SGLang 仓库周报（2026 第 18 周 · 04-27 至 05-03）

sgl-project/sglang

周期：2026-04-27 至 2026-05-03

来源 PR：215 · 重点 PR：24 · 自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-04-27-to-2026-05-03

执行摘要

本周（2026-04-27 至 2026-05-03）SGLang 仓库共合并 215 个 PR，其中 24 个高重要性 PR（重要性≥8.5）。变更集中在缓存系统（HiCache/UnifiedRadixTree）、LoRA 尾部延迟优化、扩散模型扩展与基础设施重构、量化体系模块化、以及多后端（NPU/AMD/XPU/MUSA）的适配增强。此外，CI/Infra 改进密度较高，涉及技能包升级、测试恢复、权限管理和 Docker 元数据，共计 36 个 CI 标签 PR。整体来看，本周是基础设施成熟化与模型支持广度扩展并进的一周。

本周重点变化

- 缓存系统 HiCache 集成：UnifiedRadixTree 通过 PR#23316 正式集成 HiCache 框架，实现 Full、Mamba、SWA 组件的设备 - 主机分层缓存和 LRU 管理。同时 PD *decode* 侧 radix *cache*（#19746）允许 *decode* worker 复用共享前缀，减少 KV 传输，吞吐提升约 1.32x。此外，CP 同步修复（#20460）解决了 30-40 分钟运行后崩溃的问题。
- LoRA 尾部延迟优化：PR#17913 引入 LoRADrainer，通过检测饥饿适配器并强制排空运行中适配器，将 LoRA 多适配器场景的 P99 TTFT 从 40 秒降至 12 秒（降低 70%），该功能默认关闭，按需启用。
- 扩散模型能力扩展：新增 JoyAI-Image-Edit 模型支持（#22625），为 Qwen Image 扩散模型添加 ModelOpt FP8 量化（#23155），重构组件驻留管理器（#23771）统一设备生命周期，GroupNorm+SiLU 融合默认启用（#23148）加速 HunyuanVideo VAE 解码。
- 量化体系重构：AWQ 量化模块拆分为 *scheme* 架构（#21126），线性层和 MoE 层分别拥有独立 Scheme，GPU/NPU 内核调用隔离到 hardware_backend。MXFP8 MoE Group GEMM 迁移至 JIT kernel（#23833），Blackwell 典型配置延迟降低 5-15%。NVFP4 KV *cache* 策略抽象（#21954）为后续多格式支持奠定基础。
- 推测解码多后端覆盖：Apple Silicon MLX 后端实现 *decode* 异步重叠调度（#22416），NPU 适配 Qwen3.5 MTP（#20918），Mistral Medium 3.5 新增 EAGLE 支持（#24058），*accept_length* 拆分为 *num_accepted_drafts* 和 *num_accepted_tokens*（#23962）改进指标语义。
- 多后端适配加速：MUSA 平台支持 Qwen 系列模型（#23654），XPU 确定性模式（#16793）及多个 XPU 修复，AMD 修复 Grok-2、Kimi-K2.6 和 Aiter RMSNorm 等问题，NPU 支持 GLM-4.5V 和多项修复。

- CI/Infra 密集改进：技能包（SKILL）升级（#24250、#23921）引入跨框架 benchmark 和分析工具，修复夜间 CI 互相取消（#24282）、runner 池扩充（#24080）、Docker 元数据（#24090）等，提升 CI 稳定性和维护性。

模块与主题趋势

缓存与调度

HiCache 生态持续完善：除了 UnifiedRadixTree 集成，还有 HiCache L2 draft KV 缓存支持（#21125）、Prefetch 自适应内存优化（#16370）、Mamba 池泄露修复（#23496）等。PD 分离方面，decode radix cache（#19746）和 Mamba cache capping 修复（#22462）表明团队正稳步推进分布式 KV 缓存能力。

LoRA

LoRA 从性能优化走向公平性调度：LoRADrainer 是重要创新，同时模型覆盖扩展至 Qwen3.5 和 Nemotron3，EP 下 MoE LoRA 切片修复（#24171）和 DoRA 错误处理（#22125）填补了关键空白。

扩散模型

扩散模型本周新增 2 个模型（JoyAI-Image-Edit、Qwen Image FP8），组件驻留管理器（#23771）重构了设备管理架构，GroupNorm SiLU 融合（#23148、#23938）显著提升 VAE 速度。CI 方面，GT 数据源切换（#24219、#24270）和精度阈值更新表明扩散 CI 流程在规范化。

量化

量化体系进入模块化阶段：AWQ 重构（#21126）是重要里程碑，分离 scheme 和后端内核；NVFP4 KV cache 策略抽象（#21954）为 Blackwell 做准备；MXFP8 JIT kernel（#23833）展示了 JIT 编译在 MoE 场景加速效果。同时修复了多个模型的 FP8 加载问题（#23973、#23471）。

推测解码

多后端并进：MLX、NPU、CUDA 均有进展。accept_length 拆分（#23962）和 spec_accept_rate 修复（#23530）显示团队在指标准确性上下功夫。EAGLE3 mm_input 丢失修复（#23613）和 chain MTP 测试（#24192）增强覆盖。

