

PR #23161 完整报告

sgl-project/sglang

[AMD] Pin peft<0.19 in pyproject_other.toml to fix ROCm CI ImportError

合并时间: 2026-04-20 14:43

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/23161>

执行摘要

- 一句话: 为 AMD ROCm CI 锁定 peft 版本, 修复因 torchao 版本不兼容导致的导入错误。
- 推荐动作: 该 PR 是一个典型的依赖管理修复, 变更简单直接, 适合快速浏览以了解 CI 问题解决流程。值得关注的设计决策是选择仅在 pyproject_other.toml 中锁定版本而非全面更新所有配置文件, 这反映了优先解决紧急阻塞问题的务实态度。建议团队后续跟进 review 中的建议, 评估是否需要在其他平台配置文件中同步版本锁定以避免未来问题。

功能与动机

PR body 详细说明: peft 0.19.0 (发布于 2026-04-14) 在 `peft.import_utils.is_torchao_available` 中将 torchao 最低版本要求从 `>=0.4.0` 提升至 `>=0.16.0` (引用 [huggingface/peft#3101](#))。AMD ROCm CI 使用的 Docker 镜像 (`rocm/sgl-dev:v0.5.x-rocm700-mi30x-*`) 仅提供 `torchao==0.9.0`, 导致任何导入 `peft.tuners.lora.torchao.dispatch_torchao` 的测试路径都会触发 `ImportError`, 具体表现为两个 LoRA 参考比较测试失败: `test_multi_lora_backend.py` 和 `test_lora_hf_sgl_logprob_diff.py`。由于 AMD CI 在容器内将 `pyproject_other.toml` 重命名为 `pyproject.toml` 作为依赖解析文件, 因此需要在此文件中锁定版本以立即解除 CI 阻塞。

实现拆解

1. 识别问题根源: 分析 AMD CI 失败日志, 定位到 peft 0.19.0 与 torchao 0.9.0 版本不兼容, 导致 `ImportError`。
2. 修改依赖配置文件: 在 `python/pyproject_other.toml` 中, 将 `diffusion_hip` 扩展 (第 102 行) 和 `test` 扩展 (第 164 行) 中的 `peft>=0.18.0` 改为 `peft>=0.18.0,<0.19.0`, 并添加注释说明原因。
3. 验证修复效果: 通过 CI 运行验证 (引用 `run 24563832425` 和 `24599716745`), 确认两个失败的 LoRA 测试现在通过, 且未引入其他回归。
4. 配套说明: 无测试或部署配套改动, 但 PR body 指出长期修复需要升级 ROCm 镜像中的 torchao 版本, 这可以作为后续 PR 跟进。

关键文件:

- `python/pyproject_other.toml` (模块 依赖配置; 类别 `config`; 类型 `configuration`): 这是 AMD CI 实际使用的依赖配置文件, 通过锁定 peft 版本直接解决了 CI 导入错误问题。

关键符号: 未识别

关键源码片段

python/pyproject_other.toml

这是 AMD CI 实际使用的依赖配置文件，通过锁定 peft 版本直接解决了 CI 导入错误问题。

```
# HIP (Heterogeneous-computing Interface for Portability) for AMD
# => base docker rocm/vllm-dev:20250114, not from public vllm whl
diffusion_hip = [
    "sglang[diffusion_common]",
    "peft>=0.18.0,<0.19.0", # Pin to <0.19.0 due to torchao incompatibility
    "st_attn==0.0.7",
    "vsa==0.0.4",
    "runai_model_streamer>=0.15.5",
    "cache-dit==1.1.8",
]

test = [
    "accelerate",
    "expecttest",
    "gguf",
    "jsonlines",
    "matplotlib",
    "pandas",
    "peft>=0.18.0,<0.19.0", # Pin to <0.19.0 due to torchao incompatibility
    "pytest",
    "sentence_transformers",
    "tabulate",
]
```

评论区精华

review 中只有 gemini-code-assist[bot] 提出了建议：

- 建议扩展版本锁定范围：指出 peft 与 torchao 的不兼容性是通用问题，影响所有平台，建议将相同的版本锁定应用到 python/pyproject.toml 和其他变体文件（如 pyproject_cpu.toml、pyproject_npu.toml、pyproject_xpu.toml），以确保环境一致性并防止其他 CI 管道或用户出现类似 ImportError。
- 建议添加注释：在修改行添加解释性注释以提高可维护性，该建议已被采纳（在 head_excerpt 中可见添加的注释）。
- 决策结论：PR 作者仅针对 AMD CI 的紧急问题进行了最小化修复，未采纳扩展锁定范围的建议，这可能意味着团队认为其他平台当前不受影响或可后续处理。HaiShaw 批准了 PR，表明当前变更被视为可接受的临时解决方案。
- 版本锁定范围扩展建议 (design)：PR 作者未采纳此建议，仅针对 AMD CI 进行了最小化修复；HaiShaw 批准了当前变更，表明团队接受此临时方案。
- 添加注释以提高可维护性 (documentation)：建议被采纳，在 head_excerpt 中可见添加的注释“# Pin to <0.19.0 due to torchao incompatibility”。

风险与影响

- 风险：
 - 兼容性风险：锁定 peft 版本可能暂时阻止使用 peft 0.19.0 的新功能或修复，但鉴于 AMD CI 的阻塞问题，此风险是可接受的。
 - 环境不一致风险：如 review 所指出，未在其他平台配置文件中同步锁定版本，可能导致其他 CI 管道或本地环境在未来出现类似 ImportError，但当前 PR 范围仅限于 AMD CI。
 - 回归风险：极低，因为变更仅涉及依赖版本约束，不修改任何业务逻辑代码；CI 验证显示测试通过且无回归。
 - 安全风险：无直接安全影响，但依赖锁定可能延迟安全更新，不过 peft 0.19.0 本身非安全关键版本。
- 影响：
 - 对用户的影响：无直接影响，因为 pyproject_other.toml 主要用于 CI 环境，普通用户通常使用其他配置文件。
 - 对系统的影响：修复了 AMD ROCm CI 中的两个关键 LoRA 测试失败，确保 CI 管道稳定性和测试覆盖率。
 - 对团队的影响：解除 CI 阻塞，允许团队继续合并其他变更，避免开发流程中断；但需注意长期修复（升级 torchao）尚未完成。
 - 影响范围：仅限于 AMD CI 环境，特别是使用 diffusion_hip 和 test 扩展的测试任务。
 - 风险标记：环境不一致风险，依赖锁定延迟更新

关联脉络

- PR #23072 [AMD] Pin peft<0.19.0 in AMD wheel to fix CI failures: 关联 Issue 23072 是同一个问题的另一部分，旨在修复 AMD wheel 中的 peft 版本；本 PR 则针对 CI 安装路径的 pyproject_other.toml，两者共同解决 AMD 环境下的兼容性问题。