PR #29368 完整报告

sgl-project/sglang

[Mamba] Fix long-prefill accuracy drop in radix prefix-cache state restore

合并时间：2026-07-07 04:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29368

执行摘要

- 一句话：修复 Mamba 前缀缓存长 prefill 精度下降
- 推荐动作：值得立即合并。本 PR 定位准确、修改精炼、验证充分，同时可作为优秀 bugfix 范例用于团队内部复盘：演示了如何从精度问题→跟踪日志→root cause→最小修复 的完整流程。

功能与动机

Issue #29330 报告 Nemotron-3-Ultra 在 20-shot GSM8K 评估时准确率从 94% 骤降至 20%，而 8-shot 正常。分析发现根本原因是 Mamba2Metadata 中 has_initial_states 被 mamba_track_mask 错误地 AND 拦截，导致前缀缓存命中的请求无法恢复初始状态，长 prefill 时精度崩塌。PR body 明确说明 'The gate was added in #24955; the 0.5.13 release does not have it.'

实现拆解

1. 定位根因：在 python/sglang/srt/layers/attention/mamba/mamba2_metadata.py 的 prepare_mixed 方法中，has_initial_states 决定 prefill 是否从前缀缓存槽恢复 Mamba 内部状态（conv/SSM scan seed）。该变量被 mamba_track_mask 错误地 AND 过滤——mamba_track_mask 标记的是哪些步骤会快照状态到前缀缓存，而 has_initial_states 标记的是哪些请求需要从缓存恢复状态，两者语义独立。
2. 修复方案：删除 mamba_track_mask 对 has_initial_states 的 AND 操作，使得状态恢复仅依赖 context_lens > 0，恢复至 #24955 之前的正确行为。共删除 5 行代码。
3. 验证：在 Nemotron-3-Ultra 550B（NVFP4，TP4，4×B200）和 Nemotron-Nano-9B（BF16，1×H100）上测试，20-shotGSM8K准确率分别从0.617→0.980、0.680→0.945。8-shot 准确率保持不变。吞吐量无影响。

关键文件：

- python/sglang/srt/layers/attention/mamba/mamba2_metadata.py（模块 Mamba 注意力；类别 source；类型 core-logic；符号 prepare_mixed）：核心修复文件，删除 mamba_track_mask 对 has_initial_states 的错误拦截逻辑，影响混合 Mamba 模型的前缀缓存状态恢复行为。

关键符号：prepare_mixed

关键源码片段

python/sglang/srt/layers/attention/mamba/mamba2_metadata.py

核心修复文件，删除 mamba_track_mask 对 has_initial_states 的错误拦截逻辑，影响混合 Mamba 模型的前缀缓存状态恢复行为。

```
# python/sglang/srt/layers/attention/mamba/mamba2_metadata.py
# prepare_mixed 方法中的修复片段
context_lens_tensor = forward_batch.extend_prefix_lens
assert context_lens_tensor is not None
has_initial_states = context_lens_tensor > 0
# 修复前：此处曾有 mamba_track_mask 的 AND 拦截，
# 导致缓存命中的请求无法恢复 Mamba 初始状态。
# 修复后：状态恢复仅依赖 context_lens > 0，与快照语义解耦。
prep_initial_states = torch.any(has_initial_states[:num_prefills]).item()
# 后续 chunk_offsets, chunk_indices 的计算保持不变
```

评论区精华

Review 中 b8zhong 建议删除代码中多余的行内注释（'# has_initial_states marks prefills that must seed their conv/SSM scan'），作者 adityakamat24 已执行。无其他实质争议。

- 删除多余行内注释 (style): adityakamat24 已删除该注释。

风险与影响

- 风险：变更极小（删除 5 行逻辑），且回退至 #24955 之前的状态，0.5.13 release 即不含该 gate，因此回归风险极低。需确保同一文件中 snapshot/tracking 路径（mamba.py 中的 mamba_track_mask 使用）不受影响——PR body 已确认 snapshot 逻辑未改动。
- 影响：直接修复所有混合 Mamba 模型（Nemotron-3-Ultra、Nemotron-Nano 等）在启用 radix prefix cache 时因长 prefill 导致的精度断崖下降。对 decode 性能无影响。吞吐量保持不变。影响范围限于 sglang/srt/layers/attention/mamba 模块。
- 风险标记：暂无

关联脉络

- PR #24955 [Mamba] Add mamba_track_mask to track which steps snapshot state: 此 PR 引入了导致 bug 的 mamba_track_mask gate；本 PR 正是将其误用部分回退。
- PR #29699 When attention TP for linear and full attention, use Flashinfer allreduce fusion: 同样涉及 Nemotron-H 和 Mamba 注意力模块，可能共享部分代码路径。