多后端

MUSA 当前是系列第 19 个 PR，Qwen 支持落地。XPU 和 AMD 各有多个 bugfix 和新功能，NPU 除了 MTP 还修复了 OffloaderV2、Ascend 注意力后端等。

CI/Infra

本周 CI 改进数量多且分散：技能包升级、测试 runner 池扩充、nightly 并发限制、Docker 元数据、权限管理等。这反映了团队在扩大 committer 基础和提升 CI 可靠性方面的持续投入。

风险观察

1. HiCache Mooncake 接口缺失：UnifiedRadixTree 集成缺少 L3 预取方法（prefetch_from_storage、check_prefetch_progress），当前仅支持 L2，后续需跟进补齐。同时 CUDA 13 环境下的测试被跳过。
2. 量化核心路径回归：AWQ 重构将 GPU 内核移入 hardware_backend，虽已尽量保持向后兼容，但仍可能引入未发现的回归。NVFP4 和 MXFP8 等新格式依赖 FlashInfer 或 Triton 版本，兼容性需持续监控。
3. 测试覆盖缺口：42 个 PR 被标记为"缺少测试覆盖"，其中 PD 分离、EAGLE+HiCache 组合、扩散新模型等高复杂度变更尤其值得关注。部分测试阈值放宽（如 DeepSeek-V3 GSM8K 从 0.62 降至 0.60）可能掩盖精度退化。
4. CI 环境不稳定性：B200 CI 测试曾静默跳过（#24208），SMG e2e 测试因 runner 问题暂停（#24166），CUDA 13 与 HiCache 兼容问题未完全解决。这些基础设施问题可能影响周报覆盖范围的完整性。

重点 PR 速览

PR #	标题摘要	重要性	关键点
23316	UnifiedRadixTree 集成 HiCache	9.4	设备 - 主机分层缓存，组件化驱逐 / 加载，Mooncake L3 待补齐
17913	LoRADrainer 降低 P99 TTFT	9.2	P99 TTFT 降低 70%，默认关闭，按需启用
23771	扩散组件驻留管理器	9.4	策略模式管理设备生命周期，支持常驻 / 逐层卸载 / 快照
21126	AWQ 量化重构（4/N）	9.2	Scheme 架构，GPU/NPU 内核分离，修复 XPU 回归
19746	PD decode radix cache	9.2	decode 侧前缀复用，吞吐提升 1.32x，TTFT p50 降低 8.1x
22416	MLX decode 异步重叠调度	9.2	利用 lazy eval 消除 GPU 等待，Apple Silicon 性能里程碑
21954	NVFP4 KV cache 策略抽象（1/4）	9.2	两层缩放 / 块缩放两种实现，FlashInfer 内核依赖
23833	MXFP8 MoE JIT kernel（1/2）	9.2	Blackwell 加速 5-15%，JIT 编译集成模式
22997	Whisper 自动语言检测	9.2	单请求完成检测 + 转录，吞吐相对 vLLM 提升 5.8x

PR #	标题摘要	重要性	关键点
23654	MUSA 支持 Qwen 系列（19/N）	9.0	Triton TopK 内核、FlashAttention、FP8 适配
22625	JoyAI-Image-Edit 扩散模型	9.2	新模型集成，CI 烟雾测试，注意上游权重未稳定
20460	HiCache CP 同步修复	8.8	解决 30-40 分钟运行崩溃，增加 CP-aware all-reduce
24250	技能包升级	8.8	替换 benchmark，新增 incident triage、profiler 脚本
24197	设备计时器重构	9.2	计时从调度器解耦到模型执行器，新增 forward 占用率指标
23811	MiMo-V2.5 day0 支持	9.2	多模态 + 多层 EAGLE，fused QKV 权重处理
15771	Elastic EP 失败进程恢复	9.0	进程组恢复、健康检查、动态加入，依赖 Mooncake
20918	NPU MTP for Qwen3.5	9.0	GDN/ 混合线性注意力后端，线程安全改进
23594	LoRA 扩展至 Qwen3.5/Nemotron3	9.2	非门控 MoE、Mamba2 投影、修复切片 bug

后续建议

1. 测试覆盖强化：针对“缺少测试覆盖”标签最多的 PD 分离和 HiCache 组合场景，建议在下一周期集中编写集成测试，特别是 Mooncake 后端 L3 路径。
2. 量化兼容性跟踪：AWQ 重构和 NVFP4 系列 PR 刚合并，建议在 nightly 中增加更广泛的回归测试，包括不同 GPU 架构和量化组合。
3. CI 环境稳定化：CUDA 13、B200 测试跳过、SMG 测试暂停等问题需要根本解决，避免影响变更验证。
4. 扩散模型 CI 标准化：本周扩散 CI 进行了 GT 数据源迁移和官方模式引入，建议继续标准化精度阈值和性能基线。
5. 多后端对齐：随着 MUSA、XPU、AMD 后端快速迭代，建议建立跨后端的核心功能对齐测试，避免功能分化。