

PR #31610 完整报告

sgl-project/sglang

docs(cookbook): replace pinned nightly/dev images with :latest

合并时间: 2026-07-18 07:33

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31610>

执行摘要

本次 PR 将 cookbook 中 13 个 mdx 文件和 3 个 snippets 配置文件中固定的 nightly/dev Docker 镜像标签统一替换为稳定的 `lmsysorg/sglang:latest`, 解决了因镜像过期导致读者 `docker pull` 失败的问题。同时更新了相关注释和表格, 简化多 CUDA 变体说明。

功能与动机

根据 PR 描述, 许多 cookbook 页面仍然固定了 nightly 或模型 /PR 特定的 dev Docker 镜像 (如 `nightly-dev-DATE-sha`、`dev-<model>` 等)。这些镜像随着 nightly 被清理和 dev 标签轮转而腐烂, 导致读者的 docker pull 失败。此次将它们改为稳定的 `lmsysorg/sglang:latest`。

实现拆解

1. 配置文件标签替换: 在 `docs_new/src/snippets/configs/` 下的三个 jsx 文件中 (`laguna-m1.jsx`、`minimax-m3.jsx`、`laguna-xs21.jsx`), 将 `dockerImages` 对象中的所有 GPU 变体标签从 `dev/nightly` 替换为 `lmsysorg/sglang:latest`, 并更新注释消除过时的 CUDA 变体说明。
2. cookbook 文档更新: 在 `docs_new/cookbook/autoregressive/` 下的 13 个 mdx 文件中, 替换 Docker 拉取命令中的特定标签, 简化镜像表格, 将多行表格合并为一行, 突出: `latest` 的多架构特性。
3. Nemotron3-Ultra 反复调整: 该页面经历了三次提交: 最初替换为 `:latest`, 然后改为 `:dev` (担心支持未合入), 最终与 #31094 对齐改回 `:latest`。确认 #28309 是性能优化, 不阻塞模型支持。
4. 保留不变范围: 明确保留 `rolling dev` 标签、固定稳定版本标签 (`v0.5.x`)、ROCm 镜像、benchmark 来源记录以及 Inkling 页面。
5. 无测试改动: 但 PR body 列出需要运行加载验证的模型清单。

关键源码片段

`docs_new/src/snippets/configs/poolside/laguna-m1.jsx`

核心配置文件之一, 包含 5 个 GPU 变体的 `dockerImages` 映射。替换前使用固定的 nightly 镜像 (`dev-cu13-618-nightly`), 现统一为 `:latest`, 并更新注释说明。

```
// 之前这里使用了固定的 nightly 镜像 (如 `dev-cu13-618-nightly`),  
// 现在替换为稳定的 `:latest` 标签, 覆盖 H200 和所有 Blackwell GPU。
```

```
dockerImages: {
  h200: "lmsysorg/sclang:latest",
  b200: "lmsysorg/sclang:latest",
  b300: "lmsysorg/sclang:latest",
  gb200: "lmsysorg/sclang:latest",
  gb300: "lmsysorg/sclang:latest",
},
```

docs_new/src/snippets/configs/MiniMaxAI/minimax-m3.jsx

核心配置文件，包含 6 个 GPU 变体的 dockerImages 映射。替换前每个架构使用不同 dev 镜像（如 dev-minimax-m3、dev-cu13-minimax-m3 等），现统一为 :latest，并保留 AMD ROCm 部分不变。

```
// 之前每个 GPU 架构使用不同的 dev 镜像（如 `dev-minimax-m3`、`dev-cu13-minimax-m3` 等），
// 现在统一使用 `:latest`（cu13，多架构），覆盖 H200 + 所有 Blackwell。
// AMD ROCm 镜像保持不变（因为无 `:latest` 等价物）。
dockerImages: {
  b200: "lmsysorg/sclang:latest",
  b300: "lmsysorg/sclang:latest",
  gb200: "lmsysorg/sclang:latest",
  gb300: "lmsysorg/sclang:latest",
  h200: "lmsysorg/sclang:latest",
  mi300x: "aigmkt/minimax-m3-sclang-rocm700-mi30x",
  mi325x: "aigmkt/minimax-m3-sclang-rocm700-mi30x",
  mi350x: "aigmkt/minimax-m3-sclang-rocm720-mi35x",
  mi355x: "aigmkt/minimax-m3-sclang-rocm720-mi35x",
},
```

docs_new/src/snippets/configs/poolside/laguna-xs21.jsx

核心配置文件，包含 3 个 GPU 变体的 dockerImages 映射。替换前使用专用 dev 镜像（dev-cu13-laguna-xs-2-1），现为 :latest。

```
// 之前使用专用的 dev 镜像（`dev-cu13-laguna-xs-2-1`），
// 现在替换为 `:latest`，因为该构建已合并到稳定版本中。
dockerImages: {
  h200: "lmsysorg/sclang:latest",
  b300: "lmsysorg/sclang:latest",
  gb300: "lmsysorg/sclang:latest",
},
```

评论区精华

该 PR 没有引发公开 review 评论，只有一个 approval。但在 commit 历史中，Nemotron3-Ultra 的标签经历了三次反复，反映了团队在激进统一与保守保留之间的权衡。最终决策基于“#28309 是性能优化，不影响支持”，因而采用 :latest。

风险与影响

风险：部分模型（如 Nemotron3-Ultra、MiniMax-M3）的支持可能尚未合入 :latest，读者可能遇到失败。PR body 已列出需要验证的模型列表。整体风险较低，因为 :latest 应包含主线最新代码。

影响：正面影响是读者不再因镜像过期而失败，负面影响是少数模型可能需要回退到特定标签。对文档维护者减少了过时标签的维护负担。

关联脉络

本 PR 是 #31508（清理部分过时标签）的延续，将清理范围扩大到几乎所有 cookbook 页面。同时在 Nemotron3-Ultra 的处理上与 #31094（同步文档变更）对齐，保持了文档的一致性。