

# PR #33976 完整报告

sgl-project/sglang

[NPU] [DOC] Upgrade recommended sglang version on Ascend NPU

合并时间: 2026-08-07 17:45

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33976>

## 执行摘要

- 一句话: Ascend 文档升级至 v0.5.16, GLM-5.2 切换稳定镜像
- 推荐动作: 无需精读源码逻辑, 但值得浏览 glm\_5\_2.mdx 的镜像切换和 qwen3\_6\_35b\_a3b.mdx 中移除 NEXTN 失败警示的文案, 体会版本发布后文档同步的流程。若你维护 Ascend 相关部署脚本, 需要注意镜像标签命名约定已从 v-cann 变为 cann-v。

## 功能与动机

PR body 明确两个目标: 升级 Ascend NPU 上推荐的 sglang 版本、更新 Ascend NPU 支持特性清单。从提交历史看, v0.5.16 发布后 GLM-5.2 的 0Day 支持已合入主线并进入稳定版本, 原先文档中的 daily build 镜像 (swr.cn-southwest-2 私有仓、glm5.2-20260615 标签) 不再必要; 同时用户调参 DeepEP 时需要 DEEPEP\_HCCL\_BUFFSIZE 参考, 投机解码也新增了 EAGLE 选项, 需要同步到文档。

## 实现拆解

1. 版本基线升级: 在 qwen3\_5\_397b、deepseek\_r1、deepseek\_v3\_2、glm\_5\_1、kimi\_k2\_6、mimo\_v2\_flash、minimax\_m2\_5、qwen3\_235b\_a22b 等 10+ 个教程文档中, 将 validated based on SGLang v0.5.13/v0.5.14 的表述统一改为 v0.5.16, 并同步建议用户使用 v0.5.16 或更高版本。
2. 镜像标签统一替换: 拉取命令由 quay.io/ascend/sglang:v0.5.13.post1-cann9.0.0-a3 批量改为 quay.io/ascend/sglang:cann9.0.0-a3-v0.5.16, 新命名约定将 cann 版本前置、SGLang 版本后置; Atlas 800I A3 与 A2 两个 Tab 同步更新。
3. GLM-5.2 稳定化: glm\_5\_2.mdx 删除 daily build 镜像及其 Note 说明, best practice 章节将“0Day 仍在 daily build、tuning 未定稿”的表述解耦为仅保留 tuning 未定稿, 避免误导。
4. 特性文档补全: environment\_variables.mdx 新增 DEEPEP\_HCCL\_BUFFSIZE 环境变量说明 (+5 行); 并补充 EAGLE 到 --speculative-algorithm 支持值列表, 对应两个独立 commit。
5. 配套入门文档同步: installation.mdx、quick\_start.mdx、accuracy\_evaluation.mdx 等同步版本号, 确保从安装到评估的文档链路一致。
6. 无源码与测试联动: 本 PR 仅为 docs 变更, 不涉及 Python 代码、schema 或 CI 配置。

关键文件:

- docs/docs/hardware-platforms/ascend-npus/model-deployment/tutorials/glm\_5\_2.mdx (模块 NPU 文档; 类别 docs; 类型 documentation) : 变化最大 (+8/-17) : GLM-5.2 从 daily build 镜像切换为 quay.io 稳定 v0.5.16 镜像, 删除相关 Note, 并更新 best practice 表述, 是本 PR 的核心变更。
- docs/docs/hardware-platforms/ascend-npus/model-deployment/tutorials/qwen3\_5\_397b.mdx (模块 NPU 文档; 类别 docs; 类型 documentation) : 典型示例文件: 版本从 v0.5.13 升级到 v0.5.16, 镜像标签切换为新命名约定, 覆盖 A3/A2 两个 Tab, 是批量替换的代表。
- docs/docs/hardware-platforms/ascend-npus/model-deployment/tutorials/qwen3\_6\_35b\_a3b.mdx (模块 NPU 文档; 类别 docs; 类型 documentation) : 除版本升级外, 还移除了 v0.5.13(.post1) 下 NEXTN 崩溃 (NPUMHATokenToKVPool 缺少 data\_strides) 的警示, 说明新版本已修复, 对用户有明显价值。
- docs/docs/hardware-platforms/ascend-npus/reference/environment\_variables.mdx (模块 NPU 文档; 类别 docs; 类型 documentation) : 新增 DEEPEP\_HCCL\_BUFFSIZE 环境变量说明 (+5 行), 对应独立 commit, 是支持特性清单更新的关键文件。
- docs/docs/hardware-platforms/ascend-npus/getting-started/installation.mdx (模块 NPU 文档; 类别 docs; 类型 documentation) : 入门安装文档同步版本号, 确保新用户从安装开始看到一致版本, 是文档链路一致性的重要一环。

关键符号: 未识别

## 评论区精华

PR 无人工 review 评论, sglang-npu-bot 直接 APPROVED 后合并, 因此没有可见的设计权衡或未解决问题。值得注意的是镜像标签命名约定 (cann 版本前置) 属于批量替换, 缺少人工对旧标签外部引用和站内链接的逐一核查。

- 暂无高价值评论线程

## 风险与影响

- 风险: 镜像标签命名约定变更: 如果 quay.io/ascend/sglang 上 v0.5.16 系列标签不存在或与文档不一致, 用户按新文档拉取会直接失败, 且旧格式标签在文档中已被移除, 回退路径变窄。GLM-5.2 稳定镜像切换: daily build 可能包含尚未进入 stable 的修复, 切到稳定镜像后用户只能等待下个版本, 若 v0.5.16 存在 GLM-5.2 回归则文档无法指导回退。多文档一致性: 23 个文件批量替换, 需确认没有遗漏旧版本引用、锚点链接或与搜索索引 (如 Algolia) 的冲突。影响面: 仅影响文档准确性, 不涉及运行时逻辑, 整体风险较低。
- 影响: 对用户: Ascend NPU 部署用户将按新文档拉取 v0.5.16 镜像, 获得稳定的 GLM-5.2 支持, 但需适应新镜像标签命名; 对系统与代码无任何运行时影响; 对团队而言文档维护成本低, 但需要确保新标签对应镜像真实可用并已发布。影响范围限定在 Ascend 部署文档, 影响程度中低。
- 风险标记: 镜像标签命名约定变更, GLM-5.2 稳定镜像切换, 23 文件批量替换需一致性核查

## 关联脉络

- PR #32120 [AMD][DI][CI] 8/N Add GLM-5.2 MXFP4 1P1D DI/CI recipes (base + MTP + DP8/EP8): 同为 GLM-5.2 0Day 支持的多平台铺开工作（AMD 侧新增 CI recipes），本 PR 将 Ascend 侧 GLM-5.2 文档切换到稳定发布镜像，共同反映 0Day 支持收敛到正式版本。
- PR #33688 [Diffusion]Skipping tensor copying for non-BCG GLM-Image workflows: 同为 Ascend NPU 平台相关改动（NPU 性能优化与基线更新），说明 NPU 支持线持续推进，本 PR 随之把文档基线对齐到 v0.5.16。
- PR #33935 Clean GLM-5.2 NVFP4 cookbook: 同为 GLM-5.2 文档清理类 PR（NVFP4 cookbook），与本次 Ascend 教程版本升级都属于发布后的文档同步工作。