

PR #29400 完整报告

sgl-project/sglang

[Docs] Cookbook: match playground docker image resolution to deployment (hwlquant)

合并时间: 2026-06-26 15:56

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29400>

执行摘要

本 PR 修复了 Cookbook 文档中 Playground 组件 Docker 镜像选择不准确的问题。此前 PR #29380 已为 Deployment 组件添加了按 `hwlquant` 组合键解析镜像的逻辑，但 Playground 组