

PR #49739 完整报告

vllm-project/vllm

[ROCm][CI] Wait for ROCm VRAM to settle between compiled and eager LL...

合并时间: 2026-07-25 06:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49739>

执行摘要

- 一句话: 修复 ROCm 编译测试中 GPU 内存未稳定问题
- 推荐动作: 值得合并。这是一个典型的 ROCm 特定显存稳定问题修复, 类似模式在 ROCm 测试中常见。可推广到其他类似场景。

功能与动机

ROCm CI 构建 11128 在运行 `test_dynamic_shapes_compilation[False-True-0-backed-Qwen/Qwen2-7B-Instruct]` 时失败, 原因是编译 Qwen2-7B 运行后, 同一进程中启动 eager LLM 时在 V2 model runner 中发生 GPU 故障。`gc` 和 `empty_cache()` 不足以完全释放 ROCm 状态。

实现拆解

1. 导入辅助函数: 在 `tests/compile/test_dynamic_shapes_compilation.py` 的导入部分添加 `from tests.utils import wait_for_rocm_memory_to_settle`。
2. 在模型切换间插入等待: 在 `del model`、`gc.collect()`、`torch.accelerator.empty_cache()`、`torch.accelerator.synchronize()` 之后, `eager_model = LLM(...)` 之前调用 `wait_for_rocm_memory_to_settle()`, 确保 ROCm 显存状态稳定后再初始化第二个模型。

关键文件:

- `tests/compile/test_dynamic_shapes_compilation.py` (模块 测试; 类别 `test`; 类型 `test-coverage`): 唯一修改的文件, 添加了 `wait_for_rocm_memory_to_settle` 导入和调用, 修复 ROCm CI 测试失败。

关键符号: 未识别

关键源码片段

`tests/compile/test_dynamic_shapes_compilation.py`

唯一修改的文件, 添加了 `wait_for_rocm_memory_to_settle` 导入和调用, 修复 ROCm CI 测试失败。

```
# tests/compile/test_dynamic_shapes_compilation.py ( 片段 )
```

```
import gc
import tempfile
```

```
from contextlib import contextmanager

import pytest
import torch

from tests.models.utils import check_logprobs_close
from tests.utils import wait_for_rocm_memory_to_settle # 新增导入
from vllm import LLM, SamplingParams
...

# 编译模型运行后
del model
gc.collect()
torch.accelerator.empty_cache()
torch.accelerator.synchronize()
wait_for_rocm_memory_to_settle() # 等待 ROCm 显存稳定, 避免 eager 初始化时 GPU 故障

eager_model = LLM(model=model_name, enforce_eager=True, max_model_len=1024)
```

评论区精华

该 PR 无 review 讨论（只有 bot 自动评论和批准）。核心问题为 ROCm 上编译模型卸载后显存未完全释放，导致 eager 模型初始化故障。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试文件，添加 2 行代码，且函数已在 tests/utils 中定义，调用语义明确。不会影响产品代码或非 ROCm 平台。
- 影响：该修复仅影响 ROCm CI 中的 test_dynamic_shapes_compilation 测试，消除一个已知的 flaky 失败。对系统功能、性能、兼容性无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR