PR #36274 完整报告

vllm-project/vllm

[Bugfix][ROCm] Strip block_size before attention backend validation

合并时间：2026-03-12 04:37

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/36274

执行摘要

本 PR 修复了 ROCm 平台上因 attention backend 验证逻辑拒绝不规则 block_size 值导致的模型启动失败 bug，通过剥离 block_size 与 CUDA 平台保持一致，使 Qwen3 Next 等模型能正确运行。变更仅修改一个文件，两行代码，风险低且已被快速合并。

功能与动机

动机源于 ROCm attention backend refactor (#35246) 引入的 validate_configuration 调用，该调用会拒绝 block_size=544 等不规则值（因为不在 BlockSize 类型中）。CUDA 平台已通过验证前剥离 block_size 避免此问题，本 PR 对 ROCm 应用相同修复，确保跨平台一致性，解决如 Qwen3 Next 无法启动的兼容性问题。

实现拆解

仅修改 vllm/platforms/rocm.py 文件的 get_attn_backend_cls 函数，在验证前添加一行代码：

```
attn_selector_config = attn_selector_config._replace(block_size=None)
```

此变更直接复制 CUDA 平台的逻辑，剥离 block_size 使其在验证中被忽略，无其他结构性改动，属于简单的逻辑对齐。

评论区精华

review 讨论简洁且积极：

- gemini-code-assist[bot]： "This is a clean and targeted solution that resolves the issue, allowing models with non-standard block sizes to run correctly on ROCm."
- houseroad： "Looks good." 无争议点，变更被快速接受，视为正确修复。

风险与影响

风险：变更极小，仅两行代码，且模仿已有 CUDA 逻辑，风险可控；但剥离 block_size 可能掩盖某些配置问题，不过由于测试显示 Qwen3 Next 正确启动，风险低。无安全或性能回归风险。 影响：仅影响 ROCm 平台的 attention backend 选择，修复后提升模型兼容性（如 Qwen3 Next），对系统性能和安全无负面影响。影响范围有限，不涉及其他模块。

关联脉络

本 PR 直接关联 #35246（ROCm attention backend refactor），是对其引入 bug 的修复。从近期历史 PR 看，ROCm 平台持续有 bugfix 和测试改进（如 #37228、#38161），表明团队在加强 ROCm 支持，本 PR 是这一趋势中的一致性维护，反映了跨平台代码对齐的重要性。