PR #21649 完整报告

sgl-project/sglang

fix: TRT-LLM MHA CUDA illegal address with EAGLE v2 + DP attention

合并时间：2026-04-06 00:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21649

执行摘要

- 一句话：修复 TRT-LLM MHA 在 EAGLE v2 推测解码 +DP 注意力下因批次大小不一致导致的 CUDA 非法地址错误。
- 推荐动作：该 PR 值得精读，尤其关注：1) DP 注意力下批次大小不一致的根本原因分析；2) 从 forward_batch.batch_size 到元数据推导的设计决策，体现了与其他后端行为对齐的架构一致性；3) review 中关于填充目的和注意力独立性的讨论，有助于理解分布式推理中的数据流设计。

功能与动机

根据 PR body 描述，在 Qwen3.5-397B 模型上运行 MMMU-Pro VLM 评估时，启用 DP 注意力、EAGLE v2 推测解码和 TRT-LLM MHA 后端后，约 388-405/500 个问题处会一致触发 CUDA 非法地址错误。根本原因是 prepare_mlp_sync_batch 为 MLP 同步将 forward_batch.batch_size 填充至 DP 组最大批次大小，但 init_forward_metadata 已基于原始批次大小计算元数据张量，导致 TRT-LLM FMHA 内核访问越界。

实现拆解

核心改动在 python/sglang/srt/layers/attention/trtllm_mha_backend.py 的 forward_extend 函数中，将 batch_size 参数从 forward_batch.batch_size 改为 self.forward_metadata.cu_seqlens_q.shape[0] - 1，从而从元数据张量形状推导真实批次大小，与其他注意力后端（FlashInfer、Triton）行为保持一致。提交历史显示最初尝试在 TRTLLMMHAMetadata 中存储 batch_size 字段，但最终采用更简洁的推导方案。

关键文件：

- python/sglang/srt/layers/attention/trtllm_mha_backend.py（模块 attention_backend）：唯一修改文件，包含 forward_extend 函数的关键修复，将 batch_size 参数从 forward_batch.batch_size 改为从 cu_seqlens_q 推导

关键符号：forward_extend, init_forward_metadata, prepare_mlp_sync_batch

评论区精华

review 中主要讨论了修复方案的正确性：Qiaolin-Yu 质疑为何选择元数据而非 forward_batch 的填充状态；ispobock 澄清填充仅用于 MLP 通信，注意力应基于真实批次大小独立计算；gemini-code-assist[bot] 指出初始方案中 CUDA 图捕获路径缺失 batch_size 设

置，可能导致相同错误。最终结论是采用元数据推导方案，既解决非法地址问题，又避免维护额外字段。

- 修复方案正确性：为何使用元数据而非 forward_batch 的填充批次大小 (correctness): 采用从 cu_seqlens_q 推导真实批次大小的方案，确保注意力计算与元数据边界一致
- 初始方案中 CUDA 图捕获路径的完整性 (correctness): 最终方案放弃存储 batch_size 字段，直接推导，避免该问题

风险与影响

- 风险：风险较低：1) 变更仅影响 TRT-LLM MHA 后端在 DP 注意力 + 推测解码场景，其他后端或配置不受影响；2) 从 cu_seqlens_q 推导批次大小与其他后端逻辑一致，降低不一致风险；3) 已通过 MMMU-Pro 完整评估验证（1730/1730 问题无崩溃）。潜在风险：若 cu_seqlens_q 张量形状异常，可能推导错误批次大小，但该张量由同一初始化逻辑生成，风险可控。
- 影响：影响范围：1) 用户：修复了多模态评估中的稳定性问题，确保启用 DP 注意力、EAGLE v2 和 TRT-LLM MHA 后端的场景可稳定运行；2) 系统：消除 CUDA 非法地址崩溃，提升系统可靠性；3) 团队：明确了 DP 填充仅用于 MLP 通信、注意力应基于真实批次大小的设计原则，为类似问题提供参考。影响程度中等，针对特定配置的崩溃修复。
- 风险标记：核心路径变更，分布式推理边界条件

关联脉络

- PR #22146 Isolate spec V1 path in decode post-processing: 同涉及推测解码（speculative decoding）路径的修改，本 PR 修复 EAGLE v2 下 TRT-LLM MHA 问题，22146 隔离 Spec V1 后处理路径，显示推测解码模块的持续演进
- PR #22148 Unify think_end_id to model_config as single source of truth: 同属一致性（consistency）改进，本 PR 统一批次大小推导逻辑与其他后端一致，22148 统一 think_end_id 存储，体现代码库消除冗余、提升一致性的趋势
- PR #22104 [SpecV2]: Reopen kl accuracy test for qwen3 + SpecV2: 同涉及推测解码测试，本 PR 修复实际运行问题，22104 重新启用 SpecV2 测试，反映推测解码功能的测试与修复并行推进