2026 年第 23 周周报 (06-01 至 06-07)

sgl-project/sglang

周期：2026-06-01 至 2026-06-07

来源 PR：320・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/sgl-project/sglang/reports/2026-06-01-to-2026-06-07

执行摘要

本周（2026 年 6 月 1 日至 6 月 7 日）仓库共合入 320 个 PR，其中 24 个为重点 PR。整体变化主线聚焦于 推理核心性能优化、Diffusion 生态扩张、Cache 架构重构与 PD 分离增强。ch-wan 主导的注意力后端契约重构与 CUDA Graph 缓冲区统一标志着重构趋向收敛；Diffusion 模块新增 Ideogram4、LingBot World 实时管道等多项模型支持；DeepSeek V4 推理在 FP4 索引、PD 传输、稀疏预填方面持续优化，TTFT 显著降低。风险面集中在核心路径变更的回归风险和测试覆盖缺口，需重点关注。

本周重点变化

1. 推理核心：性能与可维护性并进

- 注意力后端重构：ch-wan 将 init_forward_metadata 拆分为三段 ABC 契约（#26735），消除 DS V4 长期依赖的 side channel，并统一 CUDA Graph 输入缓冲区管理（#26742, #27407）。这些重构简化了注意力初始化的正确性维护，但 DS V4 replay 路径仍存在未解决的正确性问题。
- 性能优化：FlashMLA KV 索引构建通过 page-block 并行化，长上下文延迟从 15us 降至 1-2us（#27320）；Gemma4 路由融合为单次 Triton Kernel，decode 吞吐提升 5.6%（#26502）；DeepGEMM 预热实现 PP 并行编译，启动时间减少约 60%（#26567）。
- Speculative Decoding：修复 EAGLE draft 在 topk>1 时的 CUDA Graph 内存越界（#27338），为 topk=1 添加快速路径（#26424），并重构 EAGLE 推断测试为共享 Fixture（#26871）。

2. Diffusion：从模型支持到实时全链路

- 新模型：新增 Ideogram4 NVFP4 原生支持（#27379）与张量并行（#27393）；支持 Command A Plus（#26106）和 Gemma4 Unified 多模态（#27167）。
- 实时推理：LingBot World 完整实时扩散管道（#26954）包括因果注意力、序列分片、WebSocket 接口；优化相机条件器缓存和传输层（#27297），延迟降低 10%。
- 性能优化：Cosmos3 融合 QK-norm+RoPE（#27096）加速 4 倍，去噪内存降低 7%（#26973）；WanVAE 解码输出 in-place clamp（#27086）等。
- WebUI：新增实时播放控制器（#27148）和超分辨率控件（#27026），用户体验改进。

3. DeepSeek V4 推理链路深度优化

- FP4 索引器（#26209）：可选启动，在 SM100 上端到端吞吐提升 1.08x，但精度下降风险需评估。
- PD 传输：HiSparse 直接主机传输（#24880）平均 TTFT 降 7-9%，引入 Mooncake 层优先布局。
- 稀疏预填：集成 flash_mla_sparse_fwd（#25418）替代 chunk prefill，内核加速 1.35x，但全量反量化开销待优化。
- 兼容性修复：修复 DP 注意力 + TP MoE 下的 reduce-scatter（#27191），Waterfill 与动态 EPLB 的兼容性（#27150）。

4. HiCache 与 PD 分解：缓存命中率与延迟改善

- decode 端预取：三层缓存（L1/L2/L3）集成（#26227），TTFT 平均降低 39.3%，P99 降低 51.9%，但 per-request all_reduce 可能成为瓶颈。
- draft KV 卸载：Mooncake 等后端支持 draft KV 零拷贝 offload（#24984），实现服务器重启后自动恢复。
- 乐观预填充（#26780）：预填充 worker 在 bootstrap 完成前开始计算，通过重试机制降低 TTFT。
- L3 存储扩展：SWA 和 DS V4 均支持 L3 存储（#26881），Mooncake 存储支持 layer-first 布局（#27454）。

5. 硬件平台支持扩展

- NVIDIA：SM120 Blackwell 桌面 GPU 通过 Triton 回退内核支持 DS V4 推理（#24692），MoE 和 FlashMLA 实现 Fusion。
- AMD：MI350X 在线 MXFP4 量化（#18005），AITER 自定义 all-gather 加速 TP 通信（#25093），DS V4 SWA 位置缓存（#26931）。
- Intel XPU：新增 Gemma 4 系列模型支持（#23280），融合 RoPE 内核（#25773）。
- NPU：自适应推测解码支持（#25644），Diffusion CI GT 生成（#24630），文档完善。

6. 测试与 CI 基础设施增强

- scripted-runtime 框架：新增调度器测试原语（#27411）、KV 池耗尽器（#27412）、分块预填充测试（#27413），为调度器复杂状态测试提供基础。
- 内存检查：异步断言探测默认在 CI 中启用（#27461），添加安静模式减少日志噪声（#27238）。
- CI 优化：分区窗口显示日期范围（#27457），GB300 测试套件（#27427），LM 博客同步卡片（#27322）。

模块与主题趋势

- 推理核心重构：本周重构工作进入高潮，ch-wan 多 PR 逐步统一注意力后端的分散初始化。这一系列重构虽然短期引入回归风险，但长期将显著降低认知负担。其设计模式（ABC 契约、声明式注册表）值得团队推广。
- Diffusion 模块活跃度持续领先：从单模型支持扩展到实时推理、量化集成和 WebUI 全链路，mickqian 贡献大量代码。但 review 中遗留的正确性问题（如 #27379 的 fallback）暴

露了快速迭代时的审查疏漏，建议加强对未采纳建议的追踪。

- HiCache 与 PD 功能密集演化：项目已进入多级缓存、多模型、多平台适配的复杂阶段。节点分裂数据丢失（#27072）、渲染竞态（#27285）等问题表明设计仍在打磨。后续建议建立回归测试套件。
- 硬件碎片化：AMD、NPU、XPU 贡献持续增加，但部分修复仅为临时绕过（如环境变量、硬编码路径），缺乏通用性，且 CI 覆盖不均衡，回归风险较高。
- Speculative Decoding 正确性仍是痛点：topk>1 与 CUDA Graph 结合的问题本月已多次回退，核心难点在于 KV indices 在捕获时的静态分配与动态 topk 之间的差距。本周的修复（#27338）是重要一步，但仍需更多边界测试。

风险观察

1. 核心路径变更的回归风险：ch-wan 系列重构改变了多个注意力后端和 CUDA Graph runner 的初始化流程，特别是 #26742 后 DS V4 replay 路径中 metadata 升级缺失尚未解决，可能影响 CUDA Graph 捕获与回放正确性。作者与 reviewer 达成共识可后续修复，但需列入跟踪。
2. 合并前问题未修复：Ideogram4 的 batch.preset 为 None 时无 fallback（#27379），Cosmos3 API 的客户端断连检查与 flow_shift 安全访问（#26926）等 review 建议未被采纳即合并，构成潜在生产隐患。
3. 测试覆盖严重不足：本周 48 个 PR 被标记"缺少测试覆盖"，新功能如 PD 乐观预填充、LingBot World 实时管道、Command A Plus 均无集成测试。仅依赖手动验证和 CI 基础配置，异常路径脆弱。
4. Speculative Decoding 波动：topk>1 的 CUDA Graph 越界问题已导致两次回退（#26981, #27138），当前修复条件为 page_size==1，其他配置可能仍未解决。
5. 硬件平台修复可移植性：AMD NPU 的部分修复（如 SGLANG_FORCE_GATE_MODE_INTERLEAVED、AITER_USE_CUSTOM_AG）依赖环境变量，缺乏自动化检测，可能在版本更新后失效。

重点 PR 速览

- #26735（重要性 9.4）：引入 3-method ABC 重构 init_forward_metadata，消除 DS V4 side channel，是注意力后端契约化里程碑。值得所有推理引擎开发人员了解。
- #26742（重要性 9.2）：统一 CUDA Graph runner 输入缓冲区到 CudaGraphBufferRegistry，消除手写填充 / 复制逻辑。设计模式可复用，但 DS V4 replay 正确性待补。
- #25418（重要性 9.2）：用 flash_mla_sparse_fwd 替代 DeepSeek V4 预填内核，实现 1.35x 加速并修复长序列 chunk prefill bug。全量反量化策略需后续优化。
- #26227（重要性 9.4）：decode 端 HiCache 三层缓存预取与 PD 增量传输，TTFT 降 39.3%。功能强大但 per-request all_reduce 是潜在瓶颈，建议关注批量优化。
- #27379（重要性 9.2）：Ideogram4 NVFP4 原生支持，包含 bitsandbytes 适配、Comfy 布局推断和量化线性层工厂模式。review 中遗留的 batch.preset fallback 问题需后续修复。
- #24692（重要性 9.2）：SM120 Blackwell 桌面 GPU 全面支持 DS V4 推理，通过 Triton 回退内核弥补缺失功能，是跨代兼容性工作的参考。

- #27297（重要性 9.2）：LingBot 实时传输延迟降低 10%，通过相机条件器缓存和 raw bytes 传输实现。性能优化范例，测试覆盖充分。
- #26954（重要性 9.4）：LingBot World 实时扩散管道合并，涵盖因果去噪、序列分片、WebUI 等，是 diffusion 模块实时化的重要一步。

后续建议

1. 尽快验证 DS V4 CUDA Graph 正确性：针对 #26742 遗留的 replay metadata 问题，建议安排专项测试，覆盖 DP/PP/PD 多种组合，确认无回归后关闭 issue。
2. 建立 review 建议跟踪机制：对于未在 merge 前修复的评论，可添加 followup 标签并在 PR 描述中明确后续计划，避免关键问题悬浮。
3. 补齐新功能测试：Priority 1 目标应为 PD 乐观预填充、LingBot World 实时管道、Command A Plus 等新功能编写至少一个端到端测试，并纳入 CI。
4. 评估 Spec Decoding topk>1 稳定性：在 page_size>1 配置下运行压力测试，考虑将 topk=1 设为默认以减少风险。
5. 监控硬件平台 CI 健康度：AMD/NPU/XPU 的 CI 通过率应作为项目质量指标，推动清理硬编码路径和临时环境变量，鼓励平台通用实现。