

PR #52565 完整报告

vllm-project/vllm

[ROCm][CI] Avoid forcing FlashAttention in the ColPali pooling test

合并时间: 2026-08-17 12:08

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52565>

执行摘要

- 一句话: 修复 ColPali 测试在 ROCm 上强制 FlashAttention 导致的 CI 失败
- 推荐动作: 值得快速阅读, 作为一个典型的跨平台测试修复案例: 当引入平台特定的测试配置时, 应使用 `current_platform` 进行条件判定, 避免阻塞其他硬件平台。后续若有类似测试, 建议遵循此模式。

功能与动机

PR #52050 添加了 CUDA 特定的强制 `FLASH_ATTN` 配置到跨平台 ColPali 测试, 导致 MI300 和 MI355 上的 CI 失败, 因为 ROCm 有意不报告 FlashAttention 版本。本 PR 保留 CUDA 回归覆盖, 同时允许 ROCm 选择其支持的注意力后端。

实现拆解

1. 在 `tests/models/multimodal/pooling/test_colpali.py` 中导入 `current_platform` (from `vllm.platforms import current_platform`), 用于运行时平台判断。
2. 在 `_run_multimodal_text_query_image_docs_test` 内新增变量 `attention_backend`, 仅当 `current_platform.is_cuda()` 为真时赋值为 `"FLASH_ATTN"`, 否则为 `None`。
3. 将传给 `vllm_runner` 的参数从硬编码 `"FLASH_ATTN"` 改为该变量, 使非 CUDA 平台 (如 ROCm) 走默认后端选择逻辑。

该改动仅涉及测试文件, 无产品代码变更, 也未新增测试用例。

关键文件:

- `tests/models/multimodal/pooling/test_colpali.py` (模块 池化测试; 类别 test; 类型 test-coverage; 符号 `_run_multimodal_text_query_image_docs_test`): 唯一变更文件, 通过平台条件选择注意力后端, 修复 ROCm CI 失败, 同时保留 CUDA 回归覆盖。

关键符号: `_run_multimodal_text_query_image_docs_test`

关键源码片段

`tests/models/multimodal/pooling/test_colpali.py`

唯一变更文件, 通过平台条件选择注意力后端, 修复 ROCm CI 失败, 同时保留 CUDA 回归覆盖。

```

# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
"""Tests for ColPali late interaction model for multi-modal retrieval.

ColPali is a multi-vector retrieval model based on PaliGemma backbone
(SigLIP + Gemma) with ColBERT-style late interaction scoring (MaxSim).
It produces per-token embeddings for both text and image inputs.
"""

from io import BytesIO

import pybase64 as base64
import pytest
import torch
from PIL import Image

from vllm.entrypoints.chat_utils import (
    ChatCompletionContentPartImageParam,
    ChatCompletionContentPartTextParam,
)
from vllm.entrypoints.pooling.scoring.typing import ScoreMultiModalParam
from vllm.platforms import current_platform

from ....conftest import VllmRunner

# ... 省略其他测试辅助函数 ...

def _run_multimodal_text_query_image_docs_test(
    vllm_runner: type[VllmRunner],
    model: str,
    *,
    dtype: str,
) -> None:
    """Score a text query against image documents via the multimodal path."""
    red_image = _make_base64_image(64, 64, color=(255, 0, 0))
    blue_image = _make_base64_image(64, 64, color=(0, 0, 255))

    query = "Describe the red object"
    image_docs = [
        _make_image_mm_param(red_image),
        _make_image_mm_param(blue_image),
    ]
    # 仅 CUDA 平台强制使用 FlashAttention 后端, ROCm 等平台使用默认后端
    attention_backend = "FLASH_ATTN" if current_platform.is_cuda() else None

    with vllm_runner(
        model,
        runner="pooling",
        dtype=dtype,

```

```

max_model_len=4096,
enforce_eager=True,
gpu_memory_utilization=GPU_MEMORY_UTILIZATION,
attention_backend=attention_backend,
kernel_config={"enable_flashinfer_autotune": False},
) as vllm_model:
    # 确保运行在 V2 model runner 上，验证新路径
    assert vllm_model.llm.llm_engine.vllm_config.use_v2_model_runner
    scores = vllm_model.llm.score(query, image_docs)

    assert len(scores) == 2
    for s in scores:
        assert isinstance(s.outputs.score, float)

```

评论区精华

该 PR 无实质性评审讨论。[nooooo](#) 直接批准 (APPROVED)，[claude\[bot\]](#) 仅自动提醒手动 review 配置。唯一值得注意的是作者在 Issue 评论中说明了 CI 测试已通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低，仅涉及测试代码。主要潜在风险：若存在既非 CUDA 也非 ROCm 的平台（如 XPU、CPU）运行此测试，attention_backend 将为 None，依赖默认后端；若默认后端不支持该模型，可能引入新的失败，但该测试原先在这些平台上也可能因强制 FlashAttention 失败，因此整体修复方向正确。此外，current_platform.is_cuda() 的语义在虚拟化环境下可能需留意，但 vLLM 内部已处理。
- 影响：影响范围限于 tests/models/multimodal/pooling/test_colpali.py 单个测试文件。修复了 ROCm 平台（MI300、MI355）上该测试的 CI 失败，同时保持 CUDA 上的回归覆盖。对产品运行、API 均无影响。
- 风险标记：跨平台回归，仅测试变更

关联脉络

- PR #52050 [Bugfix] Temporarily disable FA4 head-dim 256: 该 PR 为 ColPali 测试引入了强制 FLASH_ATTN 配置，是本 PR 修复的直接原因。
- PR #52570 [CI/Build] Reduce more duplicate runner startup in tests: 该 PR 也修改了 tests/models/multimodal/pooling/test_colpali.py，属于同一测试文件的演进。