PR #24924 完整报告

sgl-project/sglang

[AMD] Pin cache-dit==1.3.0 in rocm.Dockerfile + AMD CI install script

合并时间：2026-05-11 22:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24924

执行摘要

- 一句话：修复 AMD CI 因 cache-dit 版本不匹配导致的测试失败
- 推荐动作：值得精读，尤其是 PR body 对 root cause 的详细追查链条以及 review 中对 Dockerfile no-op 的指出。展示了正确的依赖修正路径——从源头的 pyproject 出发，而非在安装脚本中打补丁。

功能与动机

修复 sglang-ci-bot 每日报告中记录的 AMD CI 全红问题，根源是 cache-dit 版本锁定在 1.1.8，而新测试 case 依赖 >=1.2.0 的特性。详情见 PR body 对 root cause 的详细分析。

实现拆解

1. 修复版本锁定源头：修改 python/pyproject_other.toml 中 diffusion_hip extra 的 cache-dit 从 ==1.1.8 改为 ==1.3.0，这是 ROCm Dockerfile 在构建时实际解析的依赖来源。
2. 同步 wheel 配置：修改 3rdparty/amd/wheel/sglang/pyproject.toml 中相同的依赖锁定，确保 amd-sglang wheel 构建时也使用 1.3.0。
3. CI 安装脚本桥接：修改 scripts/ci/amd/amd_ci_install_dependency.sh，将无版本的 pip install cache-dit 改为 pip install --upgrade 'cache-dit==1.3.0'，这样即使 CI 镜像尚未重建也能强制升级到 1.3.0。一旦新镜像部署，该行将变成空操作。
4. Revert 无效 Dockerfile 修改：最初尝试在 Dockerfile 中直接安装 1.3.0，但 review 指出由于下一行 pip install 会重新读取 pyproject_other.toml 从而降级回 1.1.8，该修改为 no-op，已被 revert。

关键文件：

- python/pyproject_other.toml（模块 依赖配置；类别 config；类型 configuration）：这是版本锁定的源头：ROCm Dockerfile 在构建时依赖此文件决定 diffusion_hip 的版本。将 cache-dit 从 1.1.8 改为 1.3.0 根本上修复了镜像构建时的版本偏斜。
- 3rdparty/amd/wheel/sglang/pyproject.toml（模块 依赖配置；类别 config；类型 configuration）：amd-sglang wheel 的依赖配置，需要与 python/pyproject_other.toml 保持同步，确保 wheel 构建也使用正确版本。
- scripts/ci/amd/amd_ci_install_dependency.sh（模块 部署脚本；类别 infra；类型 infrastructure）：CI 安装脚本的桥接修复：在已有镜像上强制升级 cache-dit 到 1.3.0，

确保新镜像部署前 CI 也能通过。

关键符号：未识别

关键源码片段

python/pyproject_other.toml

这是版本锁定的源头：ROCm Dockerfile 在构建时依赖此文件决定 diffusion_hip 的版本。将 cache-dit 从 1.1.8 改为 1.3.0 根本上修复了镜像构建时的版本偏斜。

```
# python/pyproject_other.toml (diffusion_hip extra)
# 此 section 是 AMD ROCm Dockerfile 在构建时实际解析的依赖配置
# 升级 cache-dit 版本以匹配 python/pyproject.toml 中的最新锁定
[diffusion_hip]
dependencies = [
  "sglang[diffusion_common]",
  "peft>=0.18.0,<0.19.0",
  "st_attn==0.0.7",
  "vsa==0.0.4",
  "runai_model_streamer>=0.15.5",
  "cache-dit==1.3.0", # 之前是 ==1.1.8, 升级以支持 --cache-dit-config 所需的 >=1.2.0 特性
]
```

3rdparty/amd/wheel/sglang/pyproject.toml

amd-sglang wheel 的依赖配置，需要与 python/pyproject_other.toml 保持同步，确保 wheel 构建也使用正确版本。

```
# 3rdparty/amd/wheel/sglang/pyproject.toml (diffusion_hip extra)
# 与 python/pyproject_other.toml 保持同步，确保 wheel 构建版本一致
[diffusion_hip]
dependencies = [
  "PyYAML==6.0.1",
  "cloudpickle",
  "diffusers==0.37.0",
  "imageio==2.36.0",
  "imageio-ffmpeg==0.5.1",
  "moviepy>=2.0.0",
  "opencv-python-headless==4.10.0.84",
  "remote-pdb",
  "st_attn==0.0.7",
  "vsa==0.0.4",
  "runai_model_streamer>=0.15.5",
  "cache-dit==1.3.0", # 与 pyproject_other.toml 同步升级
  "addict",
]
```

评论区精华

cursor[bot] 在 review 中指出了 Dockerfile 修改的无效性："This install is overwritten a few lines below... pip will downgrade cache-dit from 1.3.0 back to 1.1.8"。这一发现直接引发了 revert 并转向修正真正的版本源头 pyproject_other.toml。

- Dockerfile 修改为 no-op (correctness): 作者接受 review 意见，revert 了 Dockerfile 修改，改为修正真正的版本源头 pyproject_other.toml。
- 安装脚本修改验证 (correctness): 修改被认可，合入最终版本。

风险与影响

- 风险：该 PR 仅涉及安装脚本和 pyproject 文件的版本号修改，不涉及任何模型代码或运行时逻辑。风险极低。但需注意 diffusion_musa 依赖尚未同步（仍锁定 1.1.8），PR body 已标识为 follow-up。
- 影响：直接影响：修复 AMD CI 中所有使用 --cache-dit-config 的 test case（如 qwen_image_t2i_cache_dit_scm_config_diffusers_1gpu 和 wan2_1_t2v）。间接影响：新构建的 ROCm 镜像将默认包含 cache-dit==1.3.0，减少 CI 安装阶段的升级开销。对用户或生产系统无影响。
- 风险标记：依赖版本偏斜，跨平台配置不同步

关联脉络

- PR #16662 Add cache-dit>=1.2.0 runtime check: 添加了 cache-dit>=1.2.0 的运行检查，但 AMD CI 镜像中的旧版本 1.1.8 未同步更新，导致后续测试失败。
- PR #19213 [diffusion] CI: add cache-dit CI tests: 新增了使用 --cache-dit-config 的测试 case，触发 AMD CI 的版本不匹配，是此 PR 的直接触发因素。
- PR #20361 Bump cache-dit to 1.3.0 in pyproject.toml: 主 pyproject.toml 将 cache-dit 升级到 1.3.0，但 AMD 的 pyproject_other.toml 未同步，导致版本偏斜。