PR #22404 完整报告

sgl-project/sglang

cuda graph: adjust capture time num-non-padded-tokens to align capture with replay

合并时间：2026-04-11 10:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22404

执行摘要

- 一句话：修复 CUDA Graph 捕获时 num_token_non_padded 计算逻辑，确保捕获与重放行为一致。
- 推荐动作：建议 CUDA Graph 和 attention TP 相关开发者精读此 PR，理解捕获与重放路径对齐的设计决策。关注条件判断逻辑和 compute_local_num_token_non_padded 函数的实现，确保在不同配置下行为正确。

功能与动机

根据 PR body 中的描述，修复动机是 "Correctly compute the num-non-padded-tokens during graph capture, which align with the behavior in replay"，并引用了代码行 https://github.com/sgl-project/sglang/blob/1b7c33a5b751dac6187367d798a7b80bd12ccaaf/python/sglang/srt/model_executor/cuda_graph_runner.py#L326。这表明需要确保 CUDA Graph 捕获时的计算逻辑与重放阶段保持一致，避免因不一致导致的问题。

实现拆解

该 PR 仅修改了一个文件 python/sglang/srt/model_executor/cuda_graph_runner.py。在 capture_one_batch_size 函数中，原本只是简单地将 buffers.num_token_non_padded[...] 设置为 num_tokens。现在增加了条件判断逻辑：当启用 enable_num_token_non_padded 且需要 gathered buffer 且未启用 NSA 预填充检查点时，调用 compute_local_num_token_non_padded 函数计算本地 token 数量，并复制到 buffers.num_token_non_padded 中。这确保了捕获路径与重放路径中的 populate_from_forward_batch 函数行为一致。

关键文件：

- python/sglang/srt/model_executor/cuda_graph_runner.py（模块 model_executor）：这是唯一修改的文件，包含了 CUDA Graph 捕获的核心逻辑修复，直接影响图执行的一致性。

关键符号：capture_one_batch_size, compute_local_num_token_non_padded

评论区精华

由于 review 评论为空，仅有的交互是 ispobock 的批准和触发 CI 的评论。从上下文看，这个修复相对直接，没有引发深入的技术讨论。ispobock 作为合并者，通过 /tag-and-rerun-ci 触发了 CI 测试，表明需要验证修复的正确性。

- CUDA Graph 捕获与重放一致性修复 (correctness): 通过添加条件判断和调用 compute_local_num_token_non_padded 函数来修复。

风险与影响

- 风险：风险较低但需关注：1. 核心路径变更：修改了 CUDA Graph 捕获的关键逻辑，如果条件判断有误或 compute_local_num_token_non_padded 函数行为不符合预期，可能导致图捕获失败或重放时行为异常。2. 条件复杂性：新增的条件涉及多个标志（enable_num_token_non_padded、require_gathered_buffer、nsa_enable_prefill_cp），需要确保这些标志在捕获和重放时状态一致。3. 缺少测试覆盖：从 PR body 看，作者未提供准确性测试或性能测试结果，依赖 CI 测试验证。
- 影响：影响范围有限但重要：1. 对系统：修复 CUDA Graph 执行的一致性问题，确保捕获的图能正确重放，提升推理稳定性。2. 对用户：透明修复，不会改变 API 或可见行为，但能避免潜在的图执行错误。3. 对团队：涉及 CUDA Graph 和 attention TP 的交互，需要相关开发者关注此修复的逻辑。
- 风险标记：核心路径变更，条件复杂性，缺少测试覆盖

关联脉络

- PR #21104 perf: precompute FA3 scheduler_metadata to eliminate per-layer prepare_varlen_num_blocks: 同样涉及 CUDA Graph 和性能优化，可能共享类似的图捕获逻辑。
- PR #22051 [MUSA][9/N] Add FA3 attention backend support through MATE (MUSA AI Tensor Engine): 涉及 attention 后端支持，与本 PR 的 attention TP 调整相关。