CI 状态显示 PR 测试已通过 (Run #31776932514、Run #31999955409), 说明镜像构建与依赖它的 CPU 测试均正常。

关键文件:

- docker/xeon.Dockerfile (模块 镜像构建; 类别 infra; 类型 infrastructure): 这是本 PR 唯一修改的文件, 通过在该镜像构建脚本中追加 sgl-eval 安装步骤, 修复 CPU CI 测试失败, 并复用共享版本锁定脚本保持跨平台一致性。

关键符号: 未识别

关键源码片段

docker/xeon.Dockerfile

这是本 PR 唯一修改的文件，通过在该镜像构建脚本中追加 sgl-eval 安装步骤，修复 CPU CI 测试失败，并复用共享版本锁定脚本保持跨平台一致性。

```
# docker/xeon.Dockerfile 构建镜像时的关键安装步骤（节选）
RUN source /opt/.venv/bin/activate && \
    uv pip install . && \
    cd sglang/kernels/aot && \
    cp pyproject_cpu.toml pyproject.toml && \
    # 加载共享的 sgl-eval 版本固定脚本，确保与 CUDA/AMD/NPU 镜像使用同一版本
    source /sgl-workspace/sglang/scripts/ci/utis/sgl_eval_ref.sh && \
    # 通过 uv 安装固定版本的 sgl-eval CLI，供 CPU 机型的 mmlu eval 使用
    uv pip install "$SGL_EVAL_SPEC"
```

评论区精华

本 PR 无实质 Review 讨论。copilot-pull-request-reviewer[bot] 给出概述性 comment，复述了变更内容；维护者 mingfeima 直接批准（APPROVED）。

- 暂无高价值评论线程

风险与影响

- 风险：
 - CI 构建失败风险：Dockerfile 的 RUN 链中新增 source 与 uv pip install，若 sgl_eval_ref.sh 不可用或 \$SGL_EVAL_SPEC 指向不存在 / 不兼容的版本，镜像构建会失败，进而影响依赖该镜像的全部 CPU CI。不过 PR 描述中 CI 已通过，实际风险低。
 - 路径硬编码风险：docker/xeon.Dockerfile 中硬编码了脚本路径 /sgl-workspace/sglang/scripts/ci/utis/sgl_eval_ref.sh，若脚本路径调整而 Dockerfile 未同步，会出现构建失败。
 - 平台兼容性：sgl-eval 可能引入 CPU 平台不支持的依赖，但鉴于 CUDA/AMD/NPU 均使用同一版本规范，此风险较小。
 - 镜像体积与构建时间：额外安装 Python 包会略微增加镜像体积和构建时长，但影响有限。
 - 影响：本 PR 影响范围仅限于 Xeon CPU Docker 镜像构建及相关 CI 测试。修复后，test_mmlu_torch_compile_cpu 等依赖 sgl-eval 的 CPU 测试不再因缺少命令而失败，CPU CI 稳定性提升。对最终用户无直接影响，团队需重新构建 Xeon 镜像后才能生效，影响面小。
 - 风险标记：CI 构建失败风险，跨平台依赖同步

关联脉络

- PR #34477 Route mmlu eval through the external sgl-eval CLI: 该 PR 引入了 sgl-eval 依赖，并更新了 CUDA/AMD/NPU 的 CI 安装脚本，是本 PR 的根因；本 PR 补全了 CPU 镜像中缺失的对应安装步骤。