SGLang 2026 第 27 周周报（06-29 至 07-05）

sgl-project/sglang

周期：2026-06-29 至 2026-07-05

来源 PR：247 • 重点 PR：24 • 自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-06-29-to-2026-07-05

执行摘要

本周（06/29 - 07/05）仓库共合并 247 个 PR，其中 24 个被标记为重点。整体呈现三大主线：配置解析架构的声明式重构正在系统性推进，标志性 PR #30137 完成 10-PR 的全栈审查；扩散模型（Diffusion）生态在性能、功能和可观测性上获得大量投入，包括 VAE 编译预热、统一序列并行、LingBot 实时窗口等；多硬件后端（XPU、AMD、NPU、CPU）持续扩展能力边界，尤以 Intel XPU 首次支持推测解码和 CUDA-Graph 捕获最为突出。风险方面，核心路径变更 PR 占比高（44 个），测试覆盖仍然存在不足（39 个），需要团队在后续合并节奏与 QA 覆盖上加强平衡。

本周重点变化

配置解析全栈重构（10-PR stack）

由 ch-wan 主导的配置解析管道重构本周完成全栈审查与 CI 验证（#30137，重要性 9.4）。该系列将分散的命令式模型覆盖逻辑迁移到声明式注册表（Arg.resolvable 机制），新增 28 个可解析字段、运行时冻结标志与双重应用技巧。后续子 PR 逐个迁移了 DisableOverlapSchedule、Mamba Radix Cache、DeepSeek 家族、MoE 后端、page_size、attention_backend、sampling_backend 等关键配置。这一变化将使配置解析更可预测、更易审计，但可能对 ROCm 环境和下游自定义入口造成冲击。

扩散模型：性能与新功能密集落地

扩散模块本周活跃度极高，重点改进包括：

- VAE decode torch.compile 预热（#29306，重要性 9.0）：自动编译 VAE 解码路径，结合 ActiveTargetCompiledCallable 避免缓存泄漏，首次请求不再承担编译延迟。
- 统一 SP 切分与零拷贝尾部填充注意力（#30107，重要性 9.2）：通过布局不变性使填充位于全局尾部，FlashAttention varlen 无需 mask 计算，实测修复了 mova 等模型的数值误差（MAE ~1.2e-2）。
- LingBot 实时功能集（#30040，重要性 9.2）：复合输入事件、交互式 KV 窗口、延迟 VAE 编码，设计上强调状态重置与原子性，确保请求隔离。
- Ideogram 4 Cache-DiT 加速（#29631，重要性 8.9）：通过 DualTransformerExecutionMode 抽象，使双 transformer 模型共享缓存优化，单图延迟从 3.55s 降至 1.66s。
- 可观测性：新增 Diffusion 请求日志（#23049，重要性 8.9），复用 SRT 日志基础设施。

多硬件后端适配

- Intel XPU：PR #29053（重要性 9.2）为 XPU 添加了 *decode* full-graph 与 *prefill piecewise* 图捕获，基于 torch.xpu.XPUGraph 实现；PR #23180（重要性 9.2）首次在 XPU 上支持 Standalone、Eagle、Eagle3 推测解码，使用 Triton 内核实现树构建与验证。虽然部分功能默认关闭或依赖外部包，但标志着 XPU 后端走向成熟。
- AMD：修复了 ROCm 7.0 上 FP8 GEMM NaN（#29918）、*aiter* 融合在 DP attention 下崩溃（#27835）、RMSNorm 批次不变性（#28787）等关键问题。新增基于 AMD mori 库的 UMBP L3 存储后端（#25377，重要性 9.2），实现 DRAM+SSD 分层缓存，MI355X 上 TTFT 降低约 3.5 倍。
- NPU：完成 DeepSeek V4 Flash MTP 支持（#28980）、DSV4 内存池接口补全（#30001）、Qwen3-VL 扩展阶段 fused 算子路由（#29505, #29627）等。文档同步更新。

核心性能优化

- FlashKDA 融合 CUTLASS 内核（#29472，重要性 8.9）：作为 KDA prefill 可选后端，在 H20/GB200 上加速 1.3-4.6x，*safe-gate* 数值前提与 *speculative verify* 路径保护的设计保证了安全性。
- 统一 MoE 路由 Triton 内核（#29771 + #25835，重要性 8.9/8.7）：替代 CUDA AOT 内核，覆盖分组 / 非分组场景，性能与 flashinfer 持平。
- SM90 Q8KV8 FP8 稀疏 MLA 预填充（#25751，重要性 9.3）：Hopper FP8 张量核心加速 1.15-1.31x，精度损失约 2.8%。
- CUDA 图执行去重（#29625，重要性 8.8）：通过 *cudaGraphExecUpdate* 共享执行句柄，节省 GPU 内存。

