PR #21563 完整报告

sgl-project/sglang

Split workflow for releasing runtime docker

合并时间：2026-03-28 06:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21563

执行摘要

此 PR 将 runtime Docker 镜像的发布流程从原有 workflow 中拆分出来，形成独立的 release-docker-runtime.yml workflow，以优化 CI 管理。变更影响 Docker 发布流程，旨在提高可维护性和构建效率，但需注意拆分后的同步风险。

功能与动机

动机是将 runtime 和 framework Docker 镜像的发布分离，简化构建流程。从修改文件 release-docker.yml 的注释变更可见（从“both framework and runtime”改为“framework Docker images”），现在只构建 framework 镜像，runtime 镜像移至新 workflow，可能为了减少复杂度或优化构建时间。

实现拆解

主要改动点：

- 新增 release-docker-runtime.yml：定义完整的 runtime 镜像发布 workflow，包括触发条件（如 tag 推送或手动触发）、版本验证、构建和推送步骤，支持 x86 架构和 CUDA 12.9.1/13.0.1 版本。
- 修改 release-docker.yml：删除所有 runtime 相关构建步骤（例如 Build and Push AMD64 Runtime），并更新文件头部注释以明确其现在只负责 framework 镜像。

评论区精华

无 review 评论或讨论，PR 直接合并。

风险与影响

风险：拆分后需确保两个 workflow 在版本发布时同步触发，避免镜像版本不一致；新 workflow 可能引入构建错误或性能问题；构建步骤变更可能影响现有 CI 流程的稳定性。具体到文件，release-docker-runtime.yml 中的构建参数（如 CUDA_VERSION）需仔细验证。 影响：用户无直接影响，Docker 镜像发布照常；系统 CI 更模块化，可能提高效率；团队需适应新 workflow 结构，更新相关文档或流程。

关联脉络

与此 PR 相关的历史 PR 包括 PR 21562（放宽 CI 测试阈值）和 PR 21527（修复 CI 监控问题），它们也涉及 CI 基础设施的优化，表明团队在持续改进 CI 流程，本 PR 是这一趋势的一部分。