

# PR #47242 完整报告

vllm-project/vllm

[CI/Build] Fix LoRA testing

合并时间: 2026-07-01 15:36

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47242>

## 执行摘要

- 一句话: 修复 LoRA 多模态测试中环境变量未恢复问题
- 推荐动作: 该 PR 属于常规测试修复, 值得关注的点是使用 monkeypatch 管理环境变量的最佳实践, 可在其他需要临时环境变量设置的测试中推广。

## 功能与动机

修复 CI 中 LoRA 测试可能因环境变量全局修改导致后续测试行为异常的问题。reviewer DarkLight1337 在评论中指出: "Use monkeypatch so it doesn't affect other tests after this test completes"。

## 实现拆解

1. 修改 `run_test` 函数签名: 增加 `monkeypatch: pytest.MonkeyPatch` 参数; 在函数体开头调用 `monkeypatch.setenv("VLLM_WORKER_MULTIPROC_METHOD", "spawn")`, 替代原先的 `os.environ` 直接赋值方式。
2. 更新所有调用处: 在 `test_active_default_mm_lora`、`test_inactive_default_mm_lora`、`test_default_mm_lora_succeeds_with_redundant_lora_request`、`test_default_mm_lora_fails_with_overridden_lora_request` 四个测试函数的参数列表和调用处均增加 `monkeypatch` 参数。
3. 完善回归测试: 在 `test_default_mm_lora_does_not_expand_string_reqs` 中同样使用 `monkeypatch.setenv` 并添加注释, 说明强制 `spawn` 模式的原因 (防止 `forked` 子进程 CUDA 重初始化崩溃)。

关键文件:

- `tests/lora/test_default_mm_loras.py` (模块 LoRA 测试; 类别 test; 类型 test-coverage; 符号 `run_test`, `test_default_mm_lora_does_not_expand_string_reqs`): 唯一修改的文件, 修复环境变量泄漏问题并完善回归测试

关键符号: `run_test`, `test_default_mm_lora_does_not_expand_string_reqs`

## 关键源码片段

`tests/lora/test_default_mm_loras.py`

唯一修改的文件, 修复环境变量泄漏问题并完善回归测试

```
# tests/lora/test_default_mm_loras.py

def run_test(
    vllm_runner, audio_assets, monkeypatch, lora_request, expected_suffix, **kwargs
):
    # 使用 monkeypatch 设置环境变量，测试结束后自动恢复，避免影响后续测试
    monkeypatch.setenv("VLLM_WORKER_MULTIPROC_METHOD", "spawn")
    inputs = [[AUDIO_PROMPT], [audio_assets[0].audio_and_sample_rate[0]]]
    vllm_runner_kwargs = {"**VLLM_RUNNER_BASE_KWARGS", **kwargs}
    with vllm_runner(**vllm_runner_kwargs) as vllm_model:
        vllm_outputs_with_default_lora = [
            vllm_model.generate_greedy(
                prompts, max_tokens=128, audios=audios, lora_request=lora_request,
            )
            for prompts, audios in inputs
        ]
        assert vllm_outputs_with_default_lora[-1][-1][-1].endswith(expected_suffix)

@create_new_process_for_each_test()
def test_default_mm_lora_does_not_expand_string_reqs(vllm_runner, monkeypatch):
    # 强制 spawn 模式以规避 forked 子进程 CUDA 重初始化崩溃
    monkeypatch.setenv("VLLM_WORKER_MULTIPROC_METHOD", "spawn")
    class MockEngineException(Exception):
        pass
    # ... 后续测试逻辑
```

## 评论区精华

DarkLight1337 建议使用 `monkeypatch` 代替直接设置 `os.environ`，以确保环境变量在测试结束后自动恢复，避免全局污染。作者 jeejeelee 回复“Makes sense”并采纳该建议。

- 使用 `monkeypatch` 替代 `os.environ` 直接赋值 (testing): 采用 `monkeypatch.setenv` 方式，修改后已满足要求。

## 风险与影响

- 风险：该 PR 仅修改测试文件，且改动量很小 (+17/-2)，不涉及生产代码。使用 `pytest` 的 `monkeypatch` fixture 是标准做法，不会引入回归风险。
- 影响：仅影响 LoRA 多模态相关测试的隔离性，确保环境变量不会跨测试泄漏。对其他模块无影响。
- 风险标记：测试隔离性修复

## 关联脉络

- PR #47193 [ROCm][CI] Enable LoRA TP Distributed Test Group In AMD CI: 同为 LoRA 测试相关，涉及 CI 中的 LoRA 测试组启用