# PR #26249 完整报告

sgl-project/sglang

[hisparse]: update user guide

合并时间：2026-05-25 17:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26249

## HiSparse 用户指南更新分析报告

### 执行摘要

该 PR 同步更新了 HiSparse 的用户指南（旧版 .md 和新版 .mdx），扩展模型支持至 DeepSeek V4，并修订了启动参数示例以匹配当前实际配置。这是一次纯粹的文档变更，无代码或逻辑修改。

### 功能与动机

HiSparse 功能已新增对 DeepSeek V4 的支持，同时默认参数和配置方式也有所演进（如 --kv-cache-dtype 和 DSA 后端自动选择）。相应地，文档需要更新以保持准确，帮助用户正确配置。

### 实现拆解

1. 更新前提条件：将模型支持从“仅 DSA 架构”扩展为“DSA 架构和 DeepSeek V4”，删除“only”等限制性措辞。
2. 补充 DeepSeek V4 传输细节：在架构图中添加说明——DeepSeek V4 的 Direct-to-Host 路径只写 C4 KV，c4_indexer 和 C128 KV 保持设备到设备传输。
3. 修订启动示例：移除 --kv-cache-dtype bfloat16 和 --dsa-decode-backend flashmla_sparse，改为 --disable-radix-cache；新增 Note 解释 KV dtype 自动选择（SM100+ 上用 fp8_e4m3，其余用 bfloat16）和 DSA 后端自动选择（bfloat16 → flashmla_sparse，fp8_e4m3 → flashmla_kv）。
4. 移除过时关键说明：删除原先列出的必须参数（--kv-cache-dtype bfloat16、--dsa-decode-backend flashmla_sparse 等），因为这些配置现在会自动处理。

### 无需展示源码。

### 评论区精华

- zijiexia：提醒文档已迁移至 docs_new，要求作者同步修改。
- hzh0425：确认已更新 docs_new 下的文档。

### 风险与影响

- 风险：无技术风险；但需注意两份文档（docs/ 和 docs_new/）内容应保持一致，避免用户困惑。

- 影响：用户可通过更新后的文档正确配置 HiSparse 以支持 DeepSeek V4，并了解最新的参数自动选择行为。

## 关联脉络

与近期 HiSparse 相关代码变更（如 #26177 修复 bug）形成配套，文档反映了功能的最新状态。