

PR #45169 完整报告

vllm-project/vllm

[Bugfix] [DSV4] [ROCm] Pin apache-tvm-ffi version to `0.1.10`

合并时间: 2026-06-11 01:41

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45169>

执行摘要

- 一句话: 固定 apache-tvm-ffi 版本至 0.1.10 解决 DSV4 启动崩溃
- 推荐动作: 此 PR 为简单的版本锁定修复, 值得合并。建议在后续 ROCm 依赖升级时注意检查 apache-tvm-ffi 与 tilelang 的版本兼容性。

功能与动机

在 ROCm 环境下启动 DeepSeek V4 时出现 `TypeAttr __ffi_repr__ is already registered` 错误, 这是由于 `apache-tvm-ffi` 被意外升级到 0.1.12 导致的兼容性问题。PR body 明确指出 “Pinning to the version 0.1.10 as what has been shown in the `requirements/test/rocm.txt` resolves this issue。”

实现拆解

1. 修复 `requirements/rocm.txt`: 在 `tilelang==0.1.10` 之后添加 `apache-tvm-ffi==0.1.10` 行, 并追加注释说明该版本需与 `tilelang` 版本匹配。
2. 修复 `requirements/test/rocm.txt`: 在 `apache-tvm-ffi==0.1.10` 那一行 (原本已有) 后面添加注释 `# -c requirements/rocm.txt`, 确保来自主文件的约束被引用。
3. 验证与测试: 在 PR body 中给出了 DeepSeek V4 启动命令、GSM8K 准确率 (约 95%) 和推测解码接受率 (平均 2.68 tokens, 接受率 83.8%) 等结果, 确认修复有效。

关键文件:

- `requirements/rocm.txt` (模块 依赖; 类别 docs; 类型 documentation): 主 requirements 文件, 添加并锁定 `apache-tvm-ffi` 版本, 是修复的核心
- `requirements/test/rocm.txt` (模块 依赖; 类别 docs; 类型 documentation): 测试 requirements 文件, 增加注释以确保主文件约束生效

关键符号: 未识别

评论区精华

该 PR 无 review 评论, 仅由维护者 AndreasKaratzas 批准, 讨论较少。社区可能默认此修复合理, 因为版本固定是常见的兼容性解决方案。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 版本冲突：固定 `apache-tvm-ffi==0.1.10` 可能与未来需要更高版本的新功能冲突，但当前已与 `tilelang==0.1.10` 匹配，风险较低。
2. 影响范围小：仅变更两个 `requirements` 文件，不影响核心代码逻辑，回归风险极低。
3. 无测试覆盖：没有新增测试来验证版本兼容性，但通过 PR body 中的实际运行结果验证了修复。

- 影响：

- 用户：ROCm 平台使用 DeepSeek V4 的用户不再遇到启动崩溃，可直接使用。
- 系统：`apache-tvm-ffi` 版本被固定，未来依赖升级需主动决策。
- 团队：仅两个文件变更，维护成本低。影响范围仅限于 ROCm + DeepSeek V4 场景。
- 风险标记：依赖版本锁定，未增加测试，影响范围小

关联脉络

- PR #44821 fix: prefix DeepSeek V4 MTP projections: 同为 DeepSeek V4 相关修复，涉及 ROCm 平台下的 DeepSeek V4 功能