

PR #37182 完整报告

sgl-project/sglang

[CI] Fix unreachable FakeReq field initialization

合并时间: 2026-08-31 11:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37182>

执行摘要

- 一句话: 修复 FakeReq 不可达初始化并补 Mamba 状态回归测试
- 推荐动作: 值得快速浏览 (无需精读)。该 PR 本身改动很小, 但有两个可借鉴点:
 1. 遇到 "测试突然 AttributeError" 时, 先检查是否有字段赋值被误放在 return 之后的不可达位置, 这比盲目补字段更能定位根因;
 2. save_from_req → restore_to_req 往返断言是一种低成本、高价值的会话状态回归测试范式。

若要进一步加固, 可考虑把新回归测试的断言扩展至全部会话状态字段, 并在 review 检查单中增加 "return 后不可达代码" 的排查项。

功能与动机

PR body 明确指出: #37108 inserted detach_kv() before the remaining _FakeReq field initialization. This left those assignments after return kv, so tests that call SessionSlot.save_from_req() fail with AttributeError: _FakeReq has no attribute mamba_next_track_idx. This was exposed by #34565. 核心诉求是修复测试辅助类中不可达初始化导致的回归, 让 mem_cache / streaming session 单测恢复稳定, 并对 Mamba 状态在会话插槽中的保存恢复提供回归保护。

实现拆解

本次变更全部落在 test/registered/unit/mem_cache/test_streaming_session_unit.py, 共四步:

1. 字段初始化移回 __init__: 把 mamba_next_track_idx、mamba_last_track_seqlen、mamba_branching_seqlen、to_finish、finished_reason、finished_len 六个赋值从 detach_kv() 的 return 之后搬到 _FakeReq.__init__ 末尾, 并新增 mamba_last_track_idx = None (SessionSlot 需要该字段)。这样即使测试从未调用过 detach_kv(), save_from_req() 也能读取到完整字段。
2. 精简 detach_kv(): 方法体收窄为仅执行 kv, self.kv = self.kv, ReqKvInfo() 与 return kv, 与之前语义保持一致。
3. 新增往返回归测试: test_session_slot_round_trip_preserves_mamba_state() 构造带四个 Mamba 追踪字段的 _FakeReq, 依次调用 save_from_req(req, is_first=True) 与 restore_to_req(next_req), 最后逐一断言四个字段数值完整保留, 防止后续改动再次丢失

初始化。

4. 验证与 CI: 本地环境无 pytest, 作者仅运行 pre-commit; 通过 /rerun-test test_streaming_session_unit.py 在 ubuntu-latest 上定向执行该文件, 1 个测试通过。

涉及文件与符号: test/registered/unit/mem_cache/test_streaming_session_unit.py (_FakeReq.__init__、detach_kv、test_session_slot_round_trip_preserves_mamba_state) ; 无配置、schema 或部署配套改动。

关键文件:

- test/registered/unit/mem_cache/test_streaming_session_unit.py (模块 流式会话; 类别 test; 类型 test-coverage; 符号 detach_kv, test_session_slot_round_trip_preserves_mamba_state) : 唯一变更文件。修复 _FakeReq 中因 detach_kv() 的 return 而不可达的字段初始化, 补上 SessionSlot 需要的 mamba_last_track_idx, 并新增 Mamba 状态往返回归测试, 是本次回归修复与测试守护的载体。

关键符号: _FakeReq.init, detach_kv, test_session_slot_round_trip_preserves_mamba_state

关键源码片段

test/registered/unit/mem_cache/test_streaming_session_unit.py

唯一变更文件。修复 _FakeReq 中因 detach_kv() 的 return 而不可达的字段初始化, 补上 SessionSlot 需要的 mamba_last_track_idx, 并新增 Mamba 状态往返回归测试, 是本次回归修复与测试守护的载体。

```
class _FakeReq:
    """测试用最小请求对象, 字段与真实 Req 对齐, 供 SessionSlot 往返读写。"""

    def __init__(
        self, session_id: str, req_pool_idx: int, committed: int, allocated: int
    ):
        # 会话与 KV 信息: save_from_req / restore_to_req 主要通过这些字段工作
        self.session = SimpleNamespace(
            session_id=session_id,
            streaming=True,
            finish_req=lambda req: None,
            abort_req=lambda: None,
            _inflight=False,
        )
        self.kv = ReqKvInfo(
            req_pool_idx=req_pool_idx,
            kv_committed_len=committed,
            kv_allocated_len=allocated,
            swa_evicted_seqlen=0,
            cache_protected_len=0,
        )
        # 以下字段原本被误放在 detach_kv() 的 return 之后 (不可达代码),
        # 导致 save_from_req() 读取 mamba_next_track_idx 时报 AttributeError;
```

```

# 本 PR 将全部初始化移回 __init__, 并补上 SessionSlot 需要的 mamba_last_track_idx
self.origin_input_ids = list(range(committed))
self.output_ids = []
self.extra_key = None
self.cache_salt = None
self.last_node = None
self.swa_uuid_for_lock = None
self.skip_lock_node_ids = {}
self.mamba_pool_idx = None
self.mamba_ping_pong_track_buffer = None
self.mamba_next_track_idx = None
self.mamba_last_track_idx = None
self.mamba_last_track_seqlen = None
self.mamba_branching_seqlen = None
self.to_finish = None
self.finished_reason = None
self.finished_len = None

def detach_kv(self):
    """剥离本轮 KV, 把新的空 ReqKvInfo 放回请求, 供后续写入复用。"""
    kv, self.kv = self.kv, ReqKvInfo()
    return kv

def test_session_slot_round_trip_preserves_mamba_state():
    """回归测试: SessionSlot 保存 / 恢复请求时必须完整保留 Mamba 追踪状态。"""
    req = _FakeReq("session-a", req_pool_idx=0, committed=4, allocated=4)
    req.mamba_next_track_idx = 1
    req.mamba_last_track_idx = 0
    req.mamba_last_track_seqlen = 3
    req.mamba_branching_seqlen = 2

    slot = SessionSlot()
    slot.save_from_req(req, is_first=True)

    next_req = _FakeReq("session-a", req_pool_idx=1, committed=0, allocated=0)
    slot.restore_to_req(next_req)

    # 恢复后的请求应保留全部 Mamba 追踪状态, 否则流式会话续跑时状态错乱
    assert next_req.mamba_next_track_idx == 1
    assert next_req.mamba_last_track_idx == 0
    assert next_req.mamba_last_track_seqlen == 3
    assert next_req.mamba_branching_seqlen == 2

```

评论区精华

该 PR 没有人工 review 评论, 可用信息来自 PR 评论与自动审查:

- Codex 自动审查在 commit 77bf5f6 上完成 Code Review，状态 Completed，无 findings (👍)，说明修改本身未引入新的可报告问题。
- 作者主动 /rerun-test test_streaming_session_unit.py，GitHub Actions 返回 🤖 (ubuntu-latest, 1 个测试通过)，验证修复有效。
- 值得注意：PR body 对回归来源的定位非常明确 (#37108 在 return kv 后留代码、#34565 暴露)，但仓库没有就 "如何系统性地避免同类不可达代码" 展开人工讨论，这一点在 review 流程上仍有改进空间。
- Codex 自动审查结果 (other): 自动审查通过，无遗留问题。
- 定向重跑受影响的单元测试 (testing): ubuntu-latest 上 1 个测试通过，修复有效。

风险与影响

- 风险：
 1. 回归风险极低：变更只落在单测文件中的 fake 类与新增测试，不触碰生产代码路径。
 2. 覆盖有限：新测试只断言 Mamba 追踪四个字段；to_finish、finished_reason、finished_len 等同样被移回初始化的字段没有独立的往返断言，SessionSlot 未来扩展字段时仍可能重演同类遗漏。
 3. 根因模式未在源码侧排查：#37108 的错误是 "在 return 之后写赋值语句"，这种不可达代码模式不一定只存在于测试文件，生产代码中若有类似情况不会被本次测试覆盖。
 4. 迭代依赖 CI：本地缺少 pytest 环境，意味着对单测文件的快速迭代仍需借助 GitHub Actions 验证。
- 影响：受影响范围：
 - 恢复因 AttributeError 崩溃的 mem_cache / streaming session 单测，解除对 #34565 等后续 PR 的 CI 阻塞；
 - 为 streaming session 的 Mamba 状态保存与恢复提供回归守护，降低 mem_cache 重构期（如 #37164）的状态丢失风险；
 - 对推理服务、最终用户、API 无任何影响，因为所有变更都在 test/ 目录内。

影响程度：低。属于测试基础设施层面的定向修复。

- 风险标记：纯测试文件变更，测试回归修复，覆盖有限，根因模式未排查

关联脉络

- PR #37108 (标题未提供，PR body 指出该 PR 引入回归)：本 PR 修复的直接回归来源：在 _FakeReq 中插入 detach_kv() 后，把字段初始化留在 return 之后形成不可达代码。
- PR #34565 (标题未提供，PR body 指出该 PR 暴露回归)：调用 SessionSlot.save_from_req() 的测试，最先暴露 mamba_next_track_idx 未初始化的 AttributeError。本 PR 为其解除 CI 阻塞。
- PR #37164 [mem_cache] Move mamba state and retraction_backup into ReqKvInfo: 同一 session 状态演进线：Mamba 追踪状态在 StreamSession / SessionSlot 中的保存与所有权逐步收敛，本 PR 为其提供稳定测试基座。

- PR #36958 [mem_cache] Keep req.kv non-optional and key KV ownership on req_pool_idx: 同样改动 streaming_session.py / mem_cache 状态所有权模型, 与本 PR 守护的 save_from_req / restore_to_req 路径直接相关。