PR #39421 完整报告

vllm-project/vllm

[ROCm][CI] Resolved nvidia package deps issue

合并时间：2026-04-10 00:06

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39421

执行摘要

该 PR 修复了 ROCm CI 和 Docker 构建因 NVIDIA 包命名变更（引入 cu12/cu13 后缀及无后缀变体）而失败的问题，通过扩展依赖排除列表和调整 Dockerfile 实现，但移除了 PyTorch ROCm 验证步骤，可能引入环境正确性风险。

功能与动机

PR body 明确指出："Fixes ROCm CI/Docker builds broken by nvidia package naming changes (cu12/cu13 suffixed and unsuffixed variants)。" 动机是解决 NVIDIA 包命名策略更新导致的构建依赖冲突，确保 ROCm 相关构建流程稳定运行。

实现拆解

实现涉及三个文件：

1. .pre-commit-config.yaml：在 uv pip compile 步骤中扩展 --no-emit-package 排除列表，新增三类 NVIDIA 包变体：
   - 无后缀（如 nvidia-cublas）
   - -cu12 后缀（如 nvidia-cublas-cu12）
   - -cu13 后缀（如 nvidia-cublas-cu13） 覆盖了 cublas、cudnn、nccl 等关键包，防止 uv 发出这些包导致依赖冲突。
2. requirements/rocm-test.txt：同步更新排除列表，确保与 .pre-commit-config.yaml 一致，例如将 --no-emit-package, nvidia-cudnn-cu13 改为 --no-emit-package, nvidia-cudnn 并添加所有变体。
3. docker/Dockerfile.rocm：将原有的 PyTorch ROCm 验证步骤（检查 torch.version.hip）替换为持久化构建的 wheel：dockerfile COPY --from=export_vllm /*.whl /opt/vllm-wheels/ 目的是让后续脚本 python_only_compile_rocm.sh 能在移除编译器后重新安装 wheel。

评论区精华

review 中仅有的讨论聚焦于 Dockerfile 修改：

- gemini-code-assist[bot]指出：

"While persisting the built wheel is a good addition, the verification step that ensures the PyTorch build is for ROCm has been removed. This check is a valuable

safeguard to prevent building a Docker image with an incorrect PyTorch version."

- tjtanaa回复：

"Let's keep the original check" 但 PR 最终被 tjtanaa 批准合并，未在提交中重新添加验证步骤，此疑虑处于未解决状态。

风险与影响

- 主要风险：移除 PyTorch ROCm 验证步骤可能导致构建的 Docker 镜像错误使用 CUDA 版本 PyTorch，影响 ROCm 环境正确性，可能引发运行时问题。
- 次要风险：大量手动排除项增加了维护负担，未来 NVIDIA 包命名再变更时需同步更新。
- 影响范围：仅限于 CI 和 Docker 构建流程，对最终用户无直接影响，但确保 ROCm 测试和部署的稳定性。

关联脉络

从近期历史 PR 看，该 PR 与以下 PR 同属基础设施修复范畴：

- #39390：修复 CI 夜间索引生成权限问题，同属 CI 脚本调整。
- #38950：在 Dockerfile 中添加 fastsafetensors 包，优化构建过程。
- #39164：为 XPU 跳过测试设置，解决 CI 死锁，同属平台特定 CI 调整。 整体趋势显示团队持续优化多平台（ROCm、XPU）的构建和测试基础设施，以应对硬件和依赖生态变化。