PR #45794 完整报告

vllm-project/vllm

[Bugfix] MiniMax-M3 (AMD): add packed_modules_mapping and pass swiglu…

合并时间：2026-06-18 03:15

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45794

执行摘要

- 一句话：修复 MiniMax-M3 在 AMD 上的权重加载与 MXFP4 MoE 激活参数传递问题
- 推荐动作：该 PR 是修复关键加载问题的必要变更，代码简洁且逻辑清晰，值得精读。其中 packed_modules_mapping 的设计模式在 vLLM 中具有通用性，可用于其他类似融合权重的模型。此外，审核者 tjtanaa 提供的长上下文测试建议值得关注，建议后续补充相关测试。

功能与动机

在 AMD 平台上加载 MiniMax-M3 的 Quark MXFP4 量化检查点时，权重加载因 param_data.shape != loaded_weight.shape 而失败。根本原因是融合的 qkv_proj / gate_up_proj 模块未声明为 packed 模块，导致加载器尝试将整个融合权重张量赋给单个分片参数，产生形状不匹配。此外，视觉塔 MLP 权重的映射规则因子串末尾多余的点号而无法匹配，以及 MXFP4 MoE 路径缺少 swiglu 参数，可能引发激活结果不正确。

实现拆解

1. 在 MiniMaxM3SparseForCausalLM 中添加 packed_modules_mapping（vllm/models/minimax_m3/amd/model.py）：定义 qkv_proj -> [q_proj, k_proj, v_proj] 和 gate_up_proj -> [gate_proj, up_proj] 的映射。这使 AutoWeightsLoader 能够将融合权重正确分片写入对应子参数，解决形状不匹配断言错误。
2. 在 MiniMaxM3SparseForConditionalGeneration 中添加相同的 packed_modules_mapping（同文件）：视觉 - 语言入口类也需要此映射，以确保多模态模型权重加载的正确性。
3. 修正 hf_to_vllm_mapper 的子串匹配规则（同文件）：将 ".mlp.fc1." 改为 ".mlp.fc1"、".mlp.fc2." 改为 ".mlp.fc2"，去除末尾点号，使映射能正确匹配检查点中实际键名（不包含末尾点号），从而正确加载视觉塔 MLP 权重。
4. 向 Quark OCP MX MoE 量化配置传递 swiglu 参数（vllm/model_executor/layers/quantization/quark/quark_moe.py）：在 get_fused_moe_quant_config 的 ocp_mx_ 分支中，通过 getattr(layer, "swiglu_alpha", None) 等方式获取 swiglu_alpha、swiglu_beta、swiglu_limit，并作为 gemm1_alpha、gemm1_beta、gemm1_clamp_limit 传递给 mxfp4_w4a8_moe_quant_config，确保 MXFP4 MoE 使用正确的激活参数。

关键文件：

- vllm/models/minimax_m3/amd/model.py（模块 模型定义；类别 source；类型 data-contract）：修复了 MiniMax-M3 模型的两个关键问题：添加 packed_modules_mapping 解决权重加载形状不匹配，修正 hf_to_vllm_mapper 子串规则使视觉塔 MLP 权重正确映射。是本次变更的核心文件。
- vllm/model_executor/layers/quantization/quark/quark_moe.py（模块 量化模块；类别 source；类型 data-contract）：为 MXFP4 MoE 量化配置传递 swiglu 参数（gemm1_alpha、gemm1_beta、gemm1_clamp_limit），确保激活结果正确。是 MoE 正确性的关键补充。

关键符号：未识别

关键源码片段

vllm/models/minimax_m3/amd/model.py

修复了 MiniMax-M3 模型的两个关键问题：添加 packed_modules_mapping 解决权重加载形状不匹配，修正 hf_to_vllm_mapper 子串规则使视觉塔 MLP 权重正确映射。是本次变更的核心文件。

```
class MiniMaxM3SparseForCausalLM(nn.Module, SupportsEagle3):
    """MiniMax M3 (sparse/dense backbone) for causal language modeling."""
    # 定义融合参数到子参数的映射，指导权重加载器正确分片
    packed_modules_mapping = {
        "qkv_proj": ["q_proj", "k_proj", "v_proj"],
        "gate_up_proj": ["gate_proj", "up_proj"],
    }
    # ... 其余代码保持不变

class MiniMaxM3SparseForConditionalGeneration(
    nn.Module, SupportsMultiModal, SupportsEagle3
):
    """Top-level (VL) entry point for MiniMax M3."""
    # 同样添加 packed_modules_mapping，确保多模态权重加载正确
    packed_modules_mapping = {
        "qkv_proj": ["q_proj", "k_proj", "v_proj"],
        "gate_up_proj": ["gate_proj", "up_proj"],
    }
    # 修正子串映射：去掉末尾点号，匹配实际检查点键名
    hf_to_vllm_mapper = WeightsMapper(
        orig_to_new_prefix={
            "multi_modal_projector.": "vision_tower.multi_modal_projector.",
            "patch_merge_mlp.": "vision_tower.patch_merge_mlp.",
        },
        orig_to_new_substr={
            ".mlp.fc1": ".fc1", # 之前为 ".mlp.fc1."，无法匹配
            ".mlp.fc2": ".fc2",
        },
    )
```

vllm/model_executor/layers/quantization/quark/quark_moe.py

为 MXFP4 MoE 量化配置传递 swiglu 参数（gemm1_alpha、gemm1_beta、gemm1_clamp_limit），确保激活结果正确。是 MoE 正确性的关键补充。

```
# 在 QuarkMoEMethod.get_fused_moe_quant_config 的 else 分支中
return ocp_mx_moe_quant_config(
    quant_dtype=self.input_dtype,
    weight_dtype=self.weight_dtype,
    w1_scale=layer.w13_weight_scale,
    w2_scale=layer.w2_weight_scale,
    w1_bias=layer.w13_bias,
    w2_bias=layer.w2_bias,
    a1_scale=None,
    a2_scale=None,
    block_shape=None,
    # 传递 swiglu 参数，用于 MXFP4 MoE 激活
    gemm1_alpha=getattr(layer, "swiglu_alpha", None),
    gemm1_beta=getattr(layer, "swiglu_beta", None),
    gemm1_clamp_limit=getattr(layer, "swiglu_limit", None),
)
```

评论区精华

1. 审核与肯定：审核者 tjtanaa 和 dllehr-amd 均批准了 PR。dllehr-amd 提到自己之前有一个几乎相同的草稿 PR（#45838），表示认可本 PR 的改动。
2. 测试建议：tjtanaa 建议在更长输入（如 num_fewshot=20）上评估性能，以触发稀疏索引器逻辑，并提供了具体的 lm_eval 命令，但目前尚未看到公开的后续测试结果。
3. 外部评论：用户 Tobi-Adesoye 在 issue 评论中提及长上下文稀疏索引器下的 MXFP4 块因子方差问题，但该评论并非直接针对此 PR 改动，而是更泛化的架构讨论。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 兼容性风险（低）：packed_modules_mapping 的添加对不使用融合权重的加载场景无影响；getattr 默认值为 None，若层对象没有 swiglu_alpha 等属性，MoE 行为保持不变。
  2. 回归风险（低）：改动仅涉及两个文件中的少量配置键和参数传递，且已被集成测试验证（PR body 显示 gsm8k 准确率 94.2%）。
  3. 长上下文退化风险（未验证）：审核者指出当前测试的 short-context 场景未触发稀疏索引器，而 MXFP4 在稀疏非连续布局下可能存在数学漂移，此 PR 未涉及该问题。
- 影响：
  1. 用户影响：修复了 AMD 平台上 MiniMax-M3 MXFP4 量化模型的权重加载崩溃，用户现在可以成功加载模型并运行推理。

2. 系统影响：无性能或稳定性副作用。
3. 团队影响：为后续 MiniMax-M3 在 AMD 上的优化和特性开发扫清了障碍。影响范围仅限于 MiniMax-M3 模型及 AMD 平台。 - 风险标记：缺少长上下文测试覆盖，特定于 AMD 平台

关联脉络

- PR #45838 Draft PR for similar MiniMax-M3 MXFP4 fixes: 审核者 dllehr-amd 提到自己有几乎相同的草稿 PR#45838，说明此问题有并行解决路径。
- PR #45854 [ROCm][Quant] Minimax-M3: Enable fp8_per_channel for bf16 weights on mi300x: 同为 MiniMax-M3 在 ROCm 上的量化改进，但针对 fp8_per_channel。本 PR 与之配合可全面支持 MiniMax-M3 的 MXFP4 和 fp8 量化。