PR #24319 完整报告

sgl-project/sglang

[AMD] fix tbo specv2 seq_lens_cpu NoneType error

合并时间：2026-05-05 16:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24319

执行摘要

- 一句话：修复 EAGLE SpecV2 + TBO 下 seq_lens_cpu 空指针问题
- 推荐动作：值得快速合入，修复明确，影响范围小。可作为 AMD 平台 SpecV2 兼容性修复的参考模式。

功能与动机

CI 任务中 TestMTPwithTBOLowLatency 测试因 AttributeError: 'NoneType' object has no attribute 'shape' 崩溃，根因是 EagleVerifyInput 的 seq_lens_cpu 为 None。关联 issue #24212 追踪此兼容性问题，并参考了 PR #24205 的 workaround。

实现拆解

仅在 python/sglang/srt/speculative/eagle_info_v2.py 的 prepare_for_v2_verify 方法中，在 ForwardBatch.init_new 调用之前，增加了两行赋值语句：

1. self.seq_lens_cpu = batch.seq_lens_cpu
2. self.seq_lens_sum = batch.seq_lens_sum

这些赋值被放置在一个 if not batch.forward_mode.is_idle(): 条件块内，与之前的 draft token 处理逻辑条件一致，确保仅在非空闲模式下填充。该修复同时覆盖了 multi_layer_eagle_worker_v2.py 中的路径，因为它们共用 EagleVerifyInputV2Mixin。

关键文件：

- python/sglang/srt/speculative/eagle_info_v2.py（模块 推测解码；类别 source；类型 core-logic；符号 prepare_for_v2_verify）：唯一修改的文件，在 prepare_for_v2_verify 方法中填充 seq_lens_cpu/seq_lens_sum 以修复 TBO 路径下的空指针崩溃。

关键符号：prepare_for_v2_verify

关键源码片段

python/sglang/srt/speculative/eagle_info_v2.py

唯一修改的文件，在 prepare_for_v2_verify 方法中填充 seq_lens_cpu/seq_lens_sum 以修复 TBO 路径下的空指针崩溃。

```
# python/sglang/srt/speculative/eagle_info_v2.py
```

```
# 在 prepare_for_v2_verify 中，于 ForwardBatch.init_new 之前填充
if not batch.forward_mode.is_idle():
    # ... 之前已有的 draft token 处理逻辑 ...
    batch.mamba_track_mask = None
    batch.mamba_track_seqlens = None

    # ---- 新增：从 batch 复制 seq_lens_cpu / seq_lens_sum 到 self ----
    # 保证 TBO 的 split_spec_info 在切片 custom_mask 时能读取到有效值
    self.seq_lens_cpu = batch.seq_lens_cpu
    self.seq_lens_sum = batch.seq_lens_sum

# 后续创建 ForwardBatch 并返回
batch.forward_mode = (
    ForwardMode.IDLE
    if batch.forward_mode.is_idle()
    else ForwardMode.TARGET_VERIFY
)
batch.capture_hidden_mode = CaptureHiddenMode.FULL
verify_forward_batch = ForwardBatch.init_new(batch, target_worker.model_runner)
# ... 后续返回 verify_forward_batch, can_run_cuda_graph
```

评论区精华

Review 中 gemini-code-assist[bot] 提出了代码风格建议：将新增的两行赋值移到已有的 `if not batch.forward_mode.is_idle():` 条件块内部，以避免冗余的条件判断。作者在第二次提交中采纳了该建议。HaiShaw 和 hubertlu-tw 均批准了 PR。

- 将新增赋值移入已有条件块 (style): 作者在 commit 5553c474 中采纳了该建议。

风险与影响

- 风险：风险极低：变更仅在两处条件分支内增加了 GPU 张量引用赋值，不涉及新依赖或计算逻辑，且 CI 测试 TestMTPwithTBOLowLatency 已通过。回归风险主要在于 TBO 未启用时该赋值不会执行，无影响。
- 影响：仅影响 AMD 平台启用 EAGLE SpecV2 且开启 TBO 优化时的工作流，修复了 SpecV2 默认启用后的崩溃问题。对非 AMD 或非 SpecV2 场景无影响。
- 风险标记：核心路径变更

关联脉络

- PR #24205 [AMD] workaround tbo specv2 crash: 本 PR 的修复正是替代该 workaround 的正式解决方案。
- PR #23146 [AMD] Enable EAGLE speculative decoding for Qwen3.5 FP8 and MXFP4 models with aiter's unified attention: 同一功能线：AMD 平台对 EAGLE 推测解码的持续支持。