PR #26351 完整报告

sgl-project/sglang

[bugfix] commit Mamba states after NGRAM target verify

合并时间：2026-06-12 06:11

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26351

执行摘要

- 一句话：修复 NGRAM 推测解码下 Mamba 状态未提交导致的输出重复
- 推荐动作：建议精读 _mamba_verify_update 的实现，理解 Mamba 状态提交机制和 track 点更新逻辑，这对于维护 hybrid 推测解码非常重要。此修复填补了 NGRAM 路径与 eagle_worker_v2 路径之间的行为差距。

功能与动机

修复 hybrid GDN 模型（如 Qwen3.5）在 NGRAM 推测解码下的输出损坏 / 重复 bug。PR body 描述用户可见问题：响应中出现 token 循环、重复片段和格式损坏，例如输出重复的 'I am Qwen' 模式。根本原因是 target verify 后接受的每请求推测状态未提交回持久 Mamba 缓存。

实现拆解

1. 导入依赖调整：在 python/sglang/srt/speculative/ngram_worker.py 中新增 set_mamba_track_indices_from_reqs 和 get_global_server_args 导入，为后续启用 Mamba 状态跟踪做准备。
2. verify 前刷新 Mamba track indices：在 _prepare_for_speculative_decoding 方法中，当 enable_mamba_extra_buffer() 为真时，调用 set_mamba_track_indices_from_reqs(batch) 重建 track indices，并清除 mamba_track_mask 和 mamba_track_seqlens，防止过期的 extend 阶段 mask 在 TARGET_VERIFY 期间触发错误的跟踪。
3. 新增 _mamba_verify_update 方法：该方法在 forward_batch_generation 中的 verify(...) 调用之后立即执行。它计算每个请求最后一个正确步骤的索引（last_correct_step_indices），然后调用 attn_backend.update_mamba_state_after_mtp_verify(...) 将接受的 Mamba 状态提交到持久缓存。同时，如果启用了 mamba track 点更新，还根据 verify 前后的序列长度变化，计算是否需要更新 track 点并调用 set_mamba_track_indices_from_reqs。
4. 集成到 forward_batch_generation：在 target verify 分支中，verify_input.verify(...) 返回 accept_lens 和 accept_index 后，直接调用 self._mamba_verify_update(batch, accept_lens, accept_index, bs)。

关键文件：

- python/sglang/srt/speculative/ngram_worker.py（模块 推测解码；类别 source；类型 core-logic；符号 _mamba_verify_update）：唯一修改的文件，新增 _mamba_verify_update 方法和相关导入与前置刷新逻辑，是修复的核心。

关键符号：_mamba_verify_update

关键源码片段

python/sglang/srt/speculative/ngram_worker.py

唯一修改的文件，新增 _mamba_verify_update 方法和相关导入与前置刷新逻辑，是修复的核心。

```
# python/sglang/srt/speculative/ngram_worker.py
# 在 forward_batch_generation 的 target verify 分支中调用

# 新增导入
from sglang.srt.managers.schedule_batch import (
    ScheduleBatch,
    set_mamba_track_indices_from_reqs, # 用于刷新 track indices
)
from sglang.srt.server_args import ServerArgs, get_global_server_args

# 在 _prepare_for_speculative_decoding 中添加前置刷新
if get_global_server_args().enable_mamba_extra_buffer():
    # 从 reqs 重建 mamba track indices，避免使用 extend 阶段过期的 mask
    set_mamba_track_indices_from_reqs(batch)
    # 清除 mask/seqlens，防止在 TARGET_VERIFY 期间错误触发 in-forward tracking
    batch.mamba_track_mask = None
    batch.mamba_track_seqlens = None

# 新增的 _mamba_verify_update 方法
def _mamba_verify_update(
    self,
    batch: ScheduleBatch,
    accept_lens: torch.Tensor,
    accept_index: torch.Tensor,
    bs: int,
) -> None:
    """Commit accepted speculative states for hybrid linear attention backends."""
    attn_backend = self.target_worker.model_runner.attn_backend
    # 如果后端不支持 Mamba 状态更新，直接返回
    if not hasattr(attn_backend, "update_mamba_state_after_mtp_verify"):
        return
    # idle 状态或没有接受 token 时无需操作
    if batch.forward_mode.is_idle() or accept_index.numel() == 0:
        return

    # 计算每个请求的最后一个正确步骤索引
    accept_indices_offset = torch.arange(
```

```
        0,
        bs * self.draft_token_num,
        step=self.draft_token_num,
        dtype=accept_lens.dtype,
        device=accept_lens.device,
    )
    req_idx = torch.arange(bs, dtype=torch.int64, device=accept_lens.device)
    last_correct_step_indices = (
        accept_index[req_idx, (accept_lens - 1).to(torch.int64)]
        - accept_indices_offset
    )

    # 如果启用了 mamba track 点更新，则准备
    if batch.mamba_track_indices is not None:
        seq_lens_pre_verify = batch.seq_lens
        seq_lens_post_verify = batch.seq_lens + accept_lens
        mamba_track_interval = self.server_args.mamba_track_interval
        to_track_mask = (
            seq_lens_pre_verify // mamba_track_interval
            != seq_lens_post_verify // mamba_track_interval
        )
        tracking_point = (
            seq_lens_post_verify // mamba_track_interval
        )
    else:
        to_track_mask = None
        tracking_point = None

    # 调用 attention backend 提交状态
    attn_backend.update_mamba_state_after_mtp_verify(
        batch,
        last_correct_step_indices,
        accept_lens,
        to_track_mask,
        tracking_point,
    )

    # 如果启用了 track 点更新，刷新 batch 中的 track indices
    if batch.mamba_track_indices is not None and to_track_mask is not None:
        set_mamba_track_indices_from_reqs(batch)
```

评论区精华

作者 xbfs 使用 AI 解决合并冲突后，reviewer hnyls2002 要求跑 NGRAM 相关测试（registered/spec/test_spec_ngram.py 和 registered/spec/test_spec_ngram_extra.py），测试结果通过。没有其他讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅限于 NGRAM 推测解码路径中的 hybrid GDN 模型，且通过条件 *enable_mamba_extra_buffer()* 保护，不影响非 hybrid 或非推测解码路径。但缺少直接针对此修复的测试用例覆盖，可能遗漏边缘情况。
- 影响：直接影响：修复 hybrid GDN 模型（如 Qwen3.5）在 NGRAM 推测解码下的输出质量，消除 token 循环和重复。影响范围小，仅涉及特定模型和推测模式组合。对系统性能无显著影响，因为 Mamba 状态提交仅在 verify 后执行，且通过条件判断避免不必要的调用。
- 风险标记：缺少测试覆盖

关联脉络

- PR #27846 fix: per-sequence last-token embedding in EAGLE3/MTP draft for batched multimodal spec decoding: 同为推测解码路径的 Mamba 状态修复，涉及 EAGLE3/MTP，与本 PR 类似但针对不同 draft 路径。
- PR #27959 [Spec] Remove the DFLASH V1 worker path: 推测解码模块重构，本 PR 的 _mamba_verify_update 思路借鉴了 *eagle_worker_v2* 中的类似逻辑。