SGLang 仓库周报（2026 第 19 周：05-04 至 05-10）

sgl-project/sglang

周期：2026-05-04 至 2026-05-10

来源 PR：228・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-05-04-to-2026-05-10

执行摘要

本周（2026 第 19 周：05-04 至 05-10）SGLang 仓库合并了 228 个 PR，其中高亮 PR 24 个。核心主线包括：DeepSeek V4 模型的里程碑式集成、推测解码模块的系统性重构、扩散模型全链条性能优化，以及多硬件后端（MUSA、AMD、NPU、Intel XPU）支持的持续扩展。同时，可观测性工具链（dumper grafter）和 CI 基础设施也获得了显著增强。变化集中在核心推理路径，风险面集中在测试覆盖不足和核心路径变更上。

本周重点变化

1. DeepSeek V4 合并（#23882）：经过 29 批 rebase，将 DeepSeek V4 全链路推理支持合入主线，新增 MQA 注意力层、压缩 128 注意力后端、MXFP4 量化 MoE 和 JIT 内核。这是本周最大变更，但测试覆盖不足，需后续补齐。
2. 推测解码架构重构：通过拆分 EagleDraftInput（#24859）、添加自定义算法注册机制（#23991）、清理死代码和重命名字段，大幅提升模块化水平和可扩展性，支持更多模型（Gemma3/4、Kimi-K2.5）的推测解码。
3. 扩散模型性能优化：帧返回路径优化（#24616）通过文件引用避免 ZMQ 序列化；CFG 并行框架重构（#23736）引入策略引擎，实测 LTX2.3 加速 35-38%；LTX2 注意力对齐官方实现（#24313）保证数值精度；VAE 解码并行策略（#23248）可配置 tiled/patch 模式。
4. 多硬件后端推进：MUSA 后端引入 CI 工作流（#20672）和大量优化内核（#23255）；AMD 后端支持深色 FP8 MLA 注意力（#20319）、双流 MoE（#24005）并增加 nightly 测试（#24203）；NPU 支持 Trinity-mini 模型（#18172）和 MXFP8 量化（#20922）；Intel XPU 修复 MLA workspace 计算（#24372）并支持 DeepSeek V3.2（#24356）。
5. 可观测性与调试：dumper grafter 新增跨系统张量嫁接、双向传输等能力（#24507-#24513），成为新的调试工具链；重量检查器增加校验和与测试（#24537）；调度指标新增 fwd_occupancy（#24458）。

模块与主题趋势

推测解码

本周重构和扩展动作频繁。除了数据拆分和注册机制，还清理了多处理器的死代码（#24865），移除了冗余的 accept_tokens 字段（#24735），并添加了命名规范文档（#24094）。算法支持方面，新添 Gemma4 MTP（#24436）和 Kimi-K2.5 EAGLE3 MLA（#24826），覆盖更多模型。总体朝着模块化、可插拔方向演进。

扩散模型

性能优化是主旋律：融合 kernel（#24411）、预计算扰动状态（#24494）、直连 all-to-all 替代功能集合（#24366）等均在提升速度。量化方面，HunyuanVideo 新增 ModelOpt FP8（#23199），Wan2.2 支持 MXFP8 在线 / 离线量化（#20922）。同时，与官方数值对齐（#24313）和 FSDP 分片修复（#24431）保证了精度和稳定性。

核心调度与 KV 缓存

HiCache 支持 SWA（#23391）是重要进展，使滑动窗口注意力缓存可在设备与主机间分层管理，吞吐量翻倍。PD 分离修复持续进行，包括状态转移（#22665）、GC 清理集中化（#24601）、更新状态处理（#24522）等。统一 radix cache 的空 match 结果缓存（#24470）减少了分配开销。

CI 与基础设施

本周对 CI 进行了大规模调整：per-commit 测试裁剪 39 个用例至 manual（#24721），新增 Arm64 CPU CI 引导（#22123），添加 bypass-fastfail 标签加速失败跳过（#24577），扩展 PyPI 发布矩阵（#24565）。这些改动在减少 CI 资源消耗的同时，提高了灵活性。

风险观察

- 测试覆盖不足：41 个 PR 明确标记缺少测试覆盖，其中 DeepSeek V4 和新模型支持最突出，可能导致回归未被及时发现。
- 核心路径变更：34 个 PR 涉及调度器、KV 缓存、注意力后端等核心改动，需要严格的回归测试。
- TP>1 兼容性未验证：多个 PR（如 #24436、#23991）未明确测试多卡场景，潜在问题可能在部署中暴露。
- dumper grafter 性能开销：广播操作（all_gather_object）可能在高吞吐场景下成为瓶颈，需监控。
- 默认行为变更：VAE CPU offload 禁用（#24315）和 FlashInfer 工作区 OOM 修复（#24172）改变了默认路径，需确保文档同步更新。

重点 PR 速览

- #23882：DeepSeek V4 集成– 全链路支持，包含 MQA 注意力、MXFP4 MoE、JIT 内核，里程碑变更。
- #24859：拆分 EagleDraftInput– 消除阶段间 in-place 修改，提升代码可维护性。
- #23991：自定义推测算法注册– 装饰器 + 全局注册表，无需修改源码即可扩展新算法。
- #20672：MUSA CI 工作流– 为摩尔线程 GPU 添加扩散和 kernel 测试，支持多后端 CI。
- #24616：扩散帧返回优化– 用文件引用避免 ZMQ 序列化，大幅降低 IPC 开销。
- #23736：CFG 并行框架重构– 引入策略模式，LTX2.3 推理提升 35-38%。
- #23967：Nixl 异步传输– 多线程队列解耦，传输延迟降低 4x。
- #23391：HiCache SWA 支持– 统一 radix cache 中 SWA 的分层管理，吞吐量翻倍。
- #24537：重量检查器校验和– 新增 /v1/checksum 端点，增强权重调试能力。

• #24721：CI 测试裁剪– 将 39 个测试移至 manual，减少 per-commit 耗时。

后续建议

1. 补齐测试覆盖：优先为 DeepSeek V4、推测解码新模型和核心路径变更补充单元测试与集成测试，尤其关注多卡场景。
2. 验证TP>1兼容性：对本周更改为模型后端和通信机制的PR，在TP环境下进行回归测试。
3. 跟踪 dumper grafter 性能：在预生产环境中评估广播操作的延迟影响，必要时添加替代方案。
4. 固化扩散模型精度基线：随着大量对齐和优化 PR 合入，应运行全量一致性测试并更新 Ground Truth。
5. 文档同步更新：确保默认行为变更（如 VAE offload）和新增 CLI 参数在 doc 中明确记录，降低用户困惑。
6. 监控 CI 稳定性：裁剪后的 test/manual 需定期轮回归，避免覆盖遗漏。