PR #46741 完整报告

vllm-project/vllm

[ROCm][Bugfix] Fix HIP fork re-init in multimodal offline examples

合并时间：2026-06-26 13:37

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46741

执行摘要

- 一句话：修复 ROCm 上 fork 安全问题的环境变量设置
- 推荐动作：该 PR 是一个典型的环境兼容性修复，技术含量不高，但解决了 CI 中的实际问题。建议在类似场景中考虑在 Dockerfile 或容器层面统一设置 PYTORCH_NVML_BASED_CUDA_CHECK=1，以避免在多个命令中重复添加。

功能与动机

AMD ROCm 的 CI 步骤（examples 和 transformers nightly models）在执行多模态离线示例时，由于 transformers 在 vllm 之前被导入，导致 torch.cuda.is_available() 触发了 cuInit，从而在父进程中创建了 HIP 上下文，使得后面 fork 出来的子进程无法重新初始化 CUDA，报错 RuntimeError: Cannot re-initialize CUDA in forked subprocess。

实现拆解

1. 定位问题根源：在 .buildkite/test-amd.yaml 中，vision_language_offline.py 和 vision_language_multi_image_offline.py 这两个示例脚本的导入顺序导致 HIP 上下文在父进程中被初始化。
2. 确定修复方案：利用 vllm/env_override.py 中已经定义的 PYTORCH_NVML_BASED_CUDA_CHECK=1 环境变量，它可以让 torch.cuda.is_available() 绕过 cuInit，转而通过 NVML 检测 GPU 可用性，从而避免创建 HIP 上下文。
3. 修改 CI 配置文件：在两个受影响的地方（mi300•examples 和 Transformers Nightly Models）的命令前加上 PYTORCH_NVML_BASED_CUDA_CHECK=1 前缀，并添加注释说明原因。

关键文件：

- .buildkite/test-amd.yaml（模块 CI 配置；类别 config；类型 configuration）：唯一的变更文件，修改了 CI 命令，添加了 PYTORCH_NVML_BASED_CUDA_CHECK=1 环境变量，修复了 ROCm 上因导入顺序导致的 fork 上下文问题。

关键符号：未识别

评论区精华

mawong-amd 提出了是否有更好方案的问题，例如在 Dockerfile 中设置该环境变量或放到测试级别，以避免代码膨胀。作者 peizhang56 随后更新了 PR，将之前的方案（可能在多个地方分散设置）改为在当前修改中更集中地处理。最终 PR 被批准。

- 暂无高价值评论线程

风险与影响

- 风险：该修复仅影响 CI 配置文件，不会对运行时行为产生影响。PYTORCH_NVML_BASED_CUDA_CHECK=1 是 vLLM 已支持的机制，且已经在 vllm/env_override.py 中使用，因此风险极低。需要注意的是，如果未来有其他示例也遵循同样的导入顺序，可能还需要应用相同的修复。
- 影响：直接修复了 ROCm CI 中两个测试阶段的失败问题，包括 Examples 步骤和 Transformers Nightly Models 步骤中的多模态离线示例。影响范围仅限于 AMD ROCm CI 环境，不影响 NVIDIA 或其他平台。修复方式是非侵入性的，不需要修改 Python 源码。
- 风险标记：暂无

关联脉络

- PR #46749 [CI][Bugfix] Spawn engine in mm cache sleep test to fix ROCm HIP error: 同样修复 ROCm 上的 fork 相关错误，使用 spawn 方法替代默认 fork。