

# PR #29576 完整报告

sgl-project/sglang

Fix DSA indexer fusion bug causing excessive memory consumption.

合并时间: 2026-06-29 05:11

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29576>

## 执行摘要

- 一句话: 修复 DSA indexer 融合中 rope buffer 每层重复创建导致内存暴增
- 推荐动作: 值得精读: 这是一个典型的高内存泄漏修复案例, 展示了将重复计算提取为跨实例共享缓存的模式。代码简洁, 逻辑清晰, 对理解 DSA indexer 的内存管理有参考价值。

## 功能与动机

由 PR #29564 发现 bug: DSA indexer 融合中 rope buffer 按每层 x 上下文长度创建, 导致显存暴增。PR body 量化了修复前的严重后果: TP=4 的 GLM-5.2 模型推理容量从正常的 ~2,400k tok/pool 下降至 ~2,020k tok/pool, 说明融合 bug 显著降低了吞吐量。

## 实现拆解

1. 提取共享函数: 在 `dsa_indexer.py` 新增 `_shared_indexer_freqs_cis(rotary_emb)` 函数, 该函数使用 `getattr+setattr` 将 `freqs_cis` 缓存到 `rotary_emb` 模块的自定义属性 `_dsa_indexer_freqs_cis` 上, 确保同一 `rotary_emb` 实例只构建一次。
2. 修改 `Indexer.init`: 移除原来每层独立构建 `cos_sin_cache`→`freqs_cis` 的 6 行内联代码, 改为调用新的共享函数, 并删除了原注释行。
3. 仅 1 个文件变更: `python/sglang/srt/layers/attention/dsa/dsa_indexer.py`, +11/-6, 无其他文件或测试配套改动。

关键文件:

- `python/sglang/srt/layers/attention/dsa/dsa_indexer.py` (模块 DSA 注意力; 类别 source; 类型 core-logic; 符号 `_shared_indexer_freqs_cis`): 唯一变更文件。新增 `_shared_indexer_freqs_cis` 函数并在 `Indexer.__init__` 中替换内联代码, 解决了每层重复创建 rope buffer 导致的内存暴增问题。

关键符号: `_shared_indexer_freqs_cis`

## 评论区精华

讨论简洁, Fridge003 触发 e2e 回归测试 (`test_dsa_glm5_dp_mtp`、`test_dsa_glm5_tp_mtp`) 和内核单元测试 (`test_dsa_indexer`)。mmangkad 回复“Nice catch”确认 bug 发现。审查者 Fridge003 直接 approve, 无实质性争议。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低：变更仅限于提取一个纯函数并替换内联代码，逻辑等价性由缓存机制保证——首次调用与旧行为完全一致，后续复用缓存。唯一风险是如果不同 Indexer 实例希望有不同的 freqs\_cis（例如不同 layer\_id 影响 rope），但根据设计所有层共享相同配置，此风险不存在。回归测试已覆盖 GLM-5.2 模型的 DP 和 TP 场景。
- 影响：正向影响：修复后每 GPU 显存释放约 0.25 GB（具体取决于模型隐藏层数），在 TP=4 的 GLM-5.2 上推理容量提升约 18%。零负面影响；无 API 变动，无配置更改。
- 风险标记：核心路径变更

## 关联脉络

- PR #29564 Original bug report: 本 PR 直接由该 issue 发现 bug 并修复，PR body 中明确引用。