PR #26225 完整报告

sgl-project/sglang

fix(swa): downgrade translate_loc_from_full_to_swa key-change log from warning to debug

合并时间：2026-05-25 02:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26225

执行摘要

- 一句话：降级 SWA 日志级别从 warning 到 debug
- 推荐动作：简单且正确的更改，建议合并。缓存键安全问题可单独追踪，但不阻碍此 PR。

功能与动机

该消息在每次元数据准备到模型前向转换过程中，当两个不同张量（如 page_table 和 out_cache_loc）在同一逻辑轮次中被翻译时都会触发。这是当前设计下的预期行为，不影响正确性。WARNING 级别日志会刷满生产日志，且在 #26184 中临时升级为 RuntimeError 后导致了 CI 失败。

实现拆解

1. 定位文件：python/sglang/srt/mem_cache/swa_memory_pool.py，第 176 行。
2. 变更操作：将 logger.warning 改为 logger.debug，其余逻辑不变。
3. 调试保留：消息内容仍可通过 --log-level debug 开启调试日志时查看，便于诊断。

关键文件：

- python/sglang/srt/mem_cache/swa_memory_pool.py（模块 内存缓存；类别 source；类型 core-logic；符号 translate_loc_from_full_to_swa）：核心变更文件：将 translate_loc_from_full_to_swa 方法中的日志级别从 warning 降级为 debug，避免误报和日志刷屏。

关键符号：translate_loc_from_full_to_swa

关键源码片段

python/sglang/srt/mem_cache/swa_memory_pool.py

核心变更文件：将 translate_loc_from_full_to_swa 方法中的日志级别从 warning 降级为 debug，避免误报和日志刷屏。

```
# python/sglang/srt/mem_cache/swa_memory_pool.py
# translate_loc_from_full_to_swa 方法中的缓存命中检查
# 当 kv_indices 张量发生变化时，会重新计算 SWA 位置映射
# 这是正常行为，但之前使用 WARNING 级别导致生产日志刷屏
key = (kv_indices.data_ptr(), kv_indices.numel())
```

```
if key != self._cached_loc_key:
    if self._cached_loc_key is not None:
        logger.debug( # 原为 logger.warning
            "translate_loc_from_full_to_swa: loc tensor changed mid-forward "
            "without invalidate_loc_cache() — possible missing call site"
        )
    self._cached_swa_loc = self.full_to_swa_index_mapping[kv_indices].to(torch.int32)
    self._cached_loc_key = key
return self._cached_swa_loc
```

评论区精华

Review 中，gemini-code-assist[bot] 提出了两个关注点：

1. 缓存键安全性：当前缓存键仅使用 (data_ptr(), numel())，对于非连续张量视图（如 t[:2] 和 t[::5]）可能产生碰撞，导致错误的 SWA 索引翻译。建议包含 shape 和 stride() 来确保正确性。但此问题与日志级别降级无直接关联。
2. 日志消息误导：消息中的 "possible missing call site" 可能误导开发者，因为这是正常操作下的预期行为。建议更新消息为更中性的描述。但 PR 未采纳该建议，仅降级了日志级别。

- 缓存键安全性 (correctness): 未解决，但 PR 目标仅降级日志，该问题可单独处理。
- 日志消息误导性 (documentation): 未采纳，PR 仅改动日志级别。

风险与影响

- 风险：风险极低。仅修改日志级别，不影响运行时行为。缓存键碰撞问题（由 gemini-code-assist[bot] 提出）是预存在的设计风险，但此 PR 未引入新风险。
- 影响：影响范围：仅限 translate_loc_from_full_to_swa 方法中的一条日志语句。 对用户：生产日志不再被无关的 warning 刷屏；需要调试时可以启用 debug 日志。 对系统：无运行时影响。 对团队：消除了因 CI 日志级别升级导致的假阳性失败（关联 #26184）。
- 风险标记：暂无

关联脉络

- PR #26184 （推测）临时将日志升级为 RuntimeError 的 PR: PR body 提到 #26184 临时升级了该日志为 RuntimeError，导致 CI 假阳性失败。