模块与主题趋势

- Diffusion 成为第二大主题：本周 diffusion 标签出现在 31 个 PR 中，仅次于 bugfix。开发者正在系统性地将 SRT 层的优化（日志、编译、序列并行）迁移到 diffusion 路径，同时为特定模型（LingBot、Ideogram、Krea-2）定制加速。
- DeepSeek 持续深入：40 个 PR 涉及 *deepseek*，其中 V4 相关占多数。C128 状态分配解耦（#28612）、FlashMLA sparse prefill 默认启用（#29775）、非分页索引器（#29619）等体现了对长上下文和 low-latency 场景的持续优化。
- 代码健康度提升：*refactor*（50）和 *test*（61）标签数量高，表明团队正在主动清理技术债务（如 #29770 大规模删除死代码）、提高 CI 可靠性（#29447 模型清单工具、#29066 JIT 测试迁移）。
- CI 基础设施加强：AMD、XPU、NPU 持续注册测试与 nighty 配置，新增 Apple Silicon MLX CI（#29691）、XPU 作业监控（#29807），多平台并行验证成为常态。

风险观察

1. 核心路径变更高密度：44 个 PR 标注此风险，集中于配置重构、内存池、注意力后端与调度器。强烈建议每次合并后运行全量回归套件，尤其是横跨 8-GPU 模型测试。
2. 测试覆盖漏洞：39 个 PR 标注“缺少测试覆盖”，部分修复仅通过现有关联测试验证，缺少针对性单元测试。例如 #28980（NPU DSV4 MTP）和 #29867（短卷积后端）均缺少新测

试文件。需要推动测试先行。

3. 硬件兼容性盲区：XPU 图后端与 FlashKDA 等新路径依赖特定库版本（tilelang、weak_ref_tensor），CI 覆盖尚不完全，可能导致生产环境静默降级。
4. 配置重构的连锁反应：#30137 等 10-PR stack 正在逐层合并，虽已验证字节一致性，但老旧 global accessor 的软废弃可能导致仍使用旧 API 的外部代码意外中断。建议各下游仓库提前适配。

重点 PR 速览

PR	标题	重要性	关键影响
#30137	[refactor] Config resolution pipeline: full-stack review	9.4	配置架构声明式升级，10-PR 集成，后续合并需谨慎
#29678	feat(mem_cache): unified memory pool for hybrid Mamba/SWA	9.3	动态内存分配架构，统一 KV/Mamba 状态，性能影响广泛
#28612	Optimize C128 state pool allocation using request state pool	9.4	修复 DeepSeek V4 精度退化，C128 索引解耦 SWA 映射
#25751	[Kernel] Add SM90 Q8KV8 FP8 Sparse MLA Prefill JIT Kernel	9.3	Hopper FP8 注意力加速，性能与精度权衡
#29053	[XPU] Enable XPU graph support	9.2	Intel XPU 图捕获基础设施，为后续性能优化铺路
#23180	Speculative decoding support on XPU	9.2	XPU 上首次推测解码，Triton 内核实现

PR	标题	重要性	关键影响
#30107	[diffusion] perf: add unified SP shard helpers and zero-copy tail-pad attention	9.2	扩散模型序列并行统一，零拷贝填充注意力
#29306	[diffusion] feat: enable compile warmup for vae decode	9.0	VAE 编译预热消除首请求延迟，设计模式可复用
#29472	[KDA] Add FlashKDA prefill backend	9.0	FlashKDA 融合内核，safe-gate 保护与回退策略
#28974	[weight checker] refactor: add precision branch; allow ULP quant err	9.2	权重比较器重构，基于 ULP 量化误差容差，chunked 比较

后续建议

1. 加快配置重构的测试覆盖：10-PR stack 已审查完毕，建议在正式合并前补充针对 ROCm 和 Edge 平台的集成测试，尤其是 disable_overlap_schedule 与 mamba_radix_cache 迁移路径。
2. 扩散模型稳定性加固：VAE 编译预热与统一 SP 虽已验证回归通过，但涉及多个模型路径，建议在 H100 与 B200 上扩大一致性测试范围，预防 FlashAttention varlen 边界问题。
3. XPU 后端 CI 常态化：鉴于 XPU 的 PR 数量增多（本周 4 个高优先），建议将基础单元测试纳入 PR-CI，避免仅依赖 nightly 导致反馈延迟。
4. 监视内存池统一对旧模型的影响：Unified Memory Pool 默认关闭，但作为长期趋势，需要与 DeepSeek V2/V3 等主流模型做 benchmark 对比，确保无端到端延迟退化。
5. 降低技术债务优先级：本周 refactor+test 标签占比高，建议继续保持。特别是调度器的缓存管理（SWA/Mamba/COW）仍存在多个修复 PR（#29352, #29350, #29354），表明核心路径仍复杂，可考虑抽象为组件接口。