vLLM 2026 第 28 周周报 (07-06 至 07-12)

vllm-project/vllm

周期：2026-07-06 至 2026-07-12

来源 PR：258 • 重点 PR：24 • 自动生成

原文链接：http://prhub.com.cn/vllm-project/vllm/reports/2026-07-06-to-2026-07-12

1. 执行摘要

本周 vLLM 仓库共计合并 258 个 PR，其中重点 PR 24 个。整体变化集中在三条主线：KV 缓存架构重构与优化、推测解码能力增强、平台扩展（特别是 ROCm）与模型清理。此外，安全修复与 CI 改进也有显著投入。核心路径变更占比最高（41 个 PR），提示回归风险需重点防范。

2. 本周重点变化

KV 缓存布局重构：PR #44455 将注意力后端的 KV 缓存形状从 5D 统一为 4D，K 和 V 打包在最后维度，通过 transpose 和 split 零成本恢复。这一变更简化了量化缓存处理、减少了 KV-connector 耦合，并为后续 PD 分离场景铺路。同时 PR #46384 为混合模型（Full Attention + Mamba）实现了 token 精度的前缀缓存命中，通过 prefix_match_unit 配置解耦哈希粒度，大幅提升前缀重用效率。

推测解码全面进化：多个 PR 补全了推测解码的关键拼图。PR #46725 引入运行时草稿权重更新机制，允许在 RL 训练循环中独立更新 draft 模型参数，且妥善处理 sleep/wake 场景。PR #47419 将 DeepSeek-V4 DSpark 推测解码移植到 AMD ROCm，在 MI350X 上实现 1.59x 吞吐加速。PR #40996 使 DCP（Decode Context Parallelism）支持混合注意力模型，非 DCP 层保持本地状态处理。

Transformers 后端性能对齐：PR #47187 新增基于 torch.fx 和 AST 的自动融合器，可检测并替换 QKV 线性层、GLU、MoE 块和 RMSNorm 为 vLLM 高效实现，使 Transformers 模型后端的性能与原生 vLLM 持平甚至更优，显著降低模型迁移成本。

模型清理与 Rust 前端增强：PR #48100 和 #48096 分别移除了 Olmo/Olmo2 和 Persimmon/Fuyu 的原生实现，转向 Transformers 后端。Rust 前端方面，PR #47959 集成视频多模态支持，并重构为每模态准备架构；PR #47454 新增端点插件框架，支持通过环境变量安全扩展 HTTP 路由。

3. 模块与主题趋势

从标签分布看，bugfix 占比最高（128 个），其次为 v1 模块（88 个）和 performance（51 个），表明团队在稳定性和性能优化上持续投入。ROCm 平台有 31 个 PR，涉及 DSpark 支持、AITER 自定义 AllReduce、MXFP8 优化和多种调优配置，是本周最活跃的平台。CI/ 构建方面 36 个 PR，包括测试并行化、镜像构建、超时调整等，体现了对 CI 效率的重视。

热点文件集中在推理流水线核心：gpu_worker.py（7 次）、gpu_model_runner.py（7 次）、scheduler.py（6 次）以及 deepseek_v2.py（5 次），这些文件也是 KV 缓存和推测解码改造

的焦点。此外，tests/models/registry.py 也被频繁修改（6 次），反映了模型注册表的重构趋势。

4. 风险观察

本周最大的风险源是 核心路径变更（41 个 PR）。特别是 KV 缓存布局重构（PR #44455）引入了 block-major 布局假设和 CoW 拷贝顺序依赖，虽已在讨论中缓解但仍有剩余风险。测试覆盖不足（36 个 PR）是第二高发的风险，例如 PR #47857 的 resume 路径缺陷已被记录但未修复。推测解码场景中，异步调度下的窗口释放（PR #47728）和 KVBlockZeroer 数据竞争（PR #48085）已被修复，但 RL 训练的实际稳定性仍需验证。此外，视频 / 媒体处理的安全风险（SSRF、路径注入、资源耗尽）部分已修复，但覆盖完整度尚需加强。

5. 重点 PR 速览

- #44455 KV-Cache Layout Refactor [2/N]：统一注意力后端 KV 缓存形状为 4D，K/V 打包在内容维度，简化量化、连接器和后端接口。涉及 FlashAttention、FlashInfer、Triton、ROCm 等全部后端。
- #46384 混合模型前缀缓存细粒度命中：为 Full Attention + Mamba 混合模型实现 token 精度前缀匹配，通过 prefix_match_unit 解耦哈希与物理块大小，提升缓存效率。讨论中记录了 SWA 部分命中和 CoW 顺序依赖等遗留问题。
- #46725 推测解码运行时草稿权重更新：引入 start_draft_weight_update API 和 WeightTransferEngine 目标切换机制，解决 RL 训练中草稿参数陈旧问题。设计上采用独立方法而非 include_draft 标志，团队达成一致。
- #47419 ROCm DeepSeek-V4 DSpark：将 DSpark 推测解码移植到 AMD 平台，通过 EAGLE3 aux hidden state 接口和 ROCm 特定 kernel 分发实现 1.59x 加速，是平台扩展的关键里程碑。
- #47187 Transformers 后端自动化融合：利用 torch.fx 追踪 forward 图，通过 AST 重写匹配 QKV、GLU、MoE 和 RMSNorm 模式并替换为 vLLM 高效实现，使 Transformers 后端性能与原生持平。
- #47959 Rust 前端视频多模态集成：在 Rust 前端中支持 OpenAI 风格 video_url，重构模态处理为 ModalitySupport + PreparedMedia 架构，方便扩展。SSRF 风险已标记但未修复。
- #47454 端点插件框架：通过 Python entry points 定义 EndpointPlugin 协议，支持在 vLLM API 服务器上添加自定义 HTTP 路由，默认不加载，通过 VLLM_PLUGINS 环境变量白名单控制，安全可控。

6. 后续建议

1. 加强核心路径回归测试：KV 缓存布局重构和混合模型前缀命中涉及多个组件的接口变更，建议在合并后增加端到端性能与正确性测试，特别是针对 PD 分离和 async scheduling 场景。
2. 填补测试覆盖缺口：针对 36 个缺少测试覆盖的 PR，优先为 resume 路径（#47857）、CoW 拷贝顺序（#46384）、推测解码并发（#47728）等关键路径补充集成测试。

3. 持续监控 ROCm 稳定性：ROCm 平台本周新增大量 PR，包括 DSpark、AITER 和 MXFP8 优化，需在 AMD CI 中启用更多测试并跟踪 MI350X 上的长期稳定性。
4. 安全修复完整覆盖：虽然 SSRF、路径泄露等问题已修复，但视频处理的安全边界（如像素限制检查）仍需系统化审查，建议启动一次安全审计。