# PR #21514 完整报告

sgl-project/sglang

[rl][sgl] fix tensor mismatch after pause

合并时间：2026-03-27 23:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21514

## 执行摘要

该 PR 修复了 scheduler.py 中 pause_generation 函数的一个 bug，该 bug 在 batch 中所有请求完成后导致 tensor 形状不匹配，引发 KV-cache 分配错误。通过添加 is_empty() 条件检查，确保 batch 处理一致性，并新增单元测试验证修复效果。

## 功能与动机

动机源于一个运行时错误：在 pause_generation 中，当 last_batch 的所有请求完成时，filter_batch 会清空 reqs 但早期返回，tensors 仍保留旧大小，随后无条件 merge_batch 破坏了 tensor 与 reqs 数量的不变量（len(reqs) == seq_lens.shape[0]），最终导致 RuntimeError: The size of tensor a must match the size of tensor b。

## 实现拆解

实现分两个关键部分：

1. 核心逻辑修复：在 python/sglang/srt/managers/scheduler.py 的 pause_generation 函数中，将原本的无条件 merge_batch 调用替换为条件检查： python if not self.last_batch.is_empty(): if self.running_batch.is_empty(): self.running_batch = self.last_batch else: self.running_batch.merge_batch(self.last_batch) 这一逻辑与同文件的 get_next_batch_to_run 函数保持一致。
2. 单元测试新增：创建 test/registered/rl/test_pause_generation_tensor_consistency.py 文件，模拟 ScheduleBatch 的 is_empty、filter_batch 和 merge_batch 方法，验证修复后 tensor 的一致性。

## 评论区精华

在 review 评论中，reviewer ispobock 提出以下建议：

- "Can we follow this guide to write and register test?" —— 关注测试编写的规范性。
- "unit test can be moved to https://github.com/sgl-project/sglang/tree/main/test/registered/unit" —— 建议优化测试文件的位置。 这些讨论侧重于测试的流程和风格，未涉及核心逻辑争议。

## 风险与影响

风险分析：修复逻辑简单，风险较低。但需确保条件检查在所有边缘情况下正确，例如当 running_batch 为空时直接赋值。新增测试覆盖了关键场景，减少了回归风险。

影响分析：该 bug 影响 RL（强化学习）相关的调度功能，修复后提高了系统的稳定性和可靠性。用户在使用 pause_generation 时不会遇到运行时错误，提升了体验。

## 关联脉络

从历史 PR 看，该 PR 与调度模块的其他改进相关：

- PR #21490：简化 flush_cache 逻辑，也修改了 scheduler.py。
- PR #21501：修复会话关闭时的内存泄漏，涉及 schedule_batch.py。
- PR #21269：修复会话中多模态输入问题，同样涉及调度管理器。 这表明团队在持续优化调度系统的稳定性和性能，本 PR 是这一系列改进中的一环。