

SGLang 2026 第 37 周 (09-07 至 09-13) 技术周报

sgl-project/sglang

周期: 2026-09-07 至 2026-09-13

来源 PR: 342 • 重点 PR: 24 • 自动生成

原文链接: <http://prhub.com.cn/sgl-project/sglang/reports/2026-09-07-to-2026-09-13>

执行摘要

本周 SGLang 共合入 342 个 PR，平均重要性 6.5，整体呈现三条主线：新模型 / 新硬件支持快速扩张（Mamba、Qwen3.8 Flash Next、DSV4 TRT-LLM、NPU arch35）、配置解析体系进入“Round 6”深水区重构（ch-wan 主导）、以及缓存 / 内存子系统的大规模改造（kv-shard、HiCache 段锁、L3 后端）。性能优化在采样、预填充、解码路径均有亮点，但 68 个 PR 标记“核心路径变更”、19 个缺少测试覆盖，回归与稳定性是本周最需要关注的工程风险。

本周重点变化

- 配置体系重构：ch-wan 连续合入 #39202（删除冗余 stamp）、#38752（统一声明写入）、#38113（Round 6.5 派生字段）等，将 ServerArgs 拆分为 21 个命名空间，并引入 resolution_hooks 可插拔机制。这些改动大幅提升可维护性，但核心启动路径频繁变动，且涉及废弃 API/CLI，外部集成方需注意升级成本。
- 模型支持扩容：Qwen3.8 Flash Next 主模型（#37500）是本周最大单体 PR（+2131 行），集成 QSA 稀疏注意力、PLE 批量嵌入、HyperConnection 与 MTP；Mamba-1/2（#34556）填补了纯 SSM 推理空白；NemotronH Omni V3、VDN-H3 diffusion、SenseNova-U1.5 等也进入 main。
- 性能优化亮点：#36631 将采样掩码捕获改为 GPU 异步路径，配合 overlap 调度实现最高 52% 解码吞吐提升；#30805 为 DSV4 在 SM100/103 引入 TRT-LLM 稀疏 MLA 内核，统一 FP8 KV 布局，prefill/decode 提升 1.2-1.45x；#33089 在 NPU 上利用 top-k 稀疏性做 KV 卸载，DeepSeek DSA 解码加速 1.6x。
- 缓存与内存：kv-shard 系列第一篇 #38356 落地逻辑页放置双射与分类分配器；HiCache 用段锁协议替换 skip_lock_node_ids（#36848），消除静默锁错配；NPU 新增 L3 MemCache 后端（#38827）。注意这些改动同时触及 Rust 与 Python 双实现，一致性维护成本上升。
- PD 与 API 一致性：#36612 把失败通知抽到公共层，NIXL 也获得快速失败能力；#38690 为 /v1/responses 补齐 custom tools、加密 reasoning 回放与模型校验，协议行为向 Codex 客户端对齐。

模块与主题趋势

从标签分布看，bugfix 仍是最多（110），其次是 performance（71）、kv-cache（50），说明修 bug 与性能调优是常态。热点文件榜单中，[aiter_backend.py](#) 以 11 次改动居首，AMD 后端活跃度很高；[unified_tree_core.py](#) 与其 Rust 版本并列上榜，说明缓存核心正在双语言同步演进；[scheduler.py](#) 仍是调度核心的常青改动点。diffusion 相关 PR 有 42 个，

LTX-2.5、MiniMax-H3、Qwen-Image 等优化与文档同步推进，显示多模态生成侧投入在持续加大。

风险观察

“核心路径变更”标记达 68 个，覆盖调度、缓存、注意力、配置，远超测试覆盖类风险（19）。值得特别关注的是配置体系的多轮重构与缓存双实现同步问题。CI 层面有多个临时性调整（GB300 禁用、MI355X nightly 暂停、测试容差放宽），实质上是把验证压力后移。Mamba-1 已知需要禁用 radix cache，吞吐受限，属于已知折中；采样掩码的 PD 侧仍为 opt-in，需显式开启。当前未见新的高危外部安全风险，但内部变更的叠加效应需要持续盯防。

重点 PR 速览

- #37500 (Qwen3.8 Flash Next)：重要性 9.5，新增 Qwen4-Exp 主模型与 MTP、QSA 索引器 / 内核，是本周模型侧最重一笔，但 discussion 为空，建议后续补公开评审记录。
- #39126 (Qwen3.8 NVFP4 on DGX Spark)：把 PDL 依赖等待修复、ModelOpt 混合精度加载、文件式 PLE 后端合入 main，解决统一内存设备的显存瓶颈，并附回归测试。
- #39202 (config stamp 删除)：在 2 卡真机上审计 30+ 调用点后删除冗余 stamp，并新增 resolution_hooks 机制，讨论中“三个调用点不能外推为全部一致”是很好的方法论样本。
- #36631 (采样掩码 overlap)：消除 GPU-CPU 同步，将掩码物化推迟，吞吐最高 +52%；评审围绕默认路径零开销做了多轮优化。
- #30805 (DSV4 TRT-LLM)：新后端提升 1.2-1.45x，但保持 opt-in，且拒绝 PD 分离组合，显示工程上对稳定性的克制。
- #36848 (HiCache 段锁)：用锚节点 + 组件位集替换节点 ID 集合，让锁释放自验证，并扩展到 Mamba 组件模糊测试。

后续建议

- 补齐确定性回归：PDL wait 顺序、配置 stamp 删除、采样掩码等关键改动应有确定性测试，避免依赖人工 review；配置钩子可借鉴 #39202 的突变测试思路。
- 加强缓存双实现一致性守护：Rust/Python TreeCore 的协议或边界改动应配套跨语言测试，本周已有多起 host 锁、段锁边界修复，建议引入模糊测试持续验证。
- 配置重构影响面评估：在 release notes 中明确列出废弃 CLI 和 API，并给外部插件迁移指南；同时关注 #38752 中 _ForeignDefaults 代理对外部写入的兼容性。
- CI 稳定性优先：恢复 GB300/MI355X nightly 或给出明确替代方案，避免覆盖缺口长期存在；放宽的容差应记录原因并设置复核周期。
- 关注 AMD/NPU 专用路径质量：aiter_backend 高频修改且部分 PR 缺测试，建议为关键路径补齐硬件 CI 覆盖，防止平台特有回归漏出。