

PR #30088 完整报告

sgl-project/sglang

[DSA] Disable indexer fusion by default to restore DeepSeek-V3.2 accuracy

合并时间: 2026-07-04 11:50

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30088>

执行摘要

- 一句话: 默认禁用 DSA indexer fusion 修复 DeepSeek-V3.2 精度回归
- 推荐动作: 该 PR 值得合并, 以快速修复 CI 回归。建议在修复 fusion 精度问题后重新评估是否默认启用。相关精度修复工作可参见后续 PR #30342。

功能与动机

根据 PR body, CI 定时全量运行中 `test/registered/8-gpu-models/test_deepseek_v32_index_cache.py::TestDeepseekV32IndexTopkPattern` 测试失败: gsm8k 精度为 0.9295 / 0.9249, 低于阈值 0.93。根因是 DSA indexer Q/K fusion (#27705) 导致 DeepSeek-V3.2 精度退化。此前 #30018 已禁用该 fusion, 但 #30025 在合并时误将默认值改回 `False`, 导致回归。

实现拆解

变更仅在 `python/sglang/srt/envron.py` 中一行修改:

1. 将 `SGLANG_DISABLE_DSA_INDEXER_FUSION` 的默认值由 `EnvBool(False)` 改为 `EnvBool(True)`。
2. 该环境变量控制 DSA indexer 路径中 Q/K fusion 的启用状态: `True` 表示禁用 fusion, `False` 表示启用。
3. 变更后, 所有使用该环境变量的模块 (如 `dsa_indexer.py`) 将在默认配置下跳过 fusion 路径, 恢复原始精度。

关键文件:

- `python/sglang/srt/envron.py` (模块 环境配置; 类别 source; 类型 core-logic) : 唯一变更文件, 修改了 `SGLANG_DISABLE_DSA_INDEXER_FUSION` 的默认值, 直接影响 DSA indexer 的行为。

关键符号: 未识别

关键源码片段

`python/sglang/srt/envron.py`

唯一变更文件, 修改了 `SGLANG_DISABLE_DSA_INDEXER_FUSION` 的默认值, 直接影响 DSA indexer 的行为。

```
# python/sglang/srt/envron.py
class Envs:
    # ... 前面的配置省略 ...

    # DSA Backend (canonical names; fall back to SGLANG_NSA_* with deprecation warning)
    SGLANG_DSA_FUSE_TOPK = EnvBoolWithAlias(True, deprecated_name="SGLANG_NSA_FUSE_TOPK")
    SGLANG_DSA_TOPK_FLASHINFER_DETERMINISTIC = EnvBool(False)
    # ... 其他 DSA 配置 ...

    # ④ 本次变更：默认禁用 DSA indexer fusion，以修复 DeepSeek-V3.2 精度回归
    SGLANG_DISABLE_DSA_INDEXER_FUSION = EnvBool(True)

    # sgl-kernel
    SGLANG_SKIP_SGL_KERNEL_VERSION_CHECK = EnvBool(False)
    # ... 后续配置 ...
```

评论区精华

暂无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅涉及默认值切换，且提供了充分的本机验证数据（8×H200，5 轮运行）：fusion 关闭时平均精度 0.9553，方差极小；fusion 开启时平均精度 0.9306，且存在跨阈值波动。两组数据范围无重叠，确认精度回归确实由 fusion 导致。副作用是关闭 fusion 可能略微增加 stream 数量（#30025 修复后仍可接受），但精度收益远超此代价。
- 影响：影响范围较窄。仅影响使用 DSA indexer fusion 的场景（主要是 DeepSeek-V3.2 等模型在特定评测配置下的推理）。默认禁用后：
 - 用户无需手动设置环境变量即可获得稳定精度。
 - 性能上可能略逊于 fusion 开启版本，但根据 #30025 的 benchmark，对 TPOT 影响在噪声范围内（+1.96%，t=-1.48）。
 - 需要 fusion 的用户仍可通过设置 SGLANG_DISABLE_DSA_INDEXER_FUSION=0 手动启用。
 - 风险标记：配置默认值变更，精度修复

关联脉络

- PR #27705 [DSA] Indexer Q/K fusion (original feature): 引入 DSA indexer Q/K fusion 的 PR，是本次精度回归的根源。
- PR #30018 [Fix] Turn off dsa indexer fusion by default: 最初禁用该 fusion 的 PR，本次 PR 恢复了其变更。
- PR #30025 fix: reorder DSA indexer dual-stream ops to avoid CUDA graph stream explosion: 意外 re-enable fusion 的 PR，本次 PR 仅回退其环境变量翻转。

- PR #30342 Re-enable DSA indexer fusion with accuracy fix?: 讨论中提及的未来 PR, 计划在修复精度问题后重新启用 fusion。