PR #42266 完整报告

vllm-project/vllm

[CI/Build] Reduce LoRA model tests.

合并时间：2026-05-11 14:49

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42266

执行摘要

- 一句话：精简 CI LoRA 测试，跳过冗余项
- 推荐动作：建议关注被跳过多 GPU 测试的死代码问题，考虑仅在 CI 特定标记而非平台级别跳过；对于 AMD CI，建议验证实际运行测试是否有效；此 PR 的设计决策值得学习，但覆盖风险需接受。

功能与动机

当前 CI 触发了过多的 LoRA 模型测试。本 PR 跳过一部分测试，保留最大覆盖以避免回归。（引用 PR body: 'Currently, CI triggers too many model tests for LoRA. This PR skips some of them — the logic is to retain the tests with the maximum coverage to avoid regression.'）

实现拆解

1. 添加平台跳过装饰器：在 tests/lora/test_default_mm_loras.py、test_chatglm3_tp.py、test_qwen35_densemodel_lora.py、test_qwenvl.py、test_olmoe_tp.py、test_llama_tp.py、test_minicpmv_tp.py 中，对选定的测试函数添加 @pytest.mark.skipif(current_platform.is_cuda_alike(), reason="Skipping to avoid redundant model tests") 装饰器，使这些测试在 CUDA/ROCm 平台上跳过。这样减少重复执行，但保留在非 CUDA 平台上（如 CPU）运行的机会。
2. 删除冗余测试函数：在 tests/lora/test_whisper.py 中，删除 test_whisper_with_and_without_lora 函数，因为该用例与 test_whisper_multi_lora 功能重叠。
3. 重命名测试文件：将 tests/lora/test_llm_with_multi_loras.py 重命名为 tests/lora/test_qwen3_with_multi_loras.py，使文件名更准确地反映其针对 Qwen3 模型。
4. 调整 CI 配置：更新 .buildkite/test-amd.yaml 和 .buildkite/test_areas/lora.yaml，对应重命名后的文件路径，并在 AMD CI 步骤中保持对这些测试的调用；但注意，由于这些测试自身添加了 skipif，AMD CI 步骤实际覆盖率降低。

关键文件：

- tests/lora/test_whisper.py（模块 Whisper 测试；类别 test；类型 test-coverage；符号 test_whisper_with_and_without_lora）：删除了冗余测试函数 test_whisper_with_and_without_lora，直接减少测试数量

- tests/lora/test_default_mm_loras.py（模块 多模态 LoRA；类别 test；类型 test-coverage；符号 test_inactive_default_mm_lora, test_default_mm_lora_succeeds_with_redundant_lora_request, test_default_mm_lora_fails_with_overridden_lora_request）：添加 skipif 装饰器跳过三个多模态 LoRA 测试，降低重复
- tests/lora/test_chatglm3_tp.py（模块 ChatGLM3 TP 测试；类别 test；类型 test-coverage；符号 test_chatglm3_lora, test_chatglm3_lora_tp4）：添加 skipif 装饰器，但导致多 GPU 测试变为死代码
- tests/lora/test_qwen35_densemodel_lora.py（模块 Qwen3.5 密集模型；类别 test；类型 test-coverage；符号 test_qwen35_text_lora）：添加 skipif 跳过 Qwen3.5 密集模型 LoRA 测试
- tests/lora/test_qwenvl.py（模块 Qwen VL 测试；类别 test；类型 test-coverage；符号 test_qwen25vl_lora, test_qwen25vl_vision_lora）：添加 skipif 跳过 Qwen2.5 VL 及 Vision LoRA 测试
- tests/lora/test_olmoe_tp.py（模块 OlMoE TP 测试；类别 test；类型 test-coverage；符号 test_olmoe_lora, test_olmoe_lora_tp2）：添加 skipif 跳过 Olmoe LoRA 测试
- tests/lora/test_llama_tp.py（模块 LLaMA TP 测试；类别 test；类型 test-coverage）：添加 skipif 跳过 LLaMA TP 测试
- tests/lora/test_minicpmv_tp.py（模块 MiniCPM-V TP 测试；类别 test；类型 test-coverage）：添加 skipif 跳过 MiniCPM-V TP 测试
- tests/lora/test_qwen3_with_multi_loras.py（模块 Qwen3 多 LoRA 测试；类别 test；类型 rename-or-move）：重命名文件以反映测试内容
- .buildkite/test-amd.yaml（模块 AMD CI 配置；类别 config；类型 configuration）：调整 AMD CI 步骤中的测试文件路径
- .buildkite/test_areas/lora.yaml（模块 LoRA CI 区域；类别 config；类型 configuration）：调整 CI 区域配置中的测试文件路径

