PR #29627 完整报告

sgl-project/sglang

[NPU] Qwen3-VL-8B use split_qkv_rmsnorm_rope for extend

合并时间：2026-06-30 15:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29627

执行摘要

- 一句话：NPU 上 Qwen3-VL extend 阶段路由到 fused 算子
- 推荐动作：建议审核者关注 review 评论中提到的 mrope_positions 问题，确认 forward_prepare_npu 是否已适配 Qwen3-VL 的 3D 位置编码需求。建议补充精度测试，确保精度不退化。

功能与动机

当 Qwen3-VL-8B 使用 torch_npu.npu_mrope 接口时出现精度问题，因此需要调整调用路径：在 extend 阶段利用 fused 算子 split_qkv_rmsnorm_rope 来绕过该接口。

实现拆解

1. 在 python/sglang/srt/models/qwen3.py 的 Qwen3Attention.forward() 方法中，简化了分支条件判断。
2. 原条件 not _is_npu or forward_batch.forward_mode.is_extend_or_draft_extend_or_mixed() 被简化为 not _is_npu。
3. 这意味着在 NPU 设备上，无论当前是 decode 模式（原本就走 forward_prepare_npu）还是 extend/draft extend/mixed 模式（原本走 forward_prepare_native），现在都路由到 forward_prepare_npu。
4. 非 NPU 设备上的行为保持不变，仍走 forward_prepare_native。

关键文件：

- python/sglang/srt/models/qwen3.py（模块 Qwen3 模型；类别 source；类型 core-logic；符号 Qwen3Attention.forward）：该文件是 PR 的唯一变更文件，修改了 Qwen3Attention.forward() 中的分支逻辑，影响 NPU 上 Qwen3-VL 的 extend 阶段算子选择。

关键符号：Qwen3Attention.forward

关键源码片段

python/sglang/srt/models/qwen3.py

该文件是 PR 的唯一变更文件，修改了 Qwen3Attention.forward() 中的分支逻辑，影响 NPU 上 Qwen3-VL 的 extend 阶段算子选择。

defforward(self, positions: torch.Tensor, hidden_states: torch.Tensor, forward_batch: ForwardBatch,) -> torch.Tensor: # ... 前置代码 ... save_kv_cache = True use_aiter_fused = (self.use_fused_qk_norm_mrope and forward_batch.forward_mode.is_decode() and get_global_server_args().rl_on_policy_target is None) if use_aiter_fused: q, k, v = self.forward_prepare_aiter_fused_mrope(positions, hidden_states, forward_batch) save_kv_cache = False elif not _is_npu: # 非 NPU：走原生路径，支持 extend / decode q, k, v = self.forward_prepare_native(positions=positions, hidden_states=hidden_states,) else: # NPU：统一走 fused 算子路径（包括 extend 阶段），以绕过 npu_mrope 精度问题 q, k, v = self.forward_prepare_npu(positions=positions, hidden_states=hidden_states, forward_batch=forward_batch,) # ... 后续注意力计算 ... return output （注意：forward_prepare_npu 内部调用 self.rotary_emb.get_cos_sin_with_position(positions) 使用 1D positions，但 Qwen3-VL 可能需要 3D mrope_positions —— 这是 review 指出的潜在风险。）

评论区精华

review 评论（来自 gemini-code-assist[bot]）指出：将 NPU extend/prefill 路径路由到 forward_prepare_npu 后，后者目前传递 1D positions 张量给 rotary embedding，但 Qwen3-VL 使用 MRotaryEmbedding（mRoPE），需要 3D 位置坐标才能正确计算多维旋转嵌入。建议使用 forward_batch.mrope_positions（如果可用）来防止精度下降。该评论的回复或后续处理未显示，但 PR 最终被 sglang-npu-bot 批准。

- mrope_positions 缺失可能导致精度退化 (correctness): 未明确结论，但 PR 被批准并合并，可能该问题已经在其他 PR 中处理或当前不触发。

风险与影响

- 风险：
  1. 精度风险：forward_prepare_npu 目前使用 1D positions 而非 3D mrope_positions，可能导致 Qwen3-VL 的旋转嵌入计算不正确，影响模型输出精度（评论已指出）。
  2. 回归风险：改动范围小，仅影响 NPU 上的 extend 路径，非 NPU 设备无影响，但缺乏测试覆盖。
  3. 缺少测试：PR 未附带测试文件变更，无法验证新路径的正确性和精度。 - 影响：影响范围限于 NPU 设备上的 Qwen3-VL-8B 模型。对于该模型，extend（预填充）阶段将使用 fused 算子 split_qkv_rmsnorm_rope 替代原始算子，预期解决精度问题，但若 mrope_positions 未被正确使用，可能引入新的精度偏差。非 NPU 设备和 NPU decode 阶段无影响。 - 风险标记：缺少测试覆盖，潜在精度退化

关联脉络

- PR #29420 [AMD][DSV4] Remove per-batch D2H syncs in MTP to avoid bubbles between 2 batches: 同为硬件平台（AMD/NPU）相关的性能 / 精度优化 PR，涉及 forward 路径调整。

- PR #29642 [AMD] Copy decode result on forward_stream instead of copy_stream: 同为硬件平台（AMD）的调度 / 前向优化，类似的本 PR 涉及 NPU 前向路径。