

PR #31622 完整报告

sgl-project/sglang

Revert "Fix mamba track-boundary seqlen under overlap scheduler (#31369)"

合并时间: 2026-07-18 08:03

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31622>

执行摘要

PR #31369 引入的 mamba track-boundary seqlen 变更导致 Qwen3-Next 模型在 decode-cache-hit 场景下 logprobs 出现确定性 KL 散度超标。本 PR 将其完全回退，恢复至已通过的基线行为。

功能与动机

PR #31369 变更了 `_mamba_check_track_boundary` 中的 seqlen 计算方式，意图修复 overlap scheduler 下的边界问题。但该变更引入了确定性正确性回归：nightly 测试 `test_qwen3_next_models.py::test_input_output_logprobs_match_decode_cache_hit_helper` 失败，KL 散度从 ≤ 0.0005 跳升至 0.0065（阈值 0.002）。二分法定位 #31369 为唯一回归源。

实现拆解

1. 完全回退提交：执行 `git revert 0675d303`，精确撤销 #31369 的变更。
2. 单文件修改：仅 `python/sglang/srt/managers/scheduler_components/batch_result_processor.py` 中的 `_mamba_check_track_boundary` 方法。
3. 逻辑还原：seq_len 计算从 `req.kv_committed_len` 恢复为 `len(req.origin_input_ids) + len(req.output_ids) - 1`。
4. 无额外配套：不含测试、配置或部署改动。

`python/sglang/srt/managers/scheduler_components/batch_result_processor.py`

核心方法 `_mamba_check_track_boundary` 的 seqlen 计算逻辑从 `kv_committed_len` 回退到 `len(origin_input_ids) + len(output_ids) - 1`，消除了回归。

```
def _mamba_check_track_boundary(self, req, batch, result, i):
    """Check if this decode step crosses a mamba track interval boundary.

    Returns (at_boundary, track_seqlen). ``track_seqlen`` must equal the
    seq_len the forward's tracking mask used, so the tracked state and its
    recorded ``mamba_last_track_seqlen`` describe the same token position.
    That seq_len is a pure function of the tokens the request has produced:
    ``len(origin_input_ids) + len(output_ids) - 1`` (the just-decoded token
    is already appended to ``output_ids`` before this runs).
```

``kv\_committed\_len`` must NOT be used here: under the overlap scheduler, ``prepare\_for\_decode`` for the *next* batch increments it before this result is processed, so it leads seq_len by a jittering lookahead (0 or 1 depending on prefill interleaving). Using it fires the boundary one decode step early on most steps, mislabeling the tracked mamba state; a later request that reuses/donates that tracked prefix then extends from a state a cold prefill recompute would not produce.

For spec decode, the boundary is detected by comparing the accepted seq_len range against interval boundaries.

"""

```
interval = get_server_args().mamba_track_interval
```

```
if batch.spec_algorithm.is_none():
    seq_len = len(req.origin_input_ids) + len(req.output_ids) - 1
    if seq_len % interval == 0:
        return True, seq_len
elif result.num_correct_drafts_per_req_cpu is not None:
    cur = req.seqlen - 1
    prev = cur - result.num_correct_drafts_per_req_cpu[i] - 1
    if cur // interval != prev // interval:
        return True, cur // interval * interval

return False, 0
```

评论区精华

- Alisonshao 通过硬件 A/B 验证：回退后 KL 散度重新低于阈值（通过 rerun-test 确认）。
- PR body 指出正确的长期修复路径是 #29792。

风险与影响

- 风险：低风险纯回退。注意 overlap scheduler 边界问题仍未解决，需等待 #29792。
- 影响：恢复 Qwen3-Next 测试通过性，无性能退化。

关联脉络

- 本 PR 回退 #31369。
- 正确的 mamba track-boundary overlap 调度修复正在 #29792 中推进。