关键符号：test_whisper_with_and_without_lora, test_inactive_default_mm_lora, test_default_mm_lora_succeeds_with_redundant_lora_request, test_default_mm_lora_fails_with_overridden_lora_request, test_chatglm3_lora, test_chatglm3_lora_tp4, test_qwen35_text_lora, test_qwen25vl_lora, test_qwen25vl_vision_lora, test_olmoe_lora, test_olmoe_lora_tp2

关键源码片段

tests/lora/test_whisper.py

删除了冗余测试函数 test_whisper_with_and_without_lora，直接减少测试数量

```
# 原测试函数被删除，以避免 CI 冗余
# 该测试验证 Whisper 模型在有 / 无 LoRA 时输出不同
@create_new_process_for_each_test()
def test_whisper_with_and_without_lora(whisper_lora_files):
    """Test that Whisper produces different outputs with and without LoRA."""
    llm = create_whisper_llm(enable_lora=True)
    outputs_with_lora = run_whisper_inference(
```

```
        llm, lora_path=whisper_lora_files, lora_id=1
    )
    outputs_without_lora = run_whisper_inference(llm, lora_path=None)
    assert len(outputs_with_lora[0]) > 0
    assert len(outputs_without_lora[0]) > 0
    print(f"Output with LoRA: {outputs_with_lora[0]}")
    print(f"Output without LoRA: {outputs_without_lora[0]}")
```

tests/lora/test_chatglm3_tp.py

添加 skipif 装饰器，但导致多 GPU 测试变为死代码

```
import pytest
from vllm.platforms import current_platform

# 此测试同时标记 skipif 和 multi_gpu_test (4 GPUs),
# 导致在任何环境下都无法执行（死代码风险）
@pytest.mark.skipif(
    current_platform.is_cuda_alike(),
    reason="Skipping to avoid redundant model tests"
)
@multi_gpu_test(num_gpus=4)
def test_chatglm3_lora_tp4(chatglm3_lora_files):
    pass
```

评论区精华

Review 讨论主要集中在三个方面：一是 skipif 与 multi_gpu_test 组合导致某些多 GPU 测试变成死代码（@gemini-code-assist 指出的 critical 问题）；二是跳过多模态 LoRA 的 modality 过滤测试（test_inactive_default_mm_lora）被认为降低了在主流 GPU 上的重要覆盖；三是 AMD CI 步骤中显式调用的测试因自身 skipif 而失效，造成 CI 步骤变成空运行。作者回应表示保留其他测试（如 Qwen3.5 VL 测试）可弥补覆盖，但合并者 DarkLight1337 最终批准了 PR。

- skipif 与 multi_gpu_test 组合导致死代码 (testing): 作者未直接回复，但 PR 已合并，可能需要后续清理。
- AMD CI 步骤测试自身跳过 (testing): 未解决，但 PR 已合并。
- 跳过 Qwen2.5 VL 测试降低覆盖 (testing): 作者认为可接受，合并者批准。

风险与影响

- 风险：风险包括： 1) 死代码风险：test_chatglm3_lora_tp4 等测试被 skipif 和 multi_gpu_test 同时标记，在任何环境下都无法执行，可能隐藏多 GPU LoRA 回归。 2) 覆盖下降：在 CUDA/ROCm 平台上跳过多项测试，包括多模态 LoRA 关键逻辑，可能遗漏回归。 3) AMD CI 覆盖矛盾：.buildkite/test-amd.yaml 中明确调用这些测试，但测试自身跳过，导致该 CI 步骤无有效验证。 4) 本地开发影响：开发者运行完整测试集时跳过较多测试，可能减少发现问题的机会。

- 影响：对用户无直接影响；对开发者：在 GPU 环境下本地运行 pytest 时会跳过较多测试，需注意 CI 中的实际验证覆盖；对 CI 系统：总测试时间减少，但 AMD CI 步骤可能无效；对团队：后续需监控是否有未被覆盖的回归。
- 风险标记：覆盖下降，死代码风险，AMD CI 矛盾

关联脉络

- PR #42196 [CI] Trigger LoRA test when changing MoE code.: 同为 LoRA 测试 CI 优化，调整触发条件