SGLang 第 17 周周报（04/20 - 04/26）：DeepSeek-V4 部署验证与 JIT 内核重构并行推进

sgl-project/sglang

周期：2026-04-20 至 2026-04-26

来源 PR：235・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-04-20-to-2026-04-26

执行摘要

本周（04/20-04/26）SGLang 仓库合并了 235 个 PR，其中高亮 PR 24 个，平均重要性 5.72。变化集中在 DeepSeek-V4 部署文档验证、JIT 内核重构、跨平台量化支持（NPU GGUF、CPU GPTQ/AWQ）以及扩散模型能力增强（LTX2.3 HQ 流水线、LoRA、OTel 追踪）。性能优化方面，KDA 融合内核、自适应推测解码、Multi-Item Scoring 预计算索引等改进带来显著吞吐提升。CI 基础设施持续加固，修复了多个构建与测试流程问题。

本周重点变化

DeepSeek-V4 部署配方全面验证

本周共有超过 20 个文档 PR 集中处理 DeepSeek-V4 的部署指南，覆盖 H200、B200、GB200、GB300 等主流平台，CP（上下文并行）和 PD-disagg（预填充 - 解码分离）两种配置均获验证。交互式命令生成器（deployment.jsx）与 Mintlify 文档站同步迁移至主仓库 docs_new/，形成统一入口。同时新增 GB200、B300 平台支持及 H200 Pro 的 mem-fraction-static 调优参数。这些文档工作降低了用户的部署门槛，但部分生成命令未经全部平台实测，需留意注释与代码的一致性。

JIT 内核替换完成关键模块

DarkSharpness 的 Reland JIT activation（#22094）是本周最核心的内核重构：重新实现了 JIT 编译的 silu_and_mul, gelu_and_mul, gelu_tanh_and_mul，替代了此前从 sgl-kernel 静态导入的版本。修复了导致回滚的 num_token=0 边界问题，并通过条件编译（_fast_math_flags）在 Blackwell 和 ROCm 上关闭快速数学以保持精度。Ch-wan 的 Deprecate act_and_mul_triton（#23707）进一步将 MoE 的 filter_expert 逻辑内联进 JIT CUDA 激活核，淘汰了冗余的 Triton 路径，同时为 AMD/XPU 保留回退机制。这两项重构共同提升了激活层的统一性和 MoE 的执行效率。

量化与多平台支持取得突破

- NPU GGUF 量化（#17883）：TheKonka 贡献了 Ascend NPU 上的 GGUF 量化全流程，包括线性、MoE、Embedding 三层的专用方法，采用预去量化策略，在模型加载时将量化权重重构为全精度。仅验证了两种模型，且 GPU GGUF MoE 的原有错误需单独修复。

- CPU GPTQ/AWQ 4-bit（#22685）：jianan-gu 为 CPU 平台添加了 GPTQ 与 AWQ 4-bit 量化，通过 AMX 格式重打包调用 Intel AMX 后端。GPTQ v2 格式的偏移问题已在前端添加检查，但 review 中未展示具体修改，需后续验证。
- Diffusion CPU 支持（#20816）：首次为 SGLang Diffusion 引入纯 CPU 推理路径，包含 PyTorch 原生回退函数和 CPUWorker 类，支持多个扩散模型在 Intel Xeon 上运行。

扩散模型创新密集

LTX2.3 模型是本周扩散子系统的焦点：

- 高质量流水线（#23366）：mickqian 加入了两阶段生成流水线，包含 res2s 采样器（RK2 中点 SDE）、分辨率感知 sigma 调度和蒸馏 LoRA 强度控制，对齐官方 HQ 输出（PSNR 20.71 dB）。
- LoRA 支持（#23649）：为 LTX2.3 添加了 LoRA 适配器权重合并、多条件图像（首/尾帧）编码、低显存模式优化。但 review 指出两个正确性风险（non-tensor 合并、DTensor shard 错误）未确认修复。
- OTel 追踪（#21254）：jh-nv 实现了多模态生成子系统的端到端 OpenTelemetry 追踪，覆盖跨分解角色的追踪上下文传播与进程内轻量级 OTLP 收集器，对生产调试有重要价值。

性能优化多点开花

- KDA 融合内核（#23038）：将 gate 激活与 chunk-local cumulative sum 融合为单个 Triton 内核，减少 50% 内存流量，端到端吞吐提升 6-11%。
- 自适应推测解码（#21599）：通过 EMA 跟踪接受长度动态调整 speculative steps，在 EAGLE topk=1 场景下零开销切换态。CUDA 图同步风险仍待验证。
- Multi-Item Scoring 预计算索引（#22544）：在 tokenization 阶段计算分隔符位置，消除 GPU 扫描，吞吐提升约 4.5%。
- DRAM 通信消除：多个 PR 优化注意力层 DtoD 拷贝（#21985）、MoE all-reduce 守卫（#23731/23732/23734）、以及 PD streaming 批处理通知（#22658）。

模块与主题趋势

模块 / 主题	趋势	说明
DeepSeek-V4	⬆ 热度极高	文档验证占主导，部署配方覆盖多平台，但仍需关注基础模型加载环境变量和部分平台未验证的配置。
JIT 内核	⬆ 重心转移	从 activation 扩展到 grouped_topk、rmsnorm_hf，JIT 化进程加快，但 AMD CI 覆盖和精度性能力仍需平衡。
NPU/CPU	⬆ 能力补齐	GGUF 量化、扩散模型 CPU 推理、Intel XPU 流水线等新功能密集落地，但测试覆盖明显不足。

模块 / 主题	趋势	说明
MoE	⚠️ 问题集中	double-reduce bug 修复（Qwen3、DeepSeek 等）和 LoRA 内存访问修复并行，同时新增 LFM2 调优配置，稳定性在改善。
性能优化	➡️ 持续	融合内核、自适应调度、通信消减等多元优化，平均收益明确，但部分实验性特性（如 BCG）需观望。
CI 基础设施	⬆️ 加固	Docker 发布工作流重用、路径过滤修复、测试分区、重试机制等，工程效能持续提升。

风险观察

1. 核心路径变更密集（39 个 PR）：调度器、内存池、CUDA 图捕获等高敏感区域同时修改，需确保组合测试覆盖。
2. 测试覆盖缺口：26 个 PR 缺少测试覆盖，尤其 NPU/CPU 新功能和 HiCache 相关修改，建议下一周期补强。
3. AMD 平台兼容性：Hunyan V3 的 CUDA Graph 崩溃、ROCm 7.0 bpreshuffle 回退、Qwen3.5 基数缓存冲突等问题仍需 AMD 团队持续投入。
4. LoRA 合并缺陷：LTX2.3 LoRA PR 中两个 review 问题未确认修复，若在生产环境使用可能引发数值错误。
5. 实验特性稳定性：Breakable Cuda Graph（BCG）依赖 mempool 引用计数，初期版本可能存在隐式内存问题。自适应推测解码的 CUDA 图同步风险也需关注。

重点 PR 速览

- #23707 [MoE] Deprecate act_and_mul_triton：ch-wan 废弃 Triton 激活核，将 filter_expert 融合进 JIT CUDA，提升 MoE 计算效率。设计清晰，但 AMD CI 覆盖缺失。
- #17883 [NPU] GGUF 量化：TheKonka 为 Ascend NPU 新增 GGUF 量化全流程，采用预去量化策略，是 NPU 推理的重要能力补全。
- #22094 JIT activation Reland：DarkSharpness 重新引入 JIT 激活内核，修复 num_token=0 边界问题，是激活层 JIT 化的里程碑。
- #23568 Parakeet nemotron encoder：yhyang201 为 Nemotron-Nano-VL 添加音频编码器和动态分辨率支持，拓展了多模态能力。
- #21254 OTel Tracing for DiffGenerator：jh-nv 实现扩散子系统 OTel 追踪，提供跨角色追踪上下文传播，对生产监控至关重要。
- #23038 KDA 融合 gate+cumsum：yuan-luo 融合 KDA 门控和累计和内核，提速 2.2-2.65 倍，是线性注意力优化的优秀案例。
- #21599 自适应推测解码：alphabetc1 实现 EMA 驱动动态步数调整，为 EAGLE 场景带来吞吐优化，但 CUDA 图同步风险待解。

- #22931 JIT rmsnorm_hf 内核：Jiminator 添加符合 HF 语义的 RMSNorm 内核，修复量化下 MMLU 精度回归，同时保持性能。
- #22685 CPU GPTQ/AWQ 量化：jianan-gu 为 CPU 添加 4-bit 量化，扩展了非 GPU 部署能力，GPTQ v2 兼容性需后续跟进。
- #23366 LTX2.3 HQ Pipeline：mickqian 实现高质量两阶段流水线，对齐官方输出，是扩散模型质量提升的代表。

后续建议

1. 补强测试覆盖：为 NPU/CPU 新功能、HiCache 变更及实验特性（BCG、自适应推测）补充单元测试和集成测试，防止回归。
2. 推进 AMD 兼容性修复：优先解决 Hunyan V3 CUDA Graph 崩溃和 LoRA 合并缺陷，确保 AMD 平台开箱即用。
3. 统一 JIT 内核调度：随着 activation、rmsnorm、topk 等内核 JIT 化，建议规划统一的内核注册和选择策略，减少条件分支。
4. 监控 DeepSeek-V4 文档质量：验证指南中的配置参数与最新代码一致，尤其是 SGLANG_FIX_DSV4_BASE_MODEL_LOAD 等环境变量的说明。
5. 扩散模块稳定性加固：LTX2.3 HQ 流水线和 LoRA 支持已合并，但 review 指出的风险未解决，应跟进修复。