2026 年第 22 周（05-25 至 05-31）仓库周报

sgl-project/sglang

周期：2026-05-25 至 2026-05-31

来源 PR：279・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-05-25-to-2026-05-31

执行摘要

本周仓库共合并 279 个 PR，其中重点 PR 24 个，覆盖范围广泛、技术深度高。最显著的变化是 KV 缓存可观测性系统 KV-canary 的首次完整落地，标志着 SGLang 在推理正确性保障方面迈出了重要一步。与此同时，DeepSeek-V4 性能优化（融合 Kernel、AMD 专用内核）和注意力后端统一重构是另外两个核心亮点。多平台（AMD、NPU、CPU）的持续修复与性能提升体现了仓库对硬件生态的重视。整体而言，本周的变化特点是"可观测性 + 性能 + 重构"三位一体，为后续稳定性和可扩展性打下了基础。

本周重点变化

- KV-canary 可观测性系统上线： 包括核心数据层（#26808）、JIT 写 / 验证 / 计划内核（#26806、#26807、#26805）、扰动框架（#26816）、全缓存扫描（#26812）、安装 API（#26809）、EAGLE 集成（#26813）等 8 个 PR。该系统能够检测 KV 缓存的静默损坏，目前默认关闭，但设计上模块化且配置灵活。
- DeepSeek-V4 性能优化： #25976 引入融合的 mHC post-pre kernel，利用 TileLang 实现，小批量解码性能提升 3.35%；#26208 为 AMD 平台添加 fused compress、fused APE+pool+norm+RoPE 等内核，大幅减少 kernel launch 开销。#26383 修复了 AMD 上 CUDA Graph 捕获失败并引入多项下调优化。
- 注意力后端 CUDA Graph 统一： #26665 将 16 个注意力后端（FlashAttention、FlashInfer、Triton 等）的 CUDA 图捕获 / 重放逻辑统一为 Pattern A/B 两种模式，删除约 1500 行重复代码，并修复了 TBO capture 缺失 prefix 的问题。
- 负载监控优化： #26348 用共享内存快照替代原有的 ZMQ 轮询，使 /v1/loads 端点延迟降低 10-100 倍，同时支持 ZMQ fallback 和多节点 DP 场景。
- MLX 后端支持 Qwen3.5： #25754 引入运行时鸭式类型检测，逐层识别注意力模块，并重构了缓存布局和辅助状态快照，使重复前缀预填充延迟从 0.416s 降至 0.092s。
- 其他重要变化： #26565 新增 Step-3.7-Flash（198B MoE VLM）支持；#24994 支持 Cosmos3 世界模型；#24667 添加 Ray bundle 索引环境变量实现细粒度部署；#22848 实现 WebSocket 实时音频输入 ASR；#26402 重构 GPTQ 量化方案；#26753 修复 ngram verify 后 seq_lens_sum 不同步导致的 CUDA 越界；#25676 升级 xgrammar 0.2.1 启用结构标签。

模块与主题趋势

- KV缓存层：除了KV-canary，还涉及UnifiedRadixCache（evictionpriority、KVevents、L3 存储框架）、Mooncake 优化（RDMA 零拷贝、TCP 后端、Dummy Client 修复）、

HiCache 策略模式重构等。KV 缓存在可观测性、分层存储和性能优化方面持续演进。

- 性能与推理优化： 大量精力放在注意力融合（FlashInfer/FA merge_state 回退、FlexAttention 暴露、Cutlass MLA 集成）、推测解码优化（topk==1 跳过 softmax、EAGLE 多步 draft 修复、NextN 路径清理）、量化重构（GPTQ 分离、NVFP4 融合 kernel）。这些优化针对 Blackwell、AMD 等不同平台定制，体现了精细的硬件调优。
- 平台适配： AMD（ROCm）贡献领袖，包含 MTP CUDA Graph 修复、融合内核、CI 测试迁移；NPU（Ascend）主要修复量化、注意力、推测解码的兼容性问题，并提升精度；CPU 侧增加 GPT-OSS 模型优化、KV-cache 写入加速；Intel XPU 修复 GDN kernel 正确性和设备分配。
- 测试与 CI 设施： 注意力后端单元测试套件（#26517）、CI 覆盖率报告优化（#26619）、/rerun-test 支持 glob（#26422）、测试目录拆分等。这些改进提升了测试的可维护性和覆盖率可见性。
- Bugfix 集中区： 修复了多 tokenizer 路由 503（#26831）、DP 注意力缓冲区溢出（#26123）、PD 跨 rank 队列发散（#26394）、EAGLE chunked prefill 发散（#26800）、LoRA overlap 加载 slot 泄漏（#25413）等关键问题。

