PR #29505 完整报告

sgl-project/sglang

[NPU] Qwen3-VL-30B use split_qkv_rmsnorm_rope for extend

合并时间：2026-06-29 14:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29505

执行摘要

- 一句话：NPU 上 Qwen3-VL 扩展阶段路由到 fused 算子
- 推荐动作：建议精读此 PR 以了解 NPU 上 Qwen3-VL 的精度修复方案。重点关注 forward_prepare_npu 对 MRoPE 的处理是否完善，建议补充针对 NPU 上 Qwen3-VL-30B 的 extend/decode 的精度与正确性测试。讨论中 AI reviewer 的质疑虽然未得到回应，但值得开发者验证。

功能与动机

当 Qwen3-VL-30B 在 NPU 上使用 torch_npu.npu_mrope 接口时遇到精度问题，需要调整调用路径，在 extend 阶段使用 fused 算子 split_qkv_rmsnorm_rope 来绕过该接口。

实现拆解

1. 修改入口：在 python/sglang/srt/models/qwen3_moe.py 的 Qwen3MoeAttention.forward_prepare 方法中，将原来的条件判断从 not _is_npu or forward_batch.forward_mode.is_extend_or_draft_extend_or_mixed() 简化为 not _is_npu。
2. 核心逻辑：原逻辑中，非 NPU 或 extend/draft-extend/mixed 模式下走 forward_prepare_native，否则走 forward_prepare_npu。改动后，非 NPU 走 native，NPU 无论何种 forward mode（extend、decode 等）都走 forward_prepare_npu。
3. 影响：NPU 上的 extend 阶段不再执行 forward_prepare_native 中的 self.rotary_emb（MRoPE）调用，而是由 forward_prepare_npu 中的 fused 算子 split_qkv_rmsnorm_rope 替代，解决了精度问题。

关键文件：

- python/sglang/srt/models/qwen3_moe.py（模块 模型模块；类别 source；类型 core-logic；符号 Qwen3MoeAttention.forward_prepare）：核心变更文件，修改了 Qwen3MoeAttention.forward_prepare 中的 NPU 分支逻辑，影响 NPU 上 Qwen3-VL-30B 的 attention 前向准备路径。

关键符号：Qwen3MoeAttention.forward_prepare

关键源码片段

python/sglang/srt/models/qwen3_moe.py

核心变更文件，修改了 Qwen3MoeAttention.forward_prepare 中的 NPU 分支逻辑，影响 NPU 上 Qwen3-VL-30B 的 attention 前向准备路径。

```
# python/sglang/srt/models/qwen3_moe.py
# 修改前：对于 NPU，仅在非 extend 模式时走 npu 路径；修改后：NPU 一律走 npu 路径
class Qwen3MoeAttention(nn.Module):
    def forward_prepare(
        self,
        positions: torch.Tensor,
        hidden_states: torch.Tensor,
        forward_batch: ForwardBatch,
    ):
        if hidden_states.shape[0] == 0:
            return hidden_states, forward_batch, None
        # 原条件：  not _is_npu or is_extend... → 非 NPU 或 extend 模式走 native
        # 现条件：  not _is_npu → 只有非 NPU 走 native，NPU 全部走 npu（含 extend）
        if not _is_npu:
            return self.forward_prepare_native(
                positions=positions,
                hidden_states=hidden_states,
                forward_batch=forward_batch,
            )
        else:
            return self.forward_prepare_npu(
                positions=positions,
                hidden_states=hidden_states,
                forward_batch=forward_batch,
            )
```

评论区精华

Review 中 gemini-code-assist[bot] 指出一个高风险问题：forward_prepare_npu 内部使用 1D positions 调用 self.rotary_emb.get_cos_sin_with_position(positions)，但 Qwen3-VL-30B 的 self.rotary_emb 是 MRotaryEmbedding 实例，期望多维 mrope_positions（形状 [3, seq_len]）。如果 MRoPE 的 get_cos_sin 方法不支持 1D positions，可能导致运行时错误或错误 embedding。但该评论未被人类 reviewer 进一步讨论，PR 最终由 sglang-npu-bot 批准合并。

- MRoPE 兼容性问题 (correctness): 未得到明确解决或回应；PR 由机器人批准合并。

风险与影响

- 风险：
  1. MRoPE 兼容性风险：forward_prepare_npu 使用 1D positions 调用旋转位置编码，对于使用 MRoPE（多维 RoPE）的模型（如 Qwen3-VL-30B），若 forward_prepare_npu 内部未适配多维 positions，可能产生错误 embedding 或运行时崩溃。目前代码中未见显式处理。

2. 回归风险：NPU 非 *extend* 模式（如 *decode*）原本也走 forward_prepare_npu，不受影响；但 NPU *extend* 模式之前走 *native* 路径，现在切换到 npu 路径，可能引入不同于精度问题的行为变更。
3. 缺少测试：本次变更未附带单元测试或集成测试，无法确认 NPU 上各 forward mode 的正确性。

- 影响：
  1. 用户影响：NPU 上使用 Qwen3-VL-30B 的用户在 *extend* 阶段将自动使用 fused 算子，精度问题预期得到修复，但若 forward_prepare_npu 未适配 MRoPE 则可能引入新问题。
  2. 系统影响：仅影响 NPU 后端，CUDA 等其他后端无变化。
  3. 影响程度：中等——单文件 5 行变更，但涉及模型前向核心路径（attention prepare 阶段），且与 MRoPE 的兼容性存在不确定性。 - 风险标记：核心路径变更，缺少测试覆盖，MRoPE 兼容性未知

关联脉络

- PR #29598 [NPU][Bugfix] Accept in_capture in Ascend replay metadata: 同为 NPU 后端的 bugfix，反映出近期 NPU 相关修复活跃，可能涉及相同 attention 或调度路径。
- PR #29029 [NPU][Bugfix] Fix a ModelSlim loading failure: 同为 NPU 相关 bugfix，表明团队正在系统性地解决 NPU 上多模型兼容性问题。