PR #50163 完整报告

vllm-project/vllm

[ROCm][CI] Stabilize ROCm audio streaming test

合并时间：2026-07-29 07:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50163

执行摘要

- 一句话：ROCm 音频流式测试稳定性修复
- 推荐动作：值得合并，变更简单明确，无副作用。

功能与动机

CI 中 ROCm 音频流式测试存在非确定性失败，如 Buildkite 构建 #11367 所示。禁用 prefix caching 可以消除因缓存导致的不一致性。

实现拆解

1. 在 tests/entrypoints/multimodal/openai/chat_completion/test_audio.py 中导入 vllm.platforms.current_platform。
2. 新增 _ROCM_ARGS 列表：若当前平台为 ROCm，则包含 --no-enable-prefix-caching，否则为空。
3. 在 server fixture 的启动参数末尾展开 _ROCM_ARGS，使 ROCm 下传递该参数，其他平台无影响。

关键文件：

- tests/entrypoints/multimodal/openai/chat_completion/test_audio.py（模块 音频测试；类别 test；类型 test-coverage）：唯一变更文件，添加 ROCm 下禁用 prefix caching 的逻辑以稳定测试。

关键符号：未识别

关键源码片段

tests/entrypoints/multimodal/openai/chat_completion/test_audio.py

唯一变更文件，添加 ROCm 下禁用 prefix caching 的逻辑以稳定测试。

```
# tests/entrypoints/multimodal/openai/chat_completion/test_audio.py
from vllm.platforms import current_platform

# 仅在 ROCm 上禁用 prefix caching，以减少流式与非流式比较中的非确定性。
_ROCM_ARGS = ["--no-enable-prefix-caching"] if current_platform.is_rocm() else []

@pytest.fixture(scope="module")
```

```
def server():
    args = [
        "--dtype", "float32",
        "--max-model-len", "2048",
        "--max-num-seqs", "5",
        "--enforce-eager",
        "--trust-remote-code",
        "--limit-mm-per-prompt", json.dumps({"audio": MAXIMUM_AUDIOS}),
        *_ROCM_ARGS, # 展开条件参数
    ]
    with RemoteOpenAIServer(MODEL_NAME, args) as remote_server:
        yield remote_server
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅当平台为 ROCm 时禁用 prefix caching，不影响其他平台；测试本身已独立运行，不会影响生产代码。
- 影响：仅影响 ROCm CI 中音频流式测试的稳定性，对其他平台无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR