PR #25204 完整报告

sgl-project/sglang

Fix frozen kv MTP crash when bonus_tokens is None

合并时间：2026-05-15 06:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25204

执行摘要

- 一句话：修复 Frozen KV MTP 在 bonus_tokens 为 None 时的崩溃
- 推荐动作：该 PR 值得快速合并，是一次精确的 bugfix，一行代码修复了一个影响推测解码可用性的崩溃。建议在代码审查时确认 select_last_verified_seed 的其他调用点不受影响。

功能与动机

修复 Issue #24912 报告的问题：当使用 Gemma-4 Assistant 草稿模型时，服务器在首次生成请求时崩溃，TypeError: 'NoneType' object is not subscriptable in frozen_kv_mtp_utils.py，因为 draft_input.bonus_tokens 为 None。该问题是在近期的推测重构 PR 中引入的。

实现拆解

1. 在 python/sglang/srt/speculative/frozen_kv_mtp_utils.py 的 select_last_verified_seed 函数中，将 draft_input.bonus_tokens[last_indices] 改为 draft_input.input_ids[last_indices]。
2. 原因：Frozen KV MTP 不会设置 bonus_tokens，但 input_ids 始终可用，并且包含最后验证的 token，因此使用 input_ids 可以正确获取种子 token。

关键文件：

- python/sglang/srt/speculative/frozen_kv_mtp_utils.py（模块 推测解码；类别 source；类型 core-logic）：核心文件，修复 select_last_verified_seed 函数中访问 bonus_tokens 而非 input_ids 导致的崩溃。

关键符号：select_last_verified_seed

关键源码片段

python/sglang/srt/speculative/frozen_kv_mtp_utils.py

核心文件，修复 select_last_verified_seed 函数中访问 bonus_tokens 而非 input_ids 导致的崩溃。

```
# python/sglang/srt/speculative/frozen_kv_mtp_utils.py

def select_last_verified_seed(
```

```
    draft_input: FrozenKVMTPDraftExtendInput,
) -> Tuple[torch.Tensor, torch.Tensor]:
    counts = draft_input.num_accept_tokens.to(torch.long)
    last_indices = torch.cumsum(counts, dim=0) - 1
    return (
        # 修复：使用 input_ids 替代 bonus_tokens，因为 bonus_tokens 在 Frozen KV MTP 中可能为
        None
        # 而 input_ids 始终包含最后验证的 token
        draft_input.input_ids[last_indices],
        draft_input.hidden_states[last_indices],
    )
```

评论区精华

无讨论，PR 由 kpham-sgl 直接批准，没有 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：该变更只有 1 行替换，从 bonus_tokens 改为 input_ids。风险极低，因为 input_ids 在 Frozen KV MTP 的上下文中始终可用，且语义上确实包含最后验证的 token。但需注意如果其他逻辑依赖 select_last_verified_seed 返回的是 bonus_tokens（即草稿模型生成的 bonus token 而非原始输入 token），则可能存在细微语义差异。不过从代码上下文看，该函数用于提取种子 token 进行下一轮推测，input_ids 是合理的。
- 影响：直接影响所有使用 Frozen KV MTP 推测解码的用户，特别是使用 Gemma-4 Assistant 草稿模型的场景。修复后服务器不再崩溃，推测解码可正常工作。影响范围窄，但修复了关键路径上的崩溃问题。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #24912 [Bug] TypeError: 'NoneType' object is not subscriptable in frozen_kv_mtp_utils.py when using Gemma 4 Assistant draft model: 本 PR 修复的 issue 就是 #24912，描述了相同的崩溃问题。