风险观察

- 核心路径变更风险： 本周标记“核心路径变更”高达 57 次，涉及调度器、注意力前向、KV 缓存管理等核心模块。Review 中发现了多个未解决的正确性问题（如 KV-canary 中的 assert 替代、输入验证缺失、CUDA Graph 内存损坏），这些若未及时修复可能在生产环境中引发偶发故障。
- 测试覆盖不足： 50 次“缺少测试覆盖”标记，尤其集中在 Step-3.7-Flash、Cosmos3、NPU 上的 DFlash 和 DeepSeek 路径。这些新功能亟需补充单元和集成测试以保障质量。
- KV-canary 配置复杂度： 启用 KV-canary 需要设置多个环境变量（如 SGLANG_KV_CANARY_PERTURB_*_PROB、SGLANG_KV_CANARY_SWEEP_INTERVAL 等），且 Review 中建议的防御性编程（异常安全检查）尚未全部落实，存在误启用或性能下降风险。
- 依赖与硬件兼容风险： DeepSeek-V4 融合内核依赖 TileLang，非预期后端可能因回退路径未经充分测试而影响性能。SM100 CuTeDSL 内核仅 Blackwell 可用，其他 GPU 需注意回退是否正确。
- 实验性组件： sgl-router（#25851）仍为草案，虽然设计文档丰富，但距离生产可用还有距离，不应在关键场景中依赖。

重点 PR 速览

PR 编号	标题	模块	关键点
#26808-1q	KV-canary 系列（8 PR）	KV-cache, observability	完整 KV 缓存校验系统，含 JIT 内核、扰动、扫描，设计值得研读，但需关注 Review 遗留问题。

PR 编号	标题	模块	关键点
#25976	DeepSeek-V4 mHC fused kernel	deepseek, performance	融合 post-pre kernel，TileLang 实现，解码 +3.35%，大小 batch 自动切换。
#26665	统一注意力后端 CUDA Graph	refactor, attention	16 个后端统一为两种模式，减少 1500 行重复代码，提升后续开发效率。
#26348	负载共享内存快照	performance, scheduling	ZMQ 替换为 mmap，延迟降低 10-100 倍，为实时负载均衡提供基础。
#25754	MLX Qwen3.5 支持	feature, mlx	鸭式类型检测，重构 MLX 缓存，预填充延迟降低 4 倍以上。
#26318	Varlen FA 加速 USPAttention	performance, diffusion	Triton 融合 pack/scatter，Qwen-Image 推理提速 15%+，显式契约设计良好。
#22587	Mooncake GPU RDMA 零拷贝	performance, multimodal	视觉嵌入直接 GPU 间传输，消除 CPU 中转，附带安全 pickle 修复。
#26402	GPTQ 量化重构	refactor, quant	按 scheme/kernel 拆分，消除平台 is_xxx 检查，架构更清晰。

后续建议

1. 跟踪 KV-canary 遗留问题： 建议团队将 Review 中未解决的 assert 替换、输入验证、CUDA Graph 缓冲区管理等问题纳入技术债追踪，并在启用前完成修复和性能基准测试。
2. 补齐测试覆盖： 优先补齐新模型（Step-3.7-Flash、Cosmos3、GLM-4.7-Flash）和 NPU/AMD 平台的单元与集成测试，确保核心路径有自动化守护。
3. 关注融合 Kernel 的兼容性： DeepSeek-V4 融合 kernel 和 SM100 CuTeDSL 内核应增加非目标硬件的自动回退测试，避免因环境不满足而静默降级。
4. 推动实验性组件成熟： sgl-router 已有良好设计，建议规划路线图，补充端到端测试和文档后逐步纳入生产环境评估。

5. 强化 CI 质量门禁： 利用新增的注意力单元测试套件和覆盖率报告，逐步提高合入门槛，尤其对“核心路径变更”类 PR 要求附带充分测试。