PR #46749 完整报告

vllm-project/vllm

[CI][Bugfix] Spawn engine in mm cache sleep test to fix ROCm HIP error

合并时间：2026-06-26 13:38

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46749

执行摘要

- 一句话：修复 ROCm 上 mm cache 测试因 fork 继承 HIP 上下文失败
- 推荐动作：值得合并。这是一个最小修复，精准针对 ROCm HIP 上下文 fork 问题，且已通过测试验证。可作为在 ROCm 测试中处理多进程测试启动方式的参考。

功能与动机

该测试是为回归验证 issue #42995 而添加的，但在 ROCm CI 中失败，报错 torch.AcceleratorError: HIP error: invalid argument。原因是在同一进程中运行多个测试后，父进程持有的 HIP 上下文在 fork 后对子进程无效。PR 描述明确说明"root cause: the test instantiates an LLM, which launches EngineCore via the default VLLM_WORKER_MULTIPROC_METHOD=fork start method... The forked child inherits an invalid HIP context"。

实现拆解

变更仅涉及一个测试文件，分两步：

1. 导入装饰器：在 tests/multimodal/test_cache.py 中添加 from ..utils import create_new_process_for_each_test 导入。
2. 添加装饰器：在测试函数 test_sleep_wake_preserves_mm_cache_consistency 上添加 @create_new_process_for_each_test() 装饰器。该装饰器在 ROCm/XPU 上默认使用 spawn 启动方式，在 CUDA 上保持 fork，从而隔离 HIP 上下文，避免冲突。

关键文件：

- tests/multimodal/test_cache.py（模块 缓存；类别 test；类型 test-coverage）：唯一变更文件：修复在 ROCm CI 上因 fork 继承无效 HIP 上下文导致的测试失败。添加了装饰器和导入。

关键符号：未识别

关键源码片段

tests/multimodal/test_cache.py

唯一变更文件：修复在 ROCm CI 上因 fork 继承无效 HIP 上下文导致的测试失败。添加了装饰器和导入。

```
# tests/multimodal/test_cache.py
# 新增导入：使用 create_new_process_for_each_test 将测试隔离到独立进程中
from ..utils import create_new_process_for_each_test

# ... 其他导入和辅助函数 ...

@create_new_process_for_each_test() # 新增装饰器：在 ROCm/XPU 上 spawn，CUDA 上 fork
@pytest.mark.skipif(
    not torch.cuda.is_available(),
    reason="sleep mode regression requires a CUDA GPU",
)
def test_sleep_wake_preserves_mm_cache_consistency():
    """Regression for vllm-project/vllm#42995."""
    from vllm import LLM, SamplingParams
    from vllm.assets.image import ImageAsset

    image = ImageAsset("stop_sign").pil_image
    prompt = {
        "prompt": _SLEEP_VISION_PROMPT,
        "multi_modal_data": {"image": image},
    }
    sampling_params = SamplingParams(temperature=0, max_tokens=8)

    llm = LLM(
        model="Qwen/Qwen2-VL-2B-Instruct",
        enable_sleep_mode=True,
        enforce_eager=True,
        gpu_memory_utilization=0.5,
        max_model_len=2048,
    )

    llm.generate([prompt], sampling_params)
    llm.sleep(level=1)
    llm.wake_up()
    output2 = llm.generate([prompt], sampling_params)
    assert output2[0].outputs[0].text
```

评论区精华

仅有两条审核评论：claude[bot] 自动评论表示 fork 仓库禁止自动审核；AndreasKaratzas 批准并致谢。无实质性讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅 3 行，且使用已有工具函数 create_new_process_for_each_test，该工具已在 tests/basic_correctness/test_mem.py 中用于类似睡眠测试。NVIDIA CUDA 平台行为不变，因为装饰器在 CUDA 上仍使用 fork。测试结果验证了 186 通过、0

失败。

- 影响：直接影响 ROCm 平台的 test_sleep_wake_preserves_mm_cache_consistency 测试，使其在 CI 多测试环境下正确运行。对用户无影响，因为仅修改测试代码。团队受益于更稳定的 ROCm CI。
- 风险标记：测试变更，平台特定（ROCm）

关联脉络

- PR #43001 Add regression test for #42995: 引入了本次修复的测试函数 test_sleep_wake_preserves_mm_cache_consistency，是本次 PR 的直接前驱。
- PR #42995 [Bug]: V1 sleep/wake leaves P0 multimodal sender cache desynced from P1 → AssertionError on next image reuse: 原始 bug 报告，本次 PR 是回归测试的一部分修复。
- PR #44543 [Bugfix] Couple audio+video in mm processor cache for use_audio_in_video (fixes #44538): PR body 提及此 PR 无关，但同属于多模态缓存修复系列。