PR #26655 完整报告

sgl-project/sglang

Fix TRTLLM MHA draft decode cache seqlens replay

合并时间：2026-05-29 11:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26655

执行摘要

- 一句话：修复 TRTLLM MHA draft decode 缓存序列长度重放
- 推荐动作：值得立即合并，修复了可能导致不正确推理结果的 bug。建议未来为该路径增加单元测试。

功能与动机

确保 draft decode 场景下 CUDA graph 重放时，缓存序列长度（cache_seqlens）正确绑定到当前 batch 子集，避免数据竞争或错误的内存写入。

实现拆解

1. 在 init_forward_metadata_replay_cuda_graph 方法的 Draft Decode 分支中，将 metadata.cache_seqlens_int32 重新绑定到 self.decode_cuda_graph_metadata["cache_seqlens"][:bs] 切片，确保指向当前 batch 的预分配缓冲区。
2. 然后调用 copy_ 将更新后的序列长度（seq_lens + self.speculative_step_id + 1）写入该切片。
3. 移除旧的 seq_lens_cpu 相关行，改用 seq_lens.max().item() 计算 max_seq_len_k，因为 seq_lens_cpu 在此处不再需要。
4. 移除了 max_len 局部变量，直接计算 max_seq_len_k 和 max_seq_pages。

关键文件：

- python/sglang/srt/layers/attention/trtllm_mha_backend.py（模块 注意力；类别 source；类型 core-logic；符号 init_forward_metadata_replay_cuda_graph）：核心注意力后端，修复 CUDA graph 重放时 cache_seqlens_int32 未绑定当前 batch 的 bug。

关键符号：init_forward_metadata_replay_cuda_graph

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险：变更仅影响 TRTLLM MHA 后端的 CUDA graph 重放路径，且逻辑简化了数据流。未涉及其他注意力后端或模型。
- 影响：影响范围限于使用 TRTLLM MHA 后端且启用 speculative decoding 的场景。修复可能影响 Blackwell 等 GPU 上的推理正确性。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR