sglang 仓库 2026 第 24 周（06-08～06-14）周报

sgl-project/sglang

周期：2026-06-08 至 2026-06-14

来源 PR：335・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-06-08-to-2026-06-14

1. 执行摘要

本周仓库共处理 335 个 PR，其中 24 个重点 PR。变化主线是 speculative-decoding V2 架构迁移完成和 diffusion 推理效率大幅提升。AMD 平台支持显著增强，多项新模型（ZAYA1、MiMo V2 ASR、Nemotron-H）和 API 兼容性修复（Responses、Anthropic）达到生产可用。基础设施方面，CUDA Graph Runner 重构、cookbook 配置化、路由器可观测性等底层改进为后续迭代奠定基础。但核心路径变更和测试覆盖不足仍是主要质量风险，需在下一阶段优先补齐。

2. 本周重点变化

- Speculative Decoding V2 全面迁移：#25464 废弃 EAGLE V1，#27959 废弃 DFLASH V1，#27607 迁移 Frozen-KV-MTP，#17260 支持 ngram V2。所有推测解码算法统一运行在 V2 orchestrator + draft worker 架构下，调度器同步路径通过 _forward_isolation 泛化实现。（#27977、#27964 移除 V1 遗留代码）
- Diffusion 性能与功能双爆发：#27524 渐进分辨率加速覆盖 5 个模型（FLUX.1、Z-Image、Wan T2V 等），最高 2.32× 加速；#27736 扩展至 Ideogram 4；#27531 集成 SANA-WM 世界模型，支持密集双向与流式自回归两种模式；#28071 实现跨 GPU 空间分片 VAE 解码；#27826 优化 FLUX.1 TP 分片；#28119 重构 warmup 请求构建。
- AMD GPU 支持规模化推进：#18182 实现 FP8→MXFP4 在线重量化，#27380 为 DeepSeek-V4 添加 unified KV attention，#27656/#27630 融合 QK RMSNorm 和 sigmoid gate 内核，#27811 恢复 PCG 支持，#22985 支持 MoRI 专家负载均衡。CI 覆盖扩展至 8 个注意力后端测试（#27817），新增 extra-a 测试层（#27822）。
- 新模型适配：#26347 Zyphra ZAYA1（CCA + MoE 混合架构，#26083 NVFP4 在线量化适配 Blackwell）。#26278 MiMo V2 ASR（支持音频转录 API）。#24955 Nemotron-H DP attention + MTP。#26924 Qwen3.5 注意力门控和 MoE 共享专家融合内核。
- API 兼容与工具链：#25881 修复 Responses API，新增函数工具和流式 SSE。#25876 修复 Anthropic Messages API，对齐 thinking/cache tokens。#28149 GLM-4.7 function calling。#26885 cookbook 配置化引擎，消除跨模型代码重复。

3. 模块与主题趋势

Speculative Decoding是本周最大架构动作。从 #23906 CUDA Graph Runner 重构到 #25464 V1 废弃，再到 #27959 等 PR 完成所有算法迁移，体现清晰的“先建后弃”策略。overlap 调度成为默认模式，NGRAM、DFlash 等算法被纳入统一 pipeline。但需警惕同步路径泛化对调度延迟的影响。

Diffusion从单模型优化转向平台化：渐进分辨率、实时控制、模型无关适配器（#27698）等基础设施在建立。SANA-WM 引入流水线 stage 链管理（#27531），标志着 diffusion 推理正追上 LLM 的工程成熟度。

AMD 平台每周新增 PR 数量稳定在 40+，覆盖量化、内核融合、CI 测试、文档等全方面。但部分功能（如 MXFP4、unified KV）仍缺少多 GPU 验证，需持续投入。

基础设施重构量级大：CUDA Graph Runner 引入 Phase/Backend 模式（#23906）、NVTX 工具统一（#28165）、JIT 内核测试框架（#26706）、sgl-router 指标系统（#27591）等，将提升长期可维护性和可观测性。

4. 风险观察

- 核心路径变更密度高：59 个PR标记“核心路径变更”，覆盖 scheduler、attention backend、worker 等关键模块。V2 迁移涉及调度器循环、worker 创建、batch 处理，任何一步的 corner case 都可能引发服务中断。建议在下一个发布前加大集成测试和长期稳定性测试。
- 测试覆盖缺口：48 个 PR 标记“缺少测试覆盖”，尤其是 AMD 和 NPU 平台的新功能。例如 #27380 在 AMD 上添加 unified KV attention 但无新增单元测试，#27811 恢复 PCG 支持仅依赖已有测试。CI 门禁应强制要求新路径至少附加 smoke test。
- 内存资源管理：#25881 的 response_store 无限制增长虽已被标注，但尚未修复；#23862 修复 EAGLE KV cache 预算问题后，仍需验证在长期运行下无内存泄漏。建议尽快实现 LRU/TTL 淘汰。
- AMD 多 GPU 验证缺乏：多个 AMD 专用功能（如 #18182 #27380 #27811）仅在单 GPU 或 CI 模拟环境下验证，多卡部署可能暴露通信同步问题。需要增加至少双 GPU 的集成测试。

5. 重点 PR 速览

PR #	标题	重要性	领域
#25464	Deprecate Spec V1	9.4	speculative-decoding: 废弃 V1 Worker，删除 ~2.3k 行
#27524	Progressive resolution growing (2X+ speedup)	9.2	diffusion: 渐进分辨率加速，5 模型支持
#27531	Add SANA-WM with streaming support	9.4	diffusion: 世界模型流式推理
#28071	Enable spatial-shard VAE decode across GPUs	9.4	diffusion: 跨 GPU VAE 空间分片

PR #	标题	重要性	领域
#26083	Implement online nvfp4 quantization	9.2	quantization: Blackwell NVFP4 在线量化
#26347	Support Zyphra zaya1 model	9.2	model: 紧凑卷积注意力混合模型
#25881	Fix Responses API request handling	9.2	API: 函数调用 / 流式 SSE 修复
#25876	Fix Anthropic Messages API compatibility	9.2	API: thinking/tokens 对齐
#23906	Cuda Graph Runner/Backend Refactor	9.4	infra: 分层重构
#26885	Cookbook renovation	9.4	docs: 配置化引擎
#18182	Online MXFP4 quantization on AMD	9.2	AMD: FP8→MXFP4 在线量化
#26924	Qwen3.5 fused sigmoid/moe kernels	9.2	performance: 内核融合

以上 PR 覆盖本周 85% 以上重要变更。speculative-decoding V2 迁移和 diffusion 性能优化尤为关键，建议团队评审相关实现并规划后续回归测试。

6. 后续建议

1. 补齐测试覆盖：针对 48 个缺少测试的 PR，优先为 AMD、NPU 平台和 speculative V2 路径添加单元测试和集成测试；建议将测试覆盖阈值纳入 CI 强制 check。
2. 内存稳定性治理：推动 response_store LRU/TTL 实现（#25881 的 follow-up），并在长期运行压力测试中验证 EAGLE KV cache 内存边界。
3. 扩散模型生产化：渐进分辨率功能已默认关闭，建议收集实际用户场景的视觉质量反馈后开启默认值；SANA-WM 的 TP 支持需尽快补齐（#27531 遗留）。
4. 跨平台回归预防：核心路径变更 PR 应强制完成至少一个非 CUDA 平台（AMD/NPU）的测试，避免 NVIDIA 专属优化影响其他后端。建议在 CI 中增加跨平台自动回测。

5. 关注用户影响：废弃 SGLANG_ENABLE_SPEC_V2 环境变量（#27964）可能影响存量脚本，需在 changelog 中明确标注；cookbook 配置化迁移后应抽样验证各模型页面正确性。