PR #26412 完整报告

sgl-project/sglang

[Bug] Forward fixed_split_size in SWA / cross-attention paths of FlashInfer backend

合并时间：2026-05-28 15:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26412

执行摘要

- 一句话：修复 FlashInfer SWA/ 交叉注意力路径遗漏 deterministic 参数
- 推荐动作：建议合并。该 PR 修复了一个隐蔽的确定性保证缺陷，改动量小，风险低，且作者已通过内部 SWA 模型验证了修复效果。

功能与动机

--enable-deterministic-inference 依赖 FlashInfer 的 prefill_split_tile_size 参数固定 split-K 规约顺序，但该参数在 SWA 和交叉注意力路径中被遗漏，导致注意力输出随批大小或组成变化而漂移。PR body 明确描述：『The bug only surfaces on SWA / cross-attention models』，并给出测试数据——修复前 22/43 logp 不匹配，修复后全部一致。

实现拆解

1. 修改 update_sliding_window（Decode）：在 call_begin_forward() 调用中增加 fixed_split_size=fixed_split_size 和 disable_split_kv=disable_split_kv 参数。
2. 修改 update_cross_attention（Decode）：同样在 call_begin_forward() 调用中补传这两个参数。
3. 修改 update_sliding_window（Prefill）：增加 fixed_split_size=fixed_split_size 参数（Prefill 侧无 disable_split_kv）。
4. 修改 update_cross_attention（Prefill）：增加 fixed_split_size=fixed_split_size 参数。所有修改均发生在 python/sglang/srt/layers/attention/flashinfer_backend.py 文件中，共 6 行新增，无删除。

关键文件：

- python/sglang/srt/layers/attention/flashinfer_backend.py（模块 注意力后端；类别 source；类型 core-logic；符号 FlashInferIndicesUpdaterDecode.update_sliding_window, FlashInferIndicesUpdaterDecode.update_cross_attention, FlashInferIndicesUpdaterPrefill.update_sliding_window, FlashInferIndicesUpdaterPrefill.update_cross_attention）：唯一的修改文件，包含 FlashInfer 注意力后端的核心逻辑，FlashInferIndicesUpdater{Decode,Prefill} 类的四个叶子方法补传了 deterministic 参数。

关键符号：FlashInferIndicesUpdaterDecode.update_sliding_window, FlashInferIndicesUpdaterDecode.update_cross_attention,

FlashInferIndicesUpdaterPrefill.update_sliding_window,FlashInferIndicesUpdaterPrefill.update_cross_attention

关键源码片段

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

唯一的修改文件，包含 FlashInfer 注意力后端的核心逻辑，FlashInferIndicesUpdater{Decode,Prefill} 类的四个叶子方法补传了 deterministic 参数。

```
# flashinfer_backend.py 中 Decode 侧的 update_sliding_window 和 update_cross_attention 补传参数

# 原代码在 call_begin_forward 调用中缺少 fixed_split_size / disable_split_kv
# 修复后传递这两个参数
self.call_begin_forward(
    decode_wrappers[wrapper_id],
    req_pool_indices,
    paged_kernel_lens_tmp,
    paged_kernel_lens_sum_tmp,
    self.kv_indptr[wrapper_id],
    kv_start_idx_tmp,
    spec_info,
    seq_lens_cpu=seq_lens_cpu_tmp,
    use_sliding_window_kv_pool=use_sliding_window_kv_pool,
    fixed_split_size=fixed_split_size, # 新增：固定 split-K 规约大小
    disable_split_kv=disable_split_kv, # 新增：禁止 split-KV
)

# 类似地，Prefill 侧的 update_sliding_window 和 update_cross_attention 也补传 fixed_split_size
self.call_begin_forward(
    ...
    fixed_split_size=fixed_split_size, # 新增
)
```

评论区精华

无 review 评论或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅传递一个原本就已声明的 Optional[int] 参数，当 --enable-deterministic-inference 未启用时值为 None，行为完全不变。需确保所有调用 call_begin_forward() 的路径都已传递参数，目前 update_single_wrapper 路径原已正确传递，本次补全了其余四条路径。
- 影响：影响范围：使用 FlashInfer 后端并启用 --enable-deterministic-inference 且模型使用滑动窗口注意力或交叉注意力的用户。此前这些场景的确定性保证是失效的，修复后可获

得正确的批间一致输出。对无 SWA/ 交叉注意力的模型（如 Qwen3-8B）无影响。

- 风险标记：暂无

关联脉络

- 暂无明显关联 PR