PR #50161 完整报告

vllm-project/vllm

[CI][ROCm] Stabilize Qwen2-VL LoRA test

合并时间：2026-07-29 12:27

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50161

执行摘要

- 一句话：稳定 Qwen2-VL LoRA 测试的确定性
- 推荐动作：值得精读，特别是需要理解 Triton LoRA kernel 非确定性根源和 VLLM_BATCH_INVARIANT 作用的工程师。此 PR 展示了如何用最小的测试配置变更解决 CI 不稳定问题。

功能与动机

Buildkite CI（如 #80664）上 Qwen2-VL LoRA 测试在 ROCm 上出现随机失败，导致 CI 不稳定。PR body 明确指出目标是 "Stabilize and re-enable LoRA %N gating mirror"。

实现拆解

1. 添加确定性辅助函数 _enable_deterministic_lora_shrink：在 tests/lora/test_qwenvl.py 中新定义此函数，通过 monkeypatch.setenv("VLLM_BATCH_INVARIANT", "1") 强制 Triton LoRA shrink kernel 使用 SPLIT_K=1 以避免 split-K 原子累加的非确定性；同时设置 VLLM_WORKER_MULTIPROC_METHOD=spawn 确保环境变量在子进程中生效。
2. 修改 test_qwen3vl_vision_lora 函数：将其签名加入 monkeypatch 参数，并在测试体开头调用 _enable_deterministic_lora_shrink(monkeypatch)。
3. 修改 test_qwen2vl_multiple_lora_types 函数：同样加入 monkeypatch 参数并调用 _enable_deterministic_lora_shrink。
4. 删除 ROCm 镜像的 soft_fail: true 配置：在 .buildkite/test_areas/lora.yaml 的 AMD 镜像配置中移除此选项，使测试失败不再被容错，从而真正拦截非确定性回归。

关键文件：

- tests/lora/test_qwenvl.py（模块 测试；类别 test；类型 test-coverage；符号 _enable_deterministic_lora_shrink, test_qwen3vl_vision_lora, test_qwen2vl_multiple_lora_types）：核心变更文件，新增确定性辅助函数，修改两个测试函数以启用确定性 LoRA shrink。
- .buildkite/test_areas/lora.yaml（模块 CI 配置；类别 config；类型 configuration）：移除 ROCm 镜像的 soft_fail: true，使测试失败不再被容错，从而真正启用稳定性检查。

关键符号：_enable_deterministic_lora_shrink, test_qwen3vl_vision_lora, test_qwen2vl_multiple_lora_types

关键源码片段

tests/lora/test_qwenvl.py

核心变更文件，新增确定性辅助函数，修改两个测试函数以启用确定性 LoRA shrink。

```
def _enable_deterministic_lora_shrink(monkeypatch: pytest.MonkeyPatch) -> None:
    # These tests assert exact greedy outputs. Force the Triton LoRA shrink
    # kernel to use SPLIT_K=1 so it stores the complete reduction directly
    # instead of accumulating split-K partial results with atomic_add. This
    # targets reduction determinism, not full batch invariance.
    monkeypatch.setenv("VLLM_BATCH_INVARIANT", "1")
    # The kernel configuration reads VLLM_BATCH_INVARIANT at import time.
    # Spawn the engine process so it observes this setting even if the LoRA
    # Triton utilities were already imported during test collection.
    monkeypatch.setenv("VLLM_WORKER_MULTIPROC_METHOD", "spawn")


def test_qwen3vl_vision_lora(
    qwen3vl_vision_lora_files,
    monkeypatch: pytest.MonkeyPatch,
):
    _enable_deterministic_lora_shrink(monkeypatch)
    config = TestConfig(
        model_path=QWEN3VL_MODEL_PATH,
        lora_path=qwen3vl_vision_lora_files,
        mm_processor_cache_gb=0,
        enable_tower_connector_lora=True,
    )
    with Qwen2VLTester(config) as tester:
        for lora_id in [1, 2]:
            tester.run_test(
                TEST_IMAGES,
                expected_outputs=EXPECTED_OUTPUTS,
                lora_id=lora_id,
            )
```

评论区精华

核心讨论点是 VLLM_BATCH_INVARIANT 在 ROCm 上的支持情况。tjtanaa 提问："Batch invariant is still not fully supported on ROCm. or am I mistaken?" AndreasKaratzas 确认并解释：他们利用该标志触发 Triton kernel 中更确定的路径（vllm/lora/ops/triton_ops/kernel_utils.py 第 336 行），而不是完全依赖完整批次不变性。作者随后更新了注释以澄清意图。

- VLLM_BATCH_INVARIANT 在 ROCm 上的支持 (correctness): 接受了作者的说明，并更新了注释以更精确描述行为。

风险与影响

- 风险：低风险。变更仅限测试代码和 CI 配置。VLLM_BATCH_INVARIANT 属于已有环境变量，对非 LoRA 场景无影响。移除 soft_fail 可能会使原本被容忍的失败直接阻断 CI，但这正是意图所在——捕获真实回归。
- 影响：影响范围仅限于 Qwen-VL 系列 LoRA 测试（Qwen2-VL, Qwen2.5-VL, Qwen3-VL）在 ROCm CI 上的执行。由于强制确定性规约，测试输出将更稳定，避免因非确定性结果导致的随机失败。其他平台（如 NVIDIA）不受影响。
- 风险标记：缺少测试覆盖（原问题为 CI 不稳定），仅 ROCm 平台受影响

关联脉络

- PR #48245 [BugFix] Fix num_output_placeholders preemption underflow: 同为 V1 scheduler 稳定性修复，但影响范围不同。