# PR #21442 完整报告

sgl-project/sglang

[AMD] Add peft>=0.18.0 to diffusion_hip deps for transformers 5.x compat for AMD diffusion model

合并时间：2026-03-29 11:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21442

## 执行摘要

本 PR 通过为 AMD ROCm 平台的扩散模型添加 peft>=0.18.0 依赖，解决了 transformers 5.x 不兼容导致的启动失败问题。变更限于配置文件，影响扩散工作负载的启动稳定性，确保了特定平台上的模型正常运行。

## 功能与动机

此变更源于 sglang 升级后 ROCm Docker 镜像中扩散模型（如 Wan2.2 T2V）无法启动的问题。具体错误是：ROCm 镜像包含 transformers==5.3.0，但 peft==0.17.0（由 diffusers 拉取）尝试导入 transformers v5 中移除的 HybridCache，引发 ImportError。PR body 中引用：“transformers v5 requires peft>=0.18.0 (huggingface/transformers#41889), which drops the HybridCache import。” 因此，更新依赖是必要的修复。

## 实现拆解

实现仅涉及单个文件修改：python/pyproject_other.toml。具体改动是在 diffusion_hip 依赖项中添加一行 + "peft>=0.18.0"。提交历史显示两个步骤：

- 首次提交添加依赖，消息为“[diffusion] Add peft>=0.18.0 to diffusion_hip deps for transformers 5.x compat”。
- 第二次提交重新定位依赖，消息为“relocate peft to diffusion_hip”，确保依赖项组织有序。

## 评论区精华

review 中唯一评论来自 HaiShaw：> "@yichiche @bingxche please make sure the change works on other platforms." 该评论强调了跨平台兼容性考量，但未引发深入讨论。最终 PR approved，表明风险通过 CI 或其他方式得到缓解，显示团队对多平台测试的重视。

## 风险与影响

风险：

- 跨平台兼容性风险：新依赖可能影响非 ROCm 环境的扩散功能，需 CI 测试验证。
- 依赖版本冲突：peft>=0.18.0 可能与其他包版本不兼容，但此版本是 transformers v5 的强制要求，降低了不匹配风险。

影响：

- 正面修复启动失败，提升 AMD ROCm 用户体验。
- 确保扩散模块在 ROCm 平台上的稳定性，减少 CI 失败和部署中断。
- 对系统影响有限，仅涉及依赖管理，不修改核心代码。

## 关联脉络

- 与 PR 21586 关联：该 PR 处理 transformers 依赖在 CI 中的兼容性问题（通过猴子补丁），强调版本管理和测试策略的共通性。
- 与 PR 21600 关联：作为扩散模块的功能增强，本 PR 确保了依赖兼容性，支撑了扩散生态的持续演进。 这些 PR 共同反映了团队在跨平台和依赖管理中注重兼容性和测试覆盖的趋势。