PR #52237 完整报告

vllm-project/vllm

[UT] fix device of test_outputs.py

合并时间：2026-08-14 11:29

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/52237

执行摘要

- 一句话：测试设备类型改为平台自适应，支持全平台运行
- 推荐动作：本 PR 规模小、逻辑简单，不值得精读，但提供了一个良好的实践：在测试中避免硬编码设备，使用平台抽象。可关注 vllm.platforms.current_platform 的用法，作为后续编写可移植测试的参考。

功能与动机

PR 描述明确指出，使用 current_platform.device_type 替代 device="cuda" 是为了让单元测试在所有平台上都能成功运行。标签为 intel-gpu，表明该改动与 Intel GPU 平台的测试兼容性相关。

实现拆解

1. 导入平台模块：在 tests/v1/test_outputs.py 中新增 from vllm.platforms import current_platform 导入。
2. 定义设备常量：在模块顶部新增 DEVICE_TYPE = current_platform.device_type，集中定义设备类型。
3. 替换硬编码设备：在 test_sampling_mask_tensors_from_logits 和 test_sampling_mask_matches_processed_top_k_top_p_support 中，将所有 device="cuda" 替换为 device=DEVICE_TYPE，涉及 torch.tensor 和 apply_top_k_top_p 的参数。
4. 测试验证：该改动仅影响测试代码，无需额外配置或部署。

关键文件：

- tests/v1/test_outputs.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，将测试中的 CUDA 硬编码替换为平台无关的设备类型，提升测试可移植性。

关键符号：test_sampling_mask_tensors_from_logits, test_sampling_mask_matches_processed_top_k_top_p_support

关键源码片段

tests/v1/test_outputs.py

唯一变更文件，将测试中的 CUDA 硬编码替换为平台无关的设备类型，提升测试可移植性。

```
# tests/v1/test_outputs.py
from vllm.platforms import current_platform # 导入平台抽象，获取当前设备类型

# 定义全局设备类型常量，替代硬编码的 'cuda'
DEVICE_TYPE = current_platform.device_type


def test_sampling_mask_tensors_from_logits():
    tensors = SamplingMaskTensors.from_logits(
        logits=torch.tensor(
            [
                [1.0, float("-inf"), 2.0],
                [3.0, 4.0, float("-inf")],
                [float("-inf"), 5.0, 6.0],
            ],
            device=DEVICE_TYPE, # 使用平台设备，支持 CUDA、XPU 等
        ),
        num_sampled_tokens=torch.tensor([1, 0, 1], device=DEVICE_TYPE),
    )

    result = tensors.tolists(np.array([1, 0, 1]))

    assert result.token_ids.tolist() == [0, 2, 1, 2]
    assert result.offsets.tolist() == [0, 2, 4]
    assert result.cu_num_generated_tokens == [0, 1, 1, 2]


def test_sampling_mask_matches_processed_top_k_top_p_support():
    processed_logits = apply_top_k_top_p(
        # 同样使用 DEVICE_TYPE，确保在非 CUDA 平台也能正确创建张量
        logits=torch.tensor(
            [[6.0, 5.0, 4.0, 4.0, 4.0, 2.0, 1.0, 0.0]], device=DEVICE_TYPE
        ),
        k=torch.tensor([3], device=DEVICE_TYPE),
        p=torch.tensor([0.9], device=DEVICE_TYPE),
    )
    # 后续逻辑不变
```

评论区精华

审查过程中，claude[bot] 因 PR 来自 fork 而自动跳过审查，提醒维护者可通过 @claude review 触发一次性审查。维护者 jikunshang 直接批准了 PR，无进一步讨论。

- 自动化代码审查禁用 (other): 未实际触发审查，维护者直接批准。

风险与影响

- 风险：风险极低。改动仅涉及测试文件中的设备参数，不影响生产代码。唯一潜在风险是 current_platform.device_type 的返回值可能与预期设备不匹配，但该 API 已广泛使用，且

CI 会覆盖多平台验证。

- 影响：影响范围限于 tests/v1/test_outputs.py 中的两个测试用例，使得这些测试能够在非 CUDA 平台（如 Intel GPU）上执行。对用户和系统无直接影响，对团队而言提高了测试的便携性。
- 风险标记：测试可移植性，低风险

关联脉络

- PR #53035 [CI][XPU] Skip test_fused_shared_expert.py on XPU: 同为提升非 CUDA 平台测试兼容性的改动，涉及设备相关处理。