

PR #37164 完整报告

sgl-project/sglang

[mem_cache] Move mamba state and `retraction\_backup` into `ReqKvInfo`

合并时间: 2026-08-31 13:03

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37164>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 核心源码路径 - `python/sglang/srt/managers/schedule_batch.py` (core-logic) : 源码主路径; 涉及符号 `is_held`, `holds_kv`, `is_released`, `holds_mamba`; 包含 控制流调整、配置键调整、符号定义调整; +79/-65; 关键符号: `is_held`, `holds_kv`, `is_released`, `holds_mamba` - `python/sglang/srt/session/streaming_session.py` (core-logic) : 源码主路径; 包含 控制流调整、配置键调整; +23/-62 - `python/sglang/srt/managers/scheduler_components/batch_result_processor.py` (core-logic) : 源码主路径; 包含 控制流调整、配置键调整; +21/-21
2. 测试与验证 - `test/registered/unit/hardware_backend/mlx/test_attention_patching.py` (test-coverage) : 测试配套; 包含 测试覆盖调整、导入关系调整、配置键调整; +33/-33 - `test/registered/unit/mem_cache/test_unified_radix_cache_unittest.py` (test-coverage) : 测试配套; 包含 测试覆盖调整、配置键调整; +46/-46 - `test/registered/unit/mem_cache/test_mamba_unittest.py` (test-coverage) : 测试配套; 包含 测试覆盖调整、配置键调整; +17/-17

关键文件:

- `python/sglang/srt/managers/schedule_batch.py` (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`; 符号 `is_held`, `holds_kv`, `is_released`, `holds_mamba`) : 源码主路径; 涉及符号 `is_held`, `holds_kv`, `is_released`, `holds_mamba`; 包含 控制流调整、配置键调整、符号定义调整; +79/-65
- `python/sglang/srt/session/streaming_session.py` (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`) : 源码主路径; 包含 控制流调整、配置键调整; +23/-62
- `python/sglang/srt/managers/scheduler_components/batch_result_processor.py` (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`) : 源码主路径; 包含 控制流调整、配置键调整; +21/-21

- python/sglang/srt/hardware_backend/mlx/kv_cache/auxiliary_state.py (模块 sglang/srt; 类别 source; 类型 core-logic) : 源码主路径; 包含 控制流调整、配置键调整; +28/-28
- python/sglang/srt/mem_cache/memory_pool.py (模块 sglang/srt; 类别 source; 类型 core-logic) : 源码主路径; 包含 控制流调整、配置键调整; +26/-24

关键符号: is_held, holds_kv, is_released, holds_mamba, is_kv_released, mark_released, mark_kv_released

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR