

PR #49733 完整报告

vllm-project/vllm

[ROCm][CI] Fix XPASS(strict) on mixed audio embeds test

合并时间: 2026-07-25 05:09

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49733>

执行摘要

- 一句话: 条件化 xfail 标记避免 XPASS 错误
- 推荐动作: 建议快速合入, 以修复 ROCm CI 阻塞。

功能与动机

ROCm CI 仍使用 PyTorch 2.11, 此时 `text-then-audio_embeds` 测试正常通过, 但代码中的无条件 `strict=True` xfail 将通过视作 `XPASS(strict)` 并导致 CI 失败。PR body 明确指出: 'Apply the strict xfail only when `is_torch_equal_or_newer("2.12.0")`', 以兼容当前 ROCm 环境并为未来升级做好准备。

实现拆解

该 PR 仅涉及对测试文件中 xfail 标记的条件化修改, 步骤如下:

1. 导入工具函数: 在文件头部新增 `from vllm.utils.torch_utils import is_torch_equal_or_newer`, 用于版本比较。
2. 修改 xfail 标记: 将 `text-then-audio_embeds` 参数上的 `pytest.mark.xfail(strict=True)` 改为添加 `condition=is_torch_equal_or_newer("2.12.0")`, 使标记仅在 PyTorch ≥ 2.12 时生效。

无其他文件变更, 测试行为保持不变: 在 PyTorch ≥ 2.12 上测试仍预期失败; 在更低版本上则正常执行并通过。

关键文件:

- `tests/entrypoints/multimodal/openai/chat_completion/test_chat_completion_with_mixed_audio_embeds.py` (模块 测试; 类别 test; 类型 test-coverage): 唯一变更文件, 通过条件化 xfail 标记修复 ROCm CI 上的 `XPASS(strict)` 失败。

关键符号: 未识别

关键源码片段

`tests/entrypoints/multimodal/openai/chat_completion/test_chat_completion_with_mixed_audio_embeds.py`

唯一变更文件, 通过条件化 xfail 标记修复 ROCm CI 上的 `XPASS(strict)` 失败。

```

# 新增导入
from vllm.utils.torch_utils import is_torch_equal_or_newer

# xfail 标记改为条件形式
@pytest.mark.parametrize(
    "audio_first",
    [
        pytest.param(True, id="audio_embeds-then-text"),
        pytest.param(
            False,
            id="text-then-audio_embeds",
            marks=pytest.mark.xfail(
                condition=is_torch_equal_or_newer("2.12.0"), # 仅在 PyTorch >= 2.12
                时标记为预期失败
                reason="torch 2.12 regression: prompt_embeds output diverges "
                "from raw-text when text precedes audio; "
                "https://github.com/pytorch/pytorch/issues/184431",
                strict=True,
            ),
        ),
    ],
)

```

评论区精华

审核者 AndreasKaratzas 最初建议直接跳过 (skip) 低版本测试，但最终采用了条件 xfail 方案，既保留了预期失败标记，又兼容当前 CI 环境。PR 作者随后采纳了带版本条件的方式，并获批准。

- 条件化 xfail 方案 vs 直接跳过 (design): 采用条件 xfail 方案，既在低版本正常执行测试，也为 PyTorch 2.12 升级保留预期失败标记。

风险与影响

- 风险：风险极低：仅更改测试标记逻辑，不影响模型行为或服务路径；若 `is_torch_equal_or_newer` 行为异常（如返回错误值），可能导致测试标记在错误版本生效，但已通过 `--collect-only` 验证。
- 影响：直接影响 ROCm CI 的稳定性，消除因环境差异导致的误报。对用户无影响，对 PyTorch 2.12+ 环境的 CI 行为无变化。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR