PR #23975 完整报告

sgl-project/sglang

Fix LFM2 ShortConv Mamba State Indexing

合并时间：2026-05-01 06:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23975

执行摘要

- 一句话：修复 LFM2 short-conv Mamba 状态索引错误
- 推荐动作：值得精读：1）展示了混合索引命名空间错误的经典模式及修复方法；2）PR body 提供了高质量的根因分析和验证数据，是优秀 bugfix 范本；3）代码修改量小（12+6-），但影响正确性显著，适合作为新人学习索引映射的案例。

功能与动机

LFM2 使用混合缓存：请求 / 令牌池索引标识请求行，Mamba/ 短卷积状态索引标识持久卷积状态行，两者是独立的命名空间。ShortConv 路径直接使用 *req_pool_indices* 作为卷积状态索引，导致读取 / 写入错误的 Mamba 状态槽位（槽 0 被保留为填充哑元槽位，真实 Mamba 槽从 1 开始分配），CUDA graph 重放时状态复用使错误立即影响生成 token。PR body 提供了详细的首 token 对比和 IFEval 评测数据证明修复必要性。

实现拆解

1. 在 lfm2.py 和 lfm2_moe.py 的 Lfm2ShortConv.forward() 和 Lfm2MoeShortConv.forward() 中，从 forward_batch.req_to_token_pool 通过 *get_mamba_indices*() 获取正确的 Mamba 状态索引。
2. 将 *decode* 分支的 *conv_state_indices* 参数从 *req_pool_indices*.to(torch.int32) 替换为 mamba_indices.to(torch.int32)。
3. 将 prefill 分支的 *cache_indices* 参数（包括多序列和单序列情况）从 *req_pool_indices* 相关切片替换为 mamba_indices 相关切片。
4. 其余逻辑（输入投影、门控计算、输出投影）保持不变，改动集中于索引映射层，影响面极小。

关键文件：

- python/sglang/srt/models/lfm2.py（模块 模型层；类别 source；类型 data-contract；符号 Lfm2ShortConv.forward）：LFM2 模型的 ShortConv 前向逻辑所在文件，包含核心索引映射修复。
- python/sglang/srt/models/lfm2_moe.py（模块 模型层；类别 source；类型 data-contract；符号 Lfm2MoeShortConv.forward）：LFM2-MoE 模型的 ShortConv 前向逻辑所在文件，与 lfm2.py 完全对称的修复。

关键符号：Lfm2ShortConv.forward, Lfm2MoeShortConv.forward

关键源码片段

python/sglang/srt/models/lfm2.py

LFM2 模型的 ShortConv 前向逻辑所在文件，包含核心索引映射修复。

```
def forward(
    self,
    hidden_states: torch.Tensor,
    forward_batch: ForwardBatch,
) -> torch.Tensor:
    if forward_batch.forward_mode.is_idle():
        return hidden_states

    layer_cache = forward_batch.req_to_token_pool.mamba2_layer_cache(self.layer_idx)
    conv_state = layer_cache.conv[0]
    req_pool_indices = forward_batch.req_pool_indices
    # 通过统一 API 将 request pool 索引映射到 Mamba 状态池索引
    mamba_indices = forward_batch.req_to_token_pool.get_mamba_indices(
        req_pool_indices
    )

    # 门控投影
    proj, _ = self.in_proj(hidden_states)
    B_gate, C_gate, x = proj.chunk(3, dim=-1)
    Bx = B_gate * x

    if forward_batch.forward_mode.is_decode():
        # decode：使用正确映射的 Mamba 索引
        conv_out = causal_conv1d_update(
            Bx,
            conv_state,
            self.conv_weight,
            self.conv_bias,
            activation=None,
            conv_state_indices=mamba_indices.to(torch.int32),
        )
    else:
        # prefill：varlen 卷积，同样使用 Mamba 索引
        T = hidden_states.shape[0]
        Bx_t = Bx.transpose(0, 1).contiguous()
        extend_start_loc = forward_batch.extend_start_loc
        if extend_start_loc is not None and len(extend_start_loc) > 1:
            query_start_loc = torch.cat([extend_start_loc,
                torch.tensor([T], dtype=torch.int32, device=hidden_states.device)])
            cache_indices = mamba_indices.to(torch.int32)
        else:
            query_start_loc = torch.tensor([0, T], dtype=torch.int32, device=hidden_states.device)
```

```
            cache_indices = mamba_indices[:1].to(torch.int32)

        conv_out = causal_conv1d_fn(
            Bx_t, self.conv_weight, self.conv_bias,
            query_start_loc=query_start_loc,
            cache_indices=cache_indices,
            has_initial_state=None,
            conv_states=conv_state,
            activation=None,
        ).transpose(0, 1)

    output, _ = self.out_proj(C_gate * conv_out)
    return output
```

评论区精华

本 PR 无 review 评论讨论。PR body 详细说明了根因、验证方法和 IFEval 评测结果。HaiShaw 在 CI 机器人评论中确认该 PR 为「小范围精准修复」，AMD CI 的 3 个失败均不相关。

- CI 失败不相关 (other): 确认 CI 失败与 PR 无关。

风险与影响

- 风险：风险极低：改动仅交换索引来源（req_pool_indices → mamba_indices），且该映射（get_mamba_indices）在其他 Mamba 路径中已被广泛使用。未引入新的控制流或数据依赖。但缺少自动化回归测试覆盖该场景，长期依赖人工验证。
- 影响：影响范围精确集中在 LFM2 和 LFM2-MoE 模型的 ShortConv 层。修复后，CUDA graph 启用的推理首 token 和生成质量与 transformers 参考实现一致，IFEval 指标达到模型卡预期。对其他模型或硬件后端无影响。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR