

PR #52303 完整报告

vllm-project/vllm

[Doc] [ROCm] Update installation documentation

合并时间: 2026-08-17 16:56

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52303>

执行摘要

- 一句话: ROCm 安装文档新增 Python-only 构建说明
- 推荐动作: 该 PR 值得快速浏览, 作为理解 vLLM 预编译 wheel 开发工作流的参考。重点看环境变量和注意事项小节, 可帮助团队在 ROCm 开发机上快速搭建免编译环境; 同时可作为后续文档变更保持 CUDA/ROCm 双文档同步的样例。

功能与动机

PR body 明确说明这是对 <https://github.com/vllm-project/vllm/pull/49365> 的文档跟进: "Right now it supports the install Python only feature." 即 #49365 已实现 ROCm 上的 Python-only 安装能力, 本 PR 负责把该能力的使用方式、环境变量和限制写入官方安装文档, 避免开发者在只修改 Python 代码时仍触发完整编译, 也防止误用导致导入错误。

实现拆解

变更仅涉及 1 个文档文件: [docs/getting_started/installation/gpu.rocm.inc.md](#)。实现过程如下:

1. 定位插入点: 在 build-wheel-from-source 区段内、pre-built-wheels 块之后, 新增 #### Set up using Python-only build (without compilation) {#python-only-build} 小节, 并在其后保留 #### Full build (with compilation) {#full-build} 小节, 保证从免编译到完整编译的叙述顺序。
2. 说明核心命令与执行流程: 给出 `VLLM_USE_PRECOMPILED=1 python3 setup.py develop`, 并分 5 步解释运行时行为——定位当前分支、找到 main 分支上的 base commit、探测 ROCm 版本、下载匹配的预编译 wheel、复用其编译库与 vllm-rs 二进制。
3. 补齐可控参数: 列出 `VLLM_PRECOMPILED_WHEEL_LOCATION`、`VLLM_PRECOMPILED_WHEEL_COMMIT`、`VLLM_PRECOMPILED_WHEEL_VARIANT` 三个环境变量, 并说明显式指定变体时必须与系统探测结果一致。
4. 增加使用限制与运维提示: 两个 note 强调修改 C++、HIP 或 kernel 代码时不可用此方式, rebase 后建议卸载重装; 一个 tip 说明可用 `./build_rust.sh` (含 `--debug`) 单独重建 Rust 前端。
5. 补充故障排查与版本一致性提醒: 提示 wheel 找不到时按 <https://wheels.vllm.ai/rocm/<commit>> 检查可用变体, 并建议源码 commit 与 wheel 保持一致以避免未知错误。

6. 一致性确认：作者在 review 中说明该小节与 docs/getting_started/installation/gpu.cuda.inc.md 的既有写法保持一致；无测试、配置或部署配套改动。

关键文件：

- docs/getting_started/installation/gpu.rocm.inc.md（模块 安装文档；类别 docs；类型 documentation）：唯一变更文件，在 ROCm 安装指南中新增 Python-only 构建章节，说明使用预编译 wheel 免编译安装的开发流程，并补充环境变量、Rust 重建方法与使用限制。

关键符号：未识别

评论区精华

评审中仅有一处格式质疑：DarkLight1337 在 [gpu.rocm.inc.md](#) 第 195 行评论 "I think this should be indented?"，针对 Python-only 小节内的列表缩进。作者 tjtananaa 回复 "Yes, this is following the [docs/getting_started/installation/gpu.cuda.inc.md](#) directly."，说明该写法是照搬 CUDA 文档的既有格式。DarkLight1337 随后批准（APPROVED），未再提出其他问题。

- Python-only 小节列表缩进 (style)：作者确认是刻意与 CUDA 文档保持一致，评审者接受解释并批准 PR。

风险与影响

- 风险：风险集中在文档与实现的同步漂移上：文档中的三个环境变量（VLLM_PRECOMPILED_WHEEL_LOCATION、VLLM_PRECOMPILED_WHEEL_COMMIT、VLLM_PRECOMPILED_WHEEL_VARIANT）和 wheel 探测逻辑若在后续代码中改名或调整，文档不会自动校验，可能误导用户；wheels.vllm.ai 的变体命名（如 rocm721）依赖发布侧节奏，ROCM 版本更新时若未同步文档会出现死链；另外该小节与 CUDA 文档共用一套说明，若 CUDA 侧调整写法，本文件需要人工跟进。整体技术风险极低。
- 影响：面向 ROCm 平台的开发者，尤其是需要迭代 vLLM Python 代码但不想每次完整编译的用户；文档降低了 Python-only 工作流的上手成本，并提前警示了改底层代码（C++、HIP、kernel）时的失败模式。对 vLLM 团队而言，该文档与 #49365 的能力形成闭环，且与 CUDA 文档保持一致，维护成本低。影响范围限定在文档，不影响运行时行为。
- 风险标记：仅文档变更，环境变量与实现易漂移，依赖远程 wheel 可用性

关联脉络

- PR #49365 Python-only install support (PR body 引用，标题未提供)：PR body 明确指定本文档是 #49365 的 follow up，对应 Python-only 安装功能的实现。
- PR #49514 [ROCM][CI] Use the same-build wheel in Python-only CI：同属 ROCm 预编译 wheel 基础设施，前者调整 CI 使用同构建 wheel，本 PR 为用户侧安装文档，两者共同完善 Python-only 工作流。