PR #24617 完整报告

sgl-project/sglang

fix(fa3): translate page table to SWA loc in EAGLE3 topk>1 spec metadata

合并时间：2026-05-09 18:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24617

执行摘要

- 一句话：修复 FA3+EAGLE3 topk>1 时 SWA 页表地址翻译
- 推荐动作：建议精读。本 PR 展示了在注意力后端中处理多级 KV pool 地址空间时的常见陷阱，值得关注 translate_loc_from_full_to_swa 的作用和调用情境。改动简洁明了，易于理解，适合作为 backend 开发参考。建议后续添加对应的单元测试或集成测试来预防回归。

功能与动机

关联 Issue #24402 报告了使用 EAGLE3 tree decoding 时出现 CUDA error: an illegal memory access 的崩溃，发生在 GPT-OSS 等 hybrid SWA 模型上。PR body 指出：prepare_swa_spec_page_table_triton() 传入的 prefix 和 draft-token 页表来自 req_to_token_pool（full KV pool），而 SWA 层直接使用结果 swa_spec_metadata.page_table 索引 swa_kv_pool，导致非法访问。

实现拆解

1. 页表变量抽取：在 _init_sliding_window_attn_spec_metadata 中，将原本直接传给 Triton kernel 的 metadata.page_table 和 metadata_expand.page_table 分别赋值给局部变量 page_table_a 和 page_table_b。
2. 地址翻译条件注入：检查 self.use_sliding_window_kv_pool 标志，若为真，则对 page_table_a 和 page_table_b 调用 self.token_to_kv_pool.translate_loc_from_full_to_swa()，将 full KV pool 的位置索引转换为 SWA pool 的索引。
3. 传入翻译后的页表：将翻译后的 page_table_a 和 page_table_b 传递给 prepare_swa_spec_page_table_triton() kernel，确保生成的 swa_spec_metadata.page_table 中的索引与 SWA pool 正确对应。
4. 无其他文件改动：只修改了 flashattention_backend.py 中的一处方法，改动量小（+12/-2），未涉及测试或配置变更。

关键文件：

- python/sglang/srt/layers/attention/flashattention_backend.py（模块 注意力层；类别 source；类型 core-logic；符号 _init_sliding_window_attn_spec_metadata）：这是唯一的修改文件，在 _init_sliding_window_attn_spec_metadata 方法中增加了页表地址从 full KV pool 到 SWA pool 的翻译逻辑，修复了崩溃。

关键符号：_init_sliding_window_attn_spec_metadata

关键源码片段

python/sglang/srt/layers/attention/flashattention_backend.py

这是唯一的修改文件，在 _init_sliding_window_attn_spec_metadata 方法中增加了页表地址从 full KV pool 到 SWA pool 的翻译逻辑，修复了崩溃。

```
# file: python/sglang/srt/layers/attention/flashattention_backend.py
# context: _init_sliding_window_attn_spec_metadata 方法

def _init_sliding_window_attn_spec_metadata(
    self,
    metadata: FlashAttentionMetadata,
    metadata_expand: FlashAttentionMetadata,
    metadata_swa: Optional[FlashAttentionMetadata] = None,
):
    # ... 前面的 cache_seqlens_int32 和 cu_seqlens_k 计算不变 ...

    page_table = (
        metadata.page_table.new_zeros(
            (bs, metadata.max_seq_len_k + metadata_expand.page_table.shape[1])
        )
        if metadata_swa is None
        else metadata_swa.page_table
    )

    # 新增：分别保存 prefix 和 draft 的页表
    page_table_a = metadata.page_table
    page_table_b = metadata_expand.page_table

    # 新增：当使用 sliding window KV pool 时，
    # 将 full pool 的 page table 索引翻译为 SWA pool 的索引
    if self.use_sliding_window_kv_pool:
        page_table_a = self.token_to_kv_pool.translate_loc_from_full_to_swa(
            page_table_a
        )
        page_table_b = self.token_to_kv_pool.translate_loc_from_full_to_swa(
            page_table_b
        )

    # 传入翻译后的页表，确保生成的 swa_spec_metadata.page_table
    # 中的索引与 SWA pool 对齐，避免越界
    prepare_swa_spec_page_table_triton(
        page_table,
        page_table_a,
        page_table_b,
        metadata.cache_seqlens_int32,
        metadata_expand.cache_seqlens_int32,
        self.speculative_num_draft_tokens,
    )
```

```
    # ... 后续 metadata_swa 的赋值不变 ...
```

评论区精华

Review 过程简单：gemini-code-assist[bot] 自动生成评论确认变更意图，ispobock 直接批准。无实质讨论或争议。

- 修复意图确认 (correctness): 无争议，ispobock 直接批准。

风险与影响

- 风险：本 PR 仅修改了 _init_sliding_window_attn_spec_metadata 中页表地址翻译逻辑，改动集中在条件分支内，默认行为（不启用 SWA pool）无变化。风险较低，但需注意：
  1. 依赖 translate_loc_from_full_to_swa 方法的正确性和性能，若该方法存在 bug 则可能引入新问题。
  2. 未添加测试覆盖，回归风险依赖于后续测试（如 Issue #24402 中的复现命令）。
  3. 该修复仅针对 FA3 backend，其他 attention backend（如 Triton 或 FA2）若存在类似问题仍需单独修复。 - 影响：影响范围：修复了 FlashAttention 后端在使用 SWA pool + EAGLE3 + topk>1 时的崩溃，对 hybrid SWA 模型（如 GPT-OSS）至关重要。影响程度：高，因为该 bug 导致模型完全不可用（CUDA illegal memory access）。修复后模型可正常运行并达到预期的准确率和性能（PR body 提供了 GSM8K 分数 0.690 和吞吐量 121 token/s 的数据）。用户：使用 FA3 + EAGLE3 且模型使用了 SWA layer 的用户将受益。系统：无负面影响。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #24402 Issue: EAGLE3 Tree Decoding CUDA error: an illegal memory access was encountered: 本 PR 直接修复该 Issue 报告的问题。
- PR #24097 Restrict fa_skip_kv_cache to non-MLA backends: 同文件修改，涉及 FlashAttention backend 的 SWA 相关修复。