PR #32254 完整报告

sgl-project/sglang

[spec decoding] fix inkling multi layer mtp draft extend cuda graph

合并时间：2026-07-24 05:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32254

执行摘要

- 一句话：修复 Inkling MTP draft extend 的 CUDA Graph 宽度不匹配
- 推荐动作：建议快速合并。对于涉及 CUDA Graph 的类似场景，应建立一致性检查（如在输入构建时断言维度匹配）。

功能与动机

当启用 boundary-KV 窗口（widened window）时，draft extend 阶段的 CUDA Graph runner 捕获的输入宽度包含了前端宽裕 token（draft_extend_num_front_tokens），但 EagleDraftExtendInput 初始化时未计入这额外 token 数，导致图执行时维度不匹配，可能引发错误或崩溃。

实现拆解

1. 定位问题：在 python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py 的 _draft_extend_for_decode 方法中，EagleDraftExtendInput 的 num_tokens_per_req 参数原本为 self.speculative_num_steps + 1，不包含 draft_extend_num_front_tokens。
2. 修正计算：将 num_tokens_per_req 改为 self.speculative_num_steps + 1 + self.draft_extend_num_front_tokens，与 CUDA Graph runner 捕获的宽度保持一致。
3. 更新注释：添加注释说明额外 token 用于 boundary-KV 前端宽裕行，并强调必须与 graph runner 捕获宽度匹配。
4. 影响范围：仅修改一个文件中的一个表达式，不涉及接口或配置变更。

关键文件：

- python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py（模块 推测解码；类别 source；类型 core-logic；符号 _draft_extend_for_decode）：核心修复文件，修正 EagleDraftExtendInput 的 num_tokens_per_req 计算，确保与 CUDA Graph 捕获宽度匹配。

关键符号：_draft_extend_for_decode

评论区精华

无 review 讨论。作者在 Issue 评论区使用 /rerun-test 触发了 test_step3p5_flash_chain_mtp.py 测试，结果通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅修正一个数值计算，且与已有的 draft_extend_num_front_tokens 变量取值一致，无兼容性问题。由于缺乏新增单元测试（仅依赖现有 MTP E2E 测试），若未来 draft_extend_num_front_tokens 计算逻辑变更需同步更新此处。
- 影响：仅影响 Inkling 多模型 MTP 推理场景下启用 boundary-KV 窗口的 draft extend 阶段。修复后 CUDA Graph 能正确执行，预计提升稳定性和准确性。对非 boundary-KV 模式无影响（draft_extend_num_front_tokens 为 0）。
- 风险标记：缺少新增测试覆盖

关联脉络

- PR #27657 [DeepSeek V4] CP decode opt: slice repeat attention weights to local TP partition: 同样涉及推测解码和多模型 MTP 推理优化，共享相似代码路径。