

PR #44674 完整报告

vllm-project/vllm

[ROCm][Kernel] Enable permute_cols for ROCm

合并时间: 2026-06-07 17:50

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44674>

执行摘要

- 一句话: ROCm 启用 permute_cols op
- 推荐动作: 该 PR 值得快速合并, 因为它去除了一个无必要的平台门槛, 提升了 ROCm 平台的功能完整性。设计上简单地移除保护宏并正确调整 CMake 配置, 可作为平台兼容性治理的优选范例。

功能与动机

此前 `permute_cols` 算子仅对 CUDA 编译, ROCm 环境被 `#ifndef USE_ROCM` 屏蔽。该 PR 在 ROCm 上验证通过 (`pytest tests/kernels/core/test_permute_cols.py` 3 个测试全部通过), 因此解除条件编译, 使 ROCm 用户也能使用该功能。

实现拆解

1. 移除头文件保护宏: 在 `csrc/libtorch_stable/ops.h` 中, 将 `permute_cols` 的声明从 `#ifndef USE_ROCM` 块内移出, 放到块外, 使其对所有平台可见。
2. 移除注册与实现保护宏: 在 `csrc/libtorch_stable/torch_bindings.cpp` 中, `permute_cols` 的 `ops.def` 和 `ops.impl` 调用也从 `#ifndef USE_ROCM` 块内移出, 使其在 ROCm 下编译和注册。
3. 修正 CMake 编译配置: 在 `CMakeLists.txt` 中, 将 `permute_cols.cu` 从 CUDA 专用文件列表移动到共享的 `VLLM_STABLE_EXT_SRC` 列表 (第 630-632 行), 使得 HIP 编译器也能参与编译。同时删除了原先 CUDA 列表中重复的 `permute_cols.cu` 以及其他两个已存在于共享列表中的文件条目, 消除了重复编译。

关键文件:

- `csrc/libtorch_stable/torch_bindings.cpp` (模块 算子绑定; 类别 source; 类型 core-logic) : 核心绑定文件, 移除了 `permute_cols` 注册和实现的 `USE_ROCM` 保护宏。
- `csrc/libtorch_stable/ops.h` (模块 算子声明; 类别 source; 类型 core-logic) : 声明文件, 移除了 `permute_cols` 声明的平台保护。
- `CMakeLists.txt` (模块 构建配置; 类别 infra; 类型 configuration) : 构建配置, 从 CUDA 专用源列表移除 `permute_cols.cu` 并添加到共享列表。

关键符号: `permute_cols`

关键源码片段

csrc/libtorch_stable/torch_bindings.cpp

核心绑定文件，移除了 permute_cols 注册和实现的 USE_ROCM 保护宏。

```
// 在 STABLE_TORCH_LIBRARY_FRAGMENT 中：
// 将 permute_cols 定义移出 #ifndef USE_ROCM 块，使 ROCm 也能注册该 op
ops.def("permute_cols(Tensor A, Tensor perm) -> Tensor");

// 在 STABLE_TORCH_LIBRARY_IMPL 中：
// 同样移出条件编译，使实现对所有平台可见
ops.impl("permute_cols", TORCH_BOX(&permute_cols));
```

csrc/libtorch_stable/ops.h

声明文件，移除了 permute_cols 声明的平台保护。

```
// 原先是 #ifndef USE_ROCM 块内的内容，现移至块外，统一对所有平台可见
torch::stable::Tensor permute_cols(torch::stable::Tensor const& A,
                                   torch::stable::Tensor const& perm);

// 其他 CUTLASS 相关声明仍保留在 #ifndef USE_ROCM 内，因为它们只支持 CUDA
#ifndef USE_ROCM
bool cutlass_scaled_mm_supports_fp8(int64_t cuda_device_capability);
...
#endif
```

CMakeLists.txt

构建配置，从 CUDA 专用源列表移除 permute_cols.cu 并添加到共享列表。

```
# ===== 共享源文件列表（同时用于 CUDA 和 HIP 编译） =====
set(VLLM_STABLE_EXT_SRC
    ...
    "csrc/libtorch_stable/permute_cols.cu" # 新增：之前只在 CUDA 专用列表中
    "csrc/libtorch_stable/quantization/gptq/q_gemm.cu"
    ...
)

# ===== CUDA 专用列表（现移除 permute_cols.cu） =====
if(VLLM_GPU_LANG STREQUAL "CUDA" OR VLLM_GPU_LANG STREQUAL "HIP")
    set(VLLM_STABLE_EXT_CUDA_SRC
        ...
        # permute_cols.cu 已在共享列表中，此处删除以避免重复编译
        # 同样删除了重复的 per_token_group_quant.cu 项
    )
endif()
```

评论区精华

AndreasKaratzas 在 CMakeLists.txt 的 review 评论中询问被从 CUDA 路径移除的源文件是否在其他地方被包含。提交者 charlifur 回复它们已经在共享的 `VLLM_STABLE_EXT_SRC` 列表中（第 630-631 行）。AndreasKaratzas 确认这一改动是安全的，因为共享列表同时被

CUDA 和 ROCm 的构建目标使用，移除 CUDA 专用列表中的重复条目不会导致构建失败。

- CMakeLists.txt 中删去的源文件是否在其他地方被包含 (design): charlifu 指出它们已经在共享的 VLLM_STABLE_EXT_SRC 列表中。AndreasKaratzas 进一步确认这些文件通过共享列表同时参与 CUDA 和 HIP 编译，移除重复项安全且合理。

风险与影响

- 风险：风险很低。仅涉及条件编译开关和 CMake 文件调整，不改变算子逻辑；已通过对应测试套件（3 个测试全部通过）。但需注意该算子目前仅在 libtorch stable 路径下，且 HIP 编译器的支持可能不及 CUDA 成熟，若遇 ROCm 平台特定 bug 可能需要在后续修复。
- 影响：对用户：ROCm（AMD GPU）用户现在可以使用 permute_cols 算子，无需绕过。对系统：减少条件编译分支，代码更清晰。对团队：消除 CUDA 和 HIP 之间的功能差异，减少维护负担。影响程度：小范围，仅涉及 3 个文件的轻微修改。
- 风险标记：小范围变更，已有测试覆盖

关联脉络

- 暂无明显关联 PR