PR #47004 完整报告

vllm-project/vllm

[ROCm][CI][Multimodal] Use ROCm-aware FA availability check for Unlimited-OCR

合并时间：2026-06-30 14:03

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47004

执行摘要

- 一句话：修复 ROCm 上 Unlimited-OCR 的 FA 导入失败
- 推荐动作：值得合并的小修复，无需深入精读。可提醒团队统一使用 fa_utils 进行 FA 版本检查，避免类似问题。

功能与动机

在 ROCm 环境下，vllm.vllm_flash_attn 是 CUDA-only 模块，导入会失败，导致 baidu/Unlimited-OCR 模型在配置验证阶段就崩溃，无法正确 fallback 到 FlexAttention。

实现拆解

1. 修改导入源：在 vllm/model_executor/models/config.py 的 UnlimitedOCRForCausalLMConfig.verify_and_update_config 方法中，将 from vllm.vllm_flash_attn import is_fa_version_supported 替换为 from vllm.v1.attention.backends.fa_utils import is_fa_version_supported。
2. 保持逻辑不变：其余代码（包括 FA4 可用性判断和 fallback 逻辑）不做任何改动，仅修复导入路径。

关键文件：

- vllm/model_executor/models/config.py（模块 模型配置；类别 source；类型 data-contract）：修复了导入路径，使得 ROCm 环境下 Unlimited-OCR 模型能够正确判断 FlashAttention 4 可用性并 fallback。

关键符号：UnlimitedOCRForCausalLMConfig.verify_and_update_config

关键源码片段

vllm/model_executor/models/config.py

修复了导入路径，使得 ROCm 环境下 Unlimited-OCR 模型能够正确判断 FlashAttention 4 可用性并 fallback。

```
# 在 UnlimitedOCRForCausalLMConfig.verify_and_update_config 中
from vllm.v1.attention.backends.fa_utils import is_fa_version_supported # 使用统一入口，支持 ROCm
from vllm.v1.attention.backends.registry import AttentionBackendEnum
```

```
attn_config = vllm_config.attention_config
fa4_available = is_fa_version_supported(4) # 行为与之前完全一致

if attn_config.backend is None:
    attn_config.backend = (
        AttentionBackendEnum.FLASH_ATTN
        if fa4_available
        else AttentionBackendEnum.FLEX_ATTENTION # 自动 fallback
    )
```

评论区精华

无讨论，PR 直接被审核者 DarkLight1337 批准合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改一行导入语句，逻辑完全不变。需验证 fa_utils.is_fa_version_supported 在 CUDA 和 ROCm 上行为一致。
- 影响：影响范围小：仅针对 Unlimited-OCR 模型在 ROCm 环境下的配置验证。CUDA 环境不受影响，导入路径仍然有效。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR