

PR #36211 完整报告

sgl-project/sglang

[k3] declare packed_modules_mapping on `KimiK3ForConditionalGeneration`

合并时间: 2026-08-28 10:22

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36211>

执行摘要

- 一句话: 为 KimiK3 声明 packed_modules_mapping, 修复量化 exclude 匹配
- 推荐动作: 该 PR 虽小, 但属于数据契约修正, 值得精读以理解 packed_modules_mapping 的约定, 并可作为其他模型类补充映射的参考。建议关注后续是否有相关测试补充。

功能与动机

KimiK3 的运行时会将 checkpoint 分片融合为合并模块, 但模型类未声明 packed_modules_mapping, 导致量化配置无法按分片准确匹配 exclude_modules。PR body 指出, NVIDIA 的 Kimi-K3-NVFP4 配置排除了 self_attn.q_conv1d 等, 但融合后模块 qkv_conv1d 仅因 'v_conv1d' 子串偶然匹配而被排除, 其他 shard 匹配不到, 导致量化错误。

实现拆解

1. 定位变更入口: 在 python/sglang/srt/models/kimi_k3.py 的 KimiK3ForConditionalGeneration 类定义中新增 packed_modules_mapping 类属性。
2. 补充映射关系: 按照 SGLang 统一的 packed modules 映射约定 (与 80+ 其他模型类一致), 将融合模块名 (如 gate_up_proj、qkv_conv1d、fused_qkvg_proj 等) 映射到其原始 checkpoint 分片名称列表。
3. 无配套测试: 本变更未新增或修改测试文件, 主要依赖现有量化配置解析逻辑的验证。

关键文件:

- python/sglang/srt/models/kimi_k3.py (模块 模型层; 类别 source; 类型 data-contract; 符号 packed_modules_mapping): 新增 packed_modules_mapping 类属性, 是本次变更的核心。

关键符号: 未识别

关键源码片段

[python/sglang/srt/models/kimi_k3.py](#)

新增 packed_modules_mapping 类属性, 是本次变更的核心。

```
# python/sglang/srt/models/kimi_k3.py
# 类 KimiK3ForConditionalGeneration 中新增 packed_modules_mapping
```

```
packed_modules_mapping = {
    # 调度器融合的 MLP 投影（up + 上投影），映射到原始分片
    "gate_up_proj": ["gate_proj", "up_proj"],
    # 标准自注意力 QKV 融合，映射到原始分片
    "qkv_proj": ["q_proj", "k_proj", "v_proj"],
    # KimiK3 自注意力中特有的 conv1d 融合模块，映射到原始分片
    "qkv_conv1d": ["q_conv1d", "k_conv1d", "v_conv1d"],
    # 融合了 Q、K、V 和门控投影的模块
    "fused_qkvg_proj": ["q_proj", "k_proj", "v_proj", "g_proj"],
    # 更大规模的融合模块，包含 Q、K、V、B 以及两个专家子投影
    "fused_qkvbfg_a_proj": ["q_proj", "k_proj", "v_proj", "b_proj", "f_a_proj", "g_a_proj"],
    # 输出侧的门控融合模块
    "fused_fg_b_proj": ["f_b_proj", "g_b_proj"],
}
```

评论区精华

该 PR 无 review 评论，仅有两位 reviewer 的 APPROVED。讨论内容主要来自 PR body 中对问题的简述，未出现明显的争议或设计权衡。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 映射变更是纯数据声明，不涉及执行路径改动，风险较低。
 - 潜在遗漏：若 packed_modules_mapping 中的映射与模型实际 fusing 逻辑不一致（例如新增的融合模块），可能导致后续量化配置匹配异常。
 - 无测试覆盖：缺少针对该映射的单元测试，无法确保所有分片映射正确，后续重构或新增融合模块时易回归。
- 影响：
 - 用户影响：主要影响使用 KimiK3 模型且启用 NVFP4 等量化配置的用户，修复后能够正确排除指定分片，避免量化错误。
 - 系统影响：对运行速度、显存使用等无明显影响，属于配置修复。
 - 团队影响：为后续 KimiK3 模型量化配置的正确性提供了基础。
 - 风险标记：缺少测试覆盖，数据契约变更，模型加载路径

关联脉络

- PR #34842 Revert "[Fix] Disable --enable-symm-mem under CUDA graphs on Kimi hybrid models": 同为 Kimi 模型相关改动，涉及模型配置与硬件后端适配。
- PR #36620 config: a parallel leaf with no live counterpart is read bare: 同为 config 数据契约相关重构，涉及模型配置读取方式的调整。