PR #23471 完整报告

sgl-project/sglang

[Fix] NVFP4 qwen3.5 quant error fix by add packed_modules_mapping

合并时间：2026-04-29 04:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23471

执行摘要

- 一句话：修复 NVFP4 量化导致 Qwen3.5 精度下降
- 推荐动作：建议精读该 PR 以理解 SGLang 量化打包映射的工作原理。虽然改动简单，但涉及量化精度关键设计，值得关注。

功能与动机

在 NVFP4 量化模式下运行 Qwen3.5-A35B 时，许多 linear_attn_in_proj 层日志中显示为被量化，导致计算精度受损。具体表现为 bench_one_batch 结果异常。PR 作者指出 'the computation precision would be damaged if these layers were quantized'。

实现拆解

1. 在 Qwen3_5ForCausalLM 类中，将 packed_modules_mapping 字典的定义从 if _is_gfx95 or _is_npu: 条件块中移出，变为无条件定义。该映射将逻辑分组（如 qkv_proj, gate_up_proj）映射到具体子层（如 q_proj, k_proj, v_proj），用于在加载权重时识别哪些子层属于同一打包组，从而应被跳过量化。
2. 在 Qwen3_5ForConditionalGeneration 和 Qwen3_5MoeForConditionalGeneration 类中，同样移除了 if _is_gfx95 or _is_npu: 条件，使 packed_modules_mapping 和 hf_to_sglang_mapper = None 赋值无条件执行。这两个类分别对应纯文本模型和 MoE 视觉语言模型。
3. 编码调整删除 13 行、新增 11 行，净减少 2 行。文件变更仅限 python/sglang/srt/models/qwen3_5.py。

关键文件：

- python/sglang/srt/models/qwen3_5.py（模块 模型定义；类别 source；类型 data-contract；符号 Qwen3_5ForCausalLM, Qwen3_5ForConditionalGeneration, Qwen3_5MoeForConditionalGeneration）：核心修复文件，移除 packed_modules_mapping 和 hf_to_sglang_mapper 的条件守卫，使其在所有环境下生效。

关键符号：未识别

关键源码片段

python/sglang/srt/models/qwen3_5.py

核心修复文件，移除 packed_modules_mapping 和 hf_to_sglang_mapper 的条件守卫，使其在所有环境下生效。

```
# python/sglang/srt/models/qwen3_5.py 中 Qwen3_5ForCausalLM 类的变更
class Qwen3_5ForCausalLM(nn.Module):
    """Qwen3.5 Model with support for dense variant."""

    # 移除条件守卫，使 packed_modules_mapping 在所有环境中生效
    packed_modules_mapping = {
        "qkv_proj": ["q_proj", "k_proj", "v_proj"],
        "gate_up_proj": ["gate_proj", "up_proj"],
        "in_proj_qkvz": ["in_proj_qkv", "in_proj_z"],
        "in_proj_ba": ["in_proj_b", "in_proj_a"],
    }

    # ...
```

评论区精华

评审过程中，gemini-code-assist[bot] 自动评论确认了变更效果：条件守卫被移除后，打包映射在所有环境中可用，不再受特定硬件限制。yizhang2077 两次 Approved。无额外讨论。

- 暂无高价值评论线程

风险与影响

- 风险：变更范围极小，仅涉及类属性的定义条件，不涉及运行时逻辑、权重加载或推理路径。风险极低。但需注意：如果 NPU 或其他后端未来需要不同的 packed_modules_mapping 值，此修改可能引入覆盖问题。目前映射定义相同，因此无实际回归风险。
- 影响：对用户：在 CUDA 等非 NPU/GFX95 环境下运行 NVFP4 量化 Qwen3.5 模型时，linear_attn 等层不再被错误量化，精度恢复正常。对系统：无额外依赖或性能影响。对团队：小的修复，影响面窄。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR