PR #23408 完整报告

sgl-project/sglang

[AMD] Fix Kimi-K2.6 Quark MXFP4 loading prefix and packed module mapping

合并时间：2026-04-27 14:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23408

执行摘要

- 一句话：修复 Kimi-K2.6 Quark MXFP4 加载的两个 bug
- 推荐动作：值得关注的设计决策：将 prefix 传递逻辑从仅限 ModelSlimConfig 扩展到 QuarkConfig，体现了类似需求应统一处理的模式。packed_modules_mapping 的扩展应逐步迁移到模型类内部声明（见 TODO 注释）。

功能与动机

Kimi-K2.6 Quark MXFP4 检查点加载失败，一是由于量化配置类不同导致 prefix 未正确传递，二是缺少必要的 packed modules 映射。PR body 明确描述了两个问题及其影响。

实现拆解

1. 在 python/sglang/srt/models/kimi_k25.py 中，KimiK25ForConditionalGeneration.__init__ 的第 665 行将 isinstance(quant_config, ModelSlimConfig) 扩展为 isinstance(quant_config, (ModelSlimConfig, QuarkConfig))，使 DeepseekV3ForCausalLM 获得 prefix="language_model"。
2. 在 python/sglang/srt/model_loader/loader.py 的 _get_quantization_config 函数中，if model_config.quantization == "quark" 分支新增 "fused_qkv_a_proj_with_mqa": ["q_a_proj", "kv_a_proj_with_mqa"] 到 packed_modules_mapping。
3. 引入依赖：在 kimi_k25.py 头部新增 from sglang.srt.layers.quantization.quark.quark import QuarkConfig。

关键文件：

- python/sglang/srt/model_loader/loader.py（模块 模型加载；类别 source；类型 data-contract；符号 _get_quantization_config）：核心改动：在 quark 量化路径中补充 fused_qkv_a_proj_with_mqa 的 packed modules 映射，使融合后的 QKV-A 投影权重能被正确加载。
- python/sglang/srt/models/kimi_k25.py（模块 模型定义；类别 source；类型 data-contract；符号 KimiK25ForConditionalGeneration）：关键改动：导入 QuarkConfig 并扩展 prefix 传递条件，确保 DeepseekV3ForCausalLM 在 Quark 量化时也获得 language_model 前缀。

关键符号：_get_quantization_config, KimiK25ForConditionalGeneration

关键源码片段

python/sglang/srt/model_loader/loader.py

核心改动：在 quark 量化路径中补充 fused_qkv_a_proj_with_mqa 的 packed modules 映射，使融合后的 QKV-A 投影权重能被正确加载。

```
def _get_quantization_config(
    model_config: ModelConfig,
    load_config: LoadConfig,
) -> Optional[QuantizationConfig]:
    """Get the quantization config."""
    model_class, _ = get_model_architecture(model_config)
    packed_modules_mapping = getattr(model_class, "packed_modules_mapping", {})
    remap_prefix = getattr(model_class, "remap_prefix", None)
    # TODO: we should remove this code and switch to the packed_modules_mapping declared
    inside the modeling files
    if model_config.quantization == "quark":
        # 新增 fused_qkv_a_proj_with_mqa 映射，用于 Kimi-K2.6 融合 QKV-A 投影权重加载
        packed_modules_mapping.update(
            {
                "gate_up_proj": ["gate_proj", "up_proj"],
                "fused_qkv_a_proj_with_mqa": ["q_a_proj", "kv_a_proj_with_mqa"],
            }
        )

    if _is_npu:
        # NPU 下已有同样的映射，此处为 quark 路径补充
        packed_modules_mapping.update(...)
```

python/sglang/srt/models/kimi_k25.py

关键改动：导入 QuarkConfig 并扩展 prefix 传递条件，确保 DeepseekV3ForCausalLM 在 Quark 量化时也获得 language_model 前缀。

```
from sglang.srt.layers.quantization.quark.quark import QuarkConfig

class KimiK25ForConditionalGeneration(nn.Module):
    def __init__(self, config, quant_config=None, prefix="", **kwargs):
        # ...
        self.language_model = None
        if not config.encoder_only:
            self.language_model = DeepseekV3ForCausalLM(
                config.text_config,
                quant_config,
                prefix=(
                    "language_model"
                    # 扩展条件：QuarkConfig 也需要传递 language_model 前缀
                    if isinstance(quant_config, (ModelSlimConfig, QuarkConfig))
                    else ""
                ),
```

)

评论区精华

无 review 评论。审核人 BowenBao 和 HaiShaw 均直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更范围小（2 个文件，共 8 行新增，2 行删除），仅影响 Quark 量化路径中的 prefix 传递和模块映射，不改变核心推理逻辑。未添加测试，但改动简单且已验证 gsm8k 准确率 93.5%。
- 影响：直接影响：使 Kimi-K2.6 Quark MXFP4 检查点在 AMD 等平台上正确加载。间接影响：packed_modules_mapping 的扩展可能与其他量化后端（如 NPU）的映射产生重叠，但本 PR 的 fused_qkv_a_proj_with_mqa 映射与 NPU 分支内容一致，冲突概率低。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR