PR #22961 完整报告

sgl-project/sglang

[NPU] Fix issue and support GLM-4.5V

合并时间：2026-04-28 09:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22961

执行摘要

- 一句话：NPU 支持 GLM-4.5V 并修复 QK Norm 参数传递
- 推荐动作：建议阅读 glm4_moe.py 中 forward_prepare 的条件分支设计，了解 NPU 后端如何处理 QK Norm 可选的情况。对于其他需要类似支持的模型可复用此模式。

功能与动机

关联 Issue#343 报告 ZhipuAI/GLM-4.5V 模型在 NPU 环境服务启动失败。PR body 指出问题在于调用 split_qkv_rmsnorm_rope 时未根据 use_qk_norm 参数传递正确的参数，而该内核已内部支持 NORMS=False 模式，需要通过条件性传入 None 值来适配无 QK Norm 的模型。

实现拆解

1. 在 python/sglang/srt/models/glm4_moe.py 的 forward_prepare 方法中，NPU decode 路径下添加 use_qk_norm 条件判断。若为 True 则从 QK Norm 层提取 eps、weight、bias；否则全部设为 None。
2. 将提取后的变量传递给 split_qkv_rmsnorm_rope 调用，替换原先直接访问属性的逻辑。内核在参数为 None 时跳过 Norm 计算，从而兼容无 QK Norm 的模型。
3. 清理了一个冗余空行，使代码风格一致。

关键文件：

- python/sglang/srt/models/glm4_moe.py（模块 模型层；类别 source；类型 data-contract；符号 forward_prepare）：唯一被修改的文件，包含对 Glm4MoeAttention.forward_prepare 中 QKV 拆分的条件逻辑调整，直接解决 GLM-4.5V 在 NPU 上启动失败问题。

关键符号：forward_prepare

关键源码片段

python/sglang/srt/models/glm4_moe.py

唯一被修改的文件，包含对 Glm4MoeAttention.forward_prepare 中 QKV 拆分的条件逻辑调整，直接解决 GLM-4.5V 在 NPU 上启动失败问题。

```
# python/sglang/srt/models/glm4_moe.py (forward_prepare 方法，NPU decode 分支 )
def forward_prepare(
```

```
    self,
    positions: torch.Tensor,
    hidden_states: torch.Tensor,
    forward_batch: ForwardBatch,
):
    # hidden_states can be a (fp8_tensor, scale) tuple from fused RMSNorm+Quant
    hs = hidden_states[0] if isinstance(hidden_states, tuple) else hidden_states
    if hs.shape[0] == 0:
        return hidden_states, forward_batch, None
    qkv, _ = self.qkv_proj(hidden_states)

    if (
        not _is_npu
        or forward_batch.forward_mode.is_extend_or_draft_extend_or_mixed()
    ):
        # GPU 路径或 extend mode，直接 split 并应用 norm
        q, k, v = qkv.split([self.q_size, self.kv_size, self.kv_size], dim=-1)
        if self.use_qk_norm:
            q, k = apply_qk_norm(
                q=q, k=k,
                q_norm=self.q_norm, k_norm=self.k_norm,
                head_dim=self.head_dim, alt_stream=self.alt_stream,
            )
        q, k = self.rotary_emb(positions, q, k)
    else:
        # NPU decode 路径：使用 fused split_qkv_rmsnorm_rope kernel
        if self.attn.layer_id == forward_batch.token_to_kv_pool.start_layer:
            self.rotary_emb.get_cos_sin_with_position(positions)
        # --- 以下是本次修复的关键逻辑 ---
        if self.use_qk_norm:
            eps = self.q_norm.variance_epsilon
            q_weight = self.q_norm.weight
            k_weight = self.k_norm.weight
            q_bias = getattr(self.q_norm, "bias", None)
            k_bias = getattr(self.k_norm, "bias", None)
        else:
            # 模型不使用 QK Norm 时，传入 None 让 kernel 跳过 norm 步骤
            eps = None
            q_weight = None
            k_weight = None
            q_bias = None
            k_bias = None
        q, k, v = split_qkv_rmsnorm_rope(
            qkv,
            self.rotary_emb.position_sin,
            self.rotary_emb.position_cos,
            self.q_size,
            self.kv_size,
            self.head_dim,
```

```
            eps=eps, # 之前直接传 self.q_norm.variance_epsilon
            q_weight=q_weight,# 之前直接传 self.q_norm.weight
            k_weight=k_weight,# 之前直接传 self.k_norm.weight
            q_bias=q_bias, # 之前直接传 getattr(self.q_norm, "bias", None)
            k_bias=k_bias, # 之前直接传 getattr(self.k_norm, "bias", None)
        )

    inner_state = q, k, v, forward_batch
    return None, forward_batch, inner_state
```

评论区精华

机器人评审 gemini-code-assist[bot] 建议将 NPU 条件分支重构为布尔变量 use_npu_decode_path 以提高可读性，但该建议未在最终代码中采纳。作者和审核人认为当前写法已足够清晰，直接合并。

- 复杂条件逻辑重构建议 (design): 未采纳，作者未回复，PR 直接合并。

风险与影响

- 风险：变更范围极小，仅修改一个文件中的单个方法的分支逻辑。风险在于条件判断的精确性：若未来有模型在 NPU decode 路径需 QK Norm 但该条件遗漏，可能导致静默错误，但当前所有 GLM 模型已覆盖。无性能影响，因为分支代码只在 decode 路径执行，且增加了条件判断但开销可忽略。
- 影响：直接影响 NPU（Ascend）后端上 GLM-4.5V 模型的启动可用性，使得原无法启动的模型可以在 NPU 上运行并达到 MMMU 准确率 0.2802。对其他后端（CUDA、AMD 等）无影响，因为条件 _is_npu 控制了 NPU 专属路径。对已有 GLM 模型无退化。
- 风险标记：NPU 专属路径，单文件改动

关联脉络

- 暂无明显关联 PR