PR #21987 完整报告

sgl-project/sglang

[Bugfix] Fix CUDA graph replay issues in trtllm_mla draft_extend

合并时间：2026-04-03 16:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21987

执行摘要

- 一句话：修复 TRTLLM MLA 后端中 CUDA 图回放路径的步长不一致问题，确保推测解码正确性。
- 推荐动作：该 PR 值得精读，特别是对关注 CUDA 图优化、推测解码或 TRTLLM MLA 后端的工程师。关注设计决策：如何通过统一步长解决布局不匹配问题，以及移除冗余路径的权衡。

功能与动机

修复测试失败（具体见 PR body 中链接的 CI 运行），根因是 cu_seqlens_q 张量布局在 CUDA 图捕获和回放路径中不匹配。这导致 EAGLE 推测解码在启用 CUDA 图时出现 NaN 断言错误，影响模型推理的可靠性。PR body 明确引用类似修复 PR #19807，旨在解决相同问题但针对不同后端。

实现拆解

主要修改集中在 python/sglang/srt/layers/attention/trtllm_mla_backend.py 文件的 init_forward_metadata_replay_cuda_graph 函数中：

1. 将 cu_seqlens_q 从基于 accept_length 的可变步长计算改为使用 torch.arange 的均匀步长，步长为 num_draft_tokens（从 server_args.speculative_num_draft_tokens 获取），匹配捕获路径。
2. 更新 max_seq_len_q 和 sum_seq_lens_q 为固定值，而非基于 accept_length 的动态值。
3. 移除 accept_length_cpu 快速路径，因为均匀步长使其不必要。 次要修改在 .github/workflows/nightly-test-nvidia.yml：添加新的 CI 作业 nightly-test-specialized-8-gpu-b200 并调整作业过滤条件，以扩展测试覆盖。

关键文件：

- python/sglang/srt/layers/attention/trtllm_mla_backend.py（模块 attention）：核心修复文件，修改了 init_forward_metadata_replay_cuda_graph 函数，统一 CUDA 图回放路径的步长布局，直接影响推测解码的 CUDA 图功能。
- .github/workflows/nightly-test-nvidia.yml（模块 infra）：次要修改，添加新的 CI 作业并调整过滤条件，扩展测试覆盖，支持修复验证。

关键符号：init_forward_metadata_replay_cuda_graph

评论区精华

review 讨论聚焦于修改是否对推测解码的 v1 和 v2 版本都正确。Qiaolin-Yu 提问："I think it makes sense for spec v2. But is this correct for spec v1?" kpham-sgl 回应并引用相关代码块，说明捕获路径代码适用于两者，且在前向扩展中查询张量被填充，从而确认修改安全。讨论结论是修改正确，并得到批准，未解决疑虑。

- 修改是否对 spec v1 和 v2 都正确 (correctness): 讨论得出结论修改正确，基于代码引用确认安全，未留下未解决疑虑。

风险与影响

- 风险：技术风险较低：
  - 核心逻辑变更在 trtllm_mla_backend.py 的 init_forward_metadata_replay_cuda_graph 函数中，仅影响 CUDA 图回放路径，不改变前向计算本身，但需确保与捕获路径严格匹配，否则可能引入新回放错误。
  - 移除 accept_length_cpu 路径可能影响性能，但基于 PR body 描述，均匀步长使其不必要，兼容性风险小。
  - 依赖 self.num_draft_tokens 的正确性，需确保 server_args.speculative_num_draft_tokens 在所有场景下可用。
  - CI 变更（添加新作业）可能影响测试流程，但仅扩展覆盖，无直接风险。
- 影响：影响范围有限但重要：
- 用户影响：修复后，使用 TRTLLM MLA 后端进行 EAGLE 推测解码且启用 CUDA 图的用户将避免测试失败和潜在推理错误，提升稳定性和正确性。
- 系统影响：确保 CUDA 图回放功能在推测解码场景下正常工作，有助于性能优化和一致性。
- 团队影响：通过关联 PR #19807 的类似修复模式，增强了代码库中对 CUDA 图布局一致性的理解，可能促进其他后端的类似修复。
- 风险标记：核心路径变更，依赖外部参数正确性

关联脉络

- PR #19807 Fix issue 19717 by making qo_indptr uniform strided instead of packed: 关联 PR，修复了类似问题（CUDA 图布局不一致），但针对不同后端或组件，PR body 中明确引用为相同根因和修复模式。