

PR #50068 完整报告

vllm-project/vllm

[Model] Enable Qwen3.8 for AMD Rocm

合并时间: 2026-08-07 21:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50068>

执行摘要

- 一句话: Qwen3.5 文本纯模型补 M-RoPE 支持, ROCm 可加载 FP8 checkpoint
- 推荐动作: 建议精读 `vllm/model_executor/models/qwen3_5.py` 中 `SupportsMRoPE` 声明与 `get_mrope_input_positions` 实现的配套关系: 这是 vLLM 能力协议 (声明 + 实现必须成对) 的一个典型样例, Codex 对重复基类和 stub 陷阱的指正也值得留意。变更小、结构清晰, 适合作为理解 M-RoPE 初始化链路和模型能力协议机制的入门阅读材料。

功能与动机

PR body 明确说明目标: register the text-only Qwen3_5ForCausalLM and Qwen3_5MoeForCausalLM architectures - advertise hybrid and M-RoPE support on the causal implementation ... so text-only Qwen3.5-compatible checkpoints such as Qwen3.8 Max FP8 can initialize through the causal LM path。背景是 Qwen3.5 系列此前以多模态路径为主, 而社区文本纯 checkpoint 常保留 VL 训练栈遗留的 `model.language_model.` 前缀 (代码中 `hf_to_vllm_mapper` 专门剥离此前缀), 要让他们走 causal LM 路径, runner 必须在请求初始化时正确生成 M-RoPE 位置。需要说明的不确定性: PR 标题写 Enable Qwen3.8 for AMD Rocm, 而代码与 body 均指向 Qwen3.5 架构; 按 body 理解 Qwen3.8 Max FP8 是兼容 Qwen3.5 架构的文本 checkpoint, 且 body 提到的 architecture 注册与 Mamba cache metadata 暴露在 base 版本中已存在, 最终净变更只有 M-RoPE 相关 10 行。

实现拆解

1. 能力声明: 在 `vllm/model_executor/models/qwen3_5.py` 的 `.interfaces` 导入中新增 `SupportsMRoPE`, 并将其加入 `Qwen3_5ForCausalLMBase` 的基类列表。该接口是模型能力协议, GPU runner 会在请求加入时依据它调用 `get_mrope_input_positions`, 因此声明必须与实现成对出现。
2. 位置生成实现: 新增 `get_mrope_input_positions`, 对文本纯输入生成 `0..seq_len-1` 的位置序列, 通过 `unsqueeze(0).expand(3, -1)` 广播成 3 行以满足 M-RoPE 的“时间、行、列”三维位置约定, 并返回 `delta = 0` 表示无多模态偏移。该方法位于基类, 同时影响 `Dense` 与 `MoE` 两个子类。
3. 正确性修复: Codex 自动 review 提出两个 P1 问题——基类列表重复声明 `IsHybrid` 导致构造时 `TypeError: duplicate base class IsHybrid`; 以及只声明 `SupportsMRoPE` 而未实现方法时, 继承到的协议 stub 返回 `Ellipsis`, 会在 `_init_mrope_positions` 解包时抛

TypeError。作者在提交 eee31f3b7 中移除重复基类并补齐实现，这也是最终 diff 收敛为 10 行净增的原因。

4. 测试与 CI：本次未新增测试文件；PR body 声明已通过 py_compile、git diff --check，并在双节点 ROCm 集群（TP = 8、PP = 2）上验证了 OpenAI 兼容 health、chat、streaming、completion 请求。pre-commit 曾多次失败，作者修复后由 ZJY0516 触发 Buildkite CI #82808 并通过。

关键文件：

- vllm/model_executor/models/qwen3_5.py（模块 模型定义；类别 source；类型 data-contract；符号 SupportsMRoPE, get_mrope_input_positions, Qwen3_5ForCausalLMBase）：唯一的变更文件，集中实现 Qwen3.5 文本纯 causal 模型的 M-RoPE 能力声明与位置生成，是 ROCm 上加载 Qwen3.8 Max FP8 类 checkpoint 的关键入口；同时包含对 Codex 指出的重复基类与协议 stub 两个正确性问题的修复。

关键符号：get_mrope_input_positions

关键源码片段

vllm/model_executor/models/qwen3_5.py

唯一的变更文件，集中实现 Qwen3.5 文本纯 causal 模型的 M-RoPE 能力声明与位置生成，是 ROCm 上加载 Qwen3.8 Max FP8 类 checkpoint 的关键入口；同时包含对 Codex 指出的重复基类与协议 stub 两个正确性问题的修复。

```
# interfaces 导入中新增 M-RoPE 能力声明
from .interfaces import (
    HasInnerState,
    IsHybrid,
    MixtureOfExperts,
    MultiModalEmbeddings,
    SupportsEagle3,
    SupportsLoRA,
    SupportsMRoPE, # 新增：声明模型具备 M-RoPE 能力
    SupportsPP,
    _require_is_multimodal,
)

class Qwen3_5ForCausalLMBase(
    nn.Module,
    HasInnerState,
    IsHybrid,
    SupportsEagle3,
    SupportsLoRA,
    SupportsMRoPE, # 新增：让 runner 走 M-RoPE 位置初始化路径
    SupportsPP,
):
    # 注意：早期提交曾在此重复声明 IsHybrid,
    # 触发 `TypeError: duplicate base class`, 已在 `eee31f3b7` 中移除。
```

```
def get_mrope_input_positions(
    self,
    input_tokens: list[int],
    mm_features: list[object],
) -> tuple[torch.Tensor, int]:
    # 文本纯场景没有多模态特征，位置直接取 0..seq_len-1；
    # M-RoPE 约定需要 3 行位置（时间、行、列），
    # 这里用 expand 把 1 行广播成 3 行，第二项 delta = 0 表示无图像偏移。
    positions = torch.arange(len(input_tokens), dtype=torch.long)
    return positions.unsqueeze(0).expand(3, -1), 0
```

评论区精华

Codex 自动 review 的两条 P1 是本次最有价值的交锋：

- 重复基类问题：引用评论“Importing qwen3_5 now fails while constructing Qwen3_5ForCausalLMBase with TypeError: duplicate base class IsHybrid”，会让整个模块的架构注册加载失败。
- 协议 stub 问题：引用评论“Implement M-RoPE positions before advertising support ... it inherits the protocol stub, which returns Ellipsis; the tuple assignment in _init_mrope_positions then raises a TypeError before inference”，即只声明能力不实现方法会在请求初始化阶段崩溃。作者在两条评论下分别回复“Fixed in eee31f3b7 by removing the duplicate IsHybrid base”与“implementing text-only M-RoPE positions with three broadcast position rows and zero delta”，最终由 ZJY0516 以 LGTM 批准。合并流程中 mergify 多次提示 pre-commit 失败，tjtanaa 要求修复后在最终提交中通过。
- 重复声明 IsHybrid 基类导致 TypeError (correctness)：作者在提交 eee31f3b7 中移除重复的 IsHybrid 声明，最终 head 版本基类列表只保留一次 IsHybrid。
- 只声明 SupportsMRoPE 未实现方法导致初始化 TypeError (correctness)：作者在 eee31f3b7 中实现 get_mrope_input_positions，返回 3 行广播位置与 delta = 0 (three broadcast position rows and zero delta)。
- pre-commit 反复失败与提交前缀修复 (style)：作者修复 pre-commit 后通过，Buildkite CI #82808 触发成功。

风险与影响

- 风险：风险点：
 - 缺少单测覆盖：get_mrope_input_positions 仅依赖 ROCm 双节点手工验证，后续 M-RoPE 位置约定或 SupportsMRoPE 协议调整时回归无自动化保护。
 - 位置生成隐含纯文本假设：方法完全忽略 mm_features 参数并固定返回 delta = 0，一旦被带视觉输入的 checkpoint 复用会静默产生错误位置，当前靠 SupportsMRoPE 声明范围间接约束。
 - 能力声明范围扩散：SupportsMRoPE 与新增方法位于 Qwen3_5ForCausalLMBase，同时影响 Dense 与 MoE 两类子类，任何一方行为差异都会沿继承链传播。

- 协议耦合点：vLLM 能力协议要求“声明 + 实现”成对出现，单独加声明不会在启动期报错，而是在运行期以 `TypeError` 形式爆炸，定位成本更高。
- 影响：对 ROCm 用户，Qwen3.5 文本纯 / Qwen3.5 MoE 以及 Qwen3.8 Max FP8 类 checkpoint 现在可以不经多模态路径直接加载推理；对系统，模型注册与位置初始化路径新增了一条 M-RoPE 分支，可能影响后续所有继承该基类的模型；对团队，这是一次典型的“能力协议补齐”小改动，为 Qwen3.5 系列统一文本与多模态两条加载路径提供了铺垫。改动本身单文件、无破坏性变更，影响范围集中于模型加载与位置初始化阶段。
- 风险标记：缺少测试覆盖，能力声明与实现不一致风险，位置生成假设纯文本

关联脉络

- PR #51293 [CI] Re-enable FI autotune in GSM8K config for Qwen3.5-35B-A3B: 同属 Qwen3.5 模型家族在 CI/ 验证侧的配套工作，反映 Qwen3.5 系列正在 ROCm 与 CI 两条线上逐步铺开，与本 PR 的模型使能形成呼应。