PR #28464 完整报告

sgl-project/sglang

[spec decoding] fix mrope_positions in draft extend

合并时间：2026-06-18 04:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28464

执行摘要

- 一句话：修复 draft-extend 阶段 mrope_positions 崩溃
- 推荐动作：建议精读。该 PR 修正了一个明确的运行时崩溃，修复逻辑清晰，通过新增分支复用已有方法，设计简洁。值得关注的是如何通过函数签名扩展（添加可选参数）来统一不同阶段的 mrope 位置计算。

功能与动机

多模态 VL 模型（如 Qwen3.5-9B）在启用 NEXTN speculative decoding 且 batch 包含图文混合请求时，draft-extend 阶段因 mrope 位置计算错误导致崩溃。PR body 提供了复现步骤，并说明 'This will crash before this pr'。

实现拆解

1. 新增分支条件：在 forward_batch_info.py 的 init_new 方法中，原有的 mrope 分支（if ret.spec_info is not None and ...）之后，新增 elif ret.forward_mode.is_draft_extend_v2() 分支，专门处理 draft-extend 模式。
2. 复用 spec mrope 路径：新的分支调用 compute_spec_mrope_positions 方法，并传入 seq_positions=ret.positions 参数——其中 ret.positions 是输入侧一致的连续位置，而非每个请求独自重建的含 mm 偏移的位置，从而避免长度不匹配。
3. 扩展方法签名：compute_spec_mrope_positions 方法增加可选参数 seq_positions=None。当该参数为 None 时，沿用原逻辑从 batch.spec_info.positions 读取（用于 target_verify/draft_decode）；当传入时，使用传入的位置张量（用于 draft_extend）。
4. 逻辑对齐：修改后的路径确保 draft-extend 模式下 mrope 位置能与图文混合请求正确对齐，避免因 mm 请求大小不匹配导致的 crash。

关键文件：

- python/sglang/srt/model_executor/forward_batch_info.py（模块 前向批处理；类别 source；类型 data-contract；符号 compute_spec_mrope_positions, init_new）：核心变更文件。新增 mrope 分支处理 draft-extend 模式，并扩展 compute_spec_mrope_positions 方法签名以支持外部传入位置参数。

关键符号：compute_spec_mrope_positions, init_new

关键源码片段

python/sglang/srt/model_executor/forward_batch_info.py

核心变更文件。新增 mrope 分支处理 draft-extend 模式，并扩展 compute_spec_mrope_positions 方法签名以支持外部传入位置参数。

```
def init_new(...):
    # ... 前面的逻辑
    if model_runner.model_is_mrope:
        if (
            ret.spec_info is not None
            and getattr(ret.spec_info, "positions", None) is not None
        ):
            ret.compute_spec_mrope_positions(model_runner, batch)
        elif ret.forward_mode.is_draft_extend_v2():
            # Draft-extend tokens are uniform text continuation; reuse the
            # spec mrope path with the input-consistent `ret.positions` rather
            # than the per-request rebuild (which mis-sizes mm requests).
            ret.compute_spec_mrope_positions(
                model_runner, batch, seq_positions=ret.positions
            )
        else:
            ret._compute_mrope_positions(model_runner, batch)
```

```
def compute_spec_mrope_positions(
    self, model_runner: ModelRunner, batch: ScheduleBatch, seq_positions=None
):
    batch_size = self.seq_lens.shape[0]
    device = model_runner.device
    mm_inputs = batch.multimodal_inputs

    # target_verify / draft_decode read spec_info.positions; draft_extend
    # passes its own positions (uniform num_draft_tokens per request).
    if seq_positions is None:
        seq_positions = batch.spec_info.positions
    seq_positions = seq_positions.view(batch_size, -1)
    # ... 后续 mrope delta 计算逻辑不变
```

评论区精华

Reviewer kpham-sgl 对新增的注释提出了简洁性建议（nit: make this more concise），但未修改代码即已批准。未发现其他争议。

- draft-extend 分支 mrope 位置修复 (correctness): 注释保持原样，PR 即被批准合并。

风险与影响

- 风险：本次变更仅影响 forward_batch_info.py 中 mrope 分支的控制流和 compute_spec_mrope_positions 的函数签名，改动范围小（+13/-3）。风险较低，但需确认 ret.forward_mode.is_draft_extend_v2() 在所有 draft-extend 场景下均被正确设置，避免遗漏。另外，新增的可选参数 seq_positions 若在其他调用点被错误传入，可能导致意外

行为。

- 影响：影响范围限于使用多模态模型（如 Qwen3.5-9B）并启用 speculative decoding 且 draft-extend 模式的用户。修复后这些场景不再崩溃，正常功能得到恢复。非多模态模型或不用 speculative decoding 的场景无影响。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #28465 Batch EAGLE draft/draft-extend replay memcpys via grouped foreach copy: 同一功能线（draft-extend 性能优化），共享相同的 forward_batch_info.py 文件，可能与本 PR 的 mrope 修复有交互。
- PR #28500 [Perf] Make spec-decode penalty H2D non-blocking and share decode cumulate path: 同样是 speculative decoding 的性能优化，影响 schedule_batch.py 和 eagle_info_v2.py，与本 PR 的 mrope 逻辑在调度层可能有间接关联。