PR #25030 完整报告

sgl-project/sglang

[Spec] Multi-layer mamba scatter cleanup; fix positional call bug

合并时间：2026-05-12 13:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25030

执行摘要

- 一句话：修复 MultiLayerEagleWorker mamba 状态更新 bug 并清理代码
- 推荐动作：建议尽快合并，并检查其他类似位置是否存在参数位置隐患。该 PR 展示了如何通过重构和对齐代码消除隐蔽 bug，值得参考。

功能与动机

修复 MultiLayerEagleWorker 在 hybrid_gdn_config 分支下调用 update_mamba_state_after_mtp_verify 时因参数错位引发的 TypeError，同时清理代码以与 EAGLEWorker 系列保持一致，提升可维护性和正确性。

实现拆解

1. 移除冗余别名：删除 num_accept_tokens = num_correct_drafts + 1，直接使用 num_correct_drafts 并在 cumsum 中内联 + 1。
2. 优化索引计算：将 last_token_indices_per_req - first_token_indices_per_req 替换为 accept_indices[cum - 1] - accepted_indices_offset，消去了一次 cat 和一次 index_select，利用 first_token_indices_per_req[i] == i * draft_token_num 的不变性提高效率。
3. 简化 else 分支：直接返回 num_correct_drafts 而非 num_accept_tokens - 1。
4. 修复参数位置 bug：将 update_mamba_state_after_mtp_verify 调用改为关键字参数形式，显式传递 mamba_track_indices=None, mamba_steps_to_track=None，确保参数对应正确。

关键文件：

- python/sglang/srt/speculative/multi_layer_eagle_worker.py（模块 投机解码；类别 source；类型 core-logic；符号 verify）：唯一修改的文件，包含 mamba 状态更新逻辑的对齐和参数 bug 修复

关键符号：verify

关键源码片段

python/sglang/srt/speculative/multi_layer_eagle_worker.py

唯一修改的文件，包含 mamba 状态更新逻辑的对齐和参数 bug 修复

```
def verify(self, batch: ScheduleBatch):
    # ... 前面的代码省略 ...
    if self.target_worker.model_runner.hybrid_gdn_config is not None:
        # 直接使用 num_correct_drafts，移除 num_accept_tokens 别名
        num_correct_drafts = torch.tensor(
            res.num_correct_drafts_per_req_cpu,
            device=logits_output.hidden_states.device,
            dtype=torch.int64,
        )

        if spec_info.topk > 1 and res.accept_indices.shape[0] > 0:
            # 用 accepted_indices_offset 替代 first_token_indices_per_req
            cumulative_num_accept_tokens = torch.cumsum(
                num_correct_drafts + 1, dim=0
            )
            accepted_indices_offset = torch.arange(
                0,
                len(batch.seq_lens) * self.speculative_num_draft_tokens,
                step=self.speculative_num_draft_tokens,
                dtype=num_correct_drafts.dtype,
                device=num_correct_drafts.device,
            )
            # 直接计算，消去 cat 和 index_select
            last_correct_step_indices = (
                res.accept_indices[cumulative_num_accept_tokens - 1]
                - accepted_indices_offset
            )
        else:
            last_correct_step_indices = num_correct_drafts

        # 修复：使用关键字参数确保与函数签名一致
        self.target_worker.model_runner.attn_backend.update_mamba_state_after_mtp_verify(
            last_correct_step_indices=last_correct_step_indices,
            mamba_track_indices=None,
            mamba_steps_to_track=None,
            model=self.target_worker.model_runner.model,
        )
```

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更集中在 MultiLayerEagleWorker 的单个路径（hybrid_gdn_config 分支），且逻辑与已验证的 EAGLEWorker 系列对齐。需确保测试覆盖该分支，尤其是 topk > 1 的路径，避免回归。

- 影响：仅影响使用 MultiLayerEagleWorker 且开启 hybrid_gdn_config 的场景（如 Mamba 与注意力混合模型）。修复后该路径能正确运行，不再抛出 TypeError。对其他用户无影响。
- 风险标记：低风险，对齐已验证逻辑

关联脉络

- PR #25029 [Spec] Mamba scatter cleanup; fix multi-layer positional bug; dflash naming: 本 PR 是对 #25029 后续清理的延续。
- PR #25038 [Spec] Rename accepted_indices -> accept_indices; drop _token_id suffix per Rule 5: 同一命名规范系列 PR，但本 PR 重命名了多文件中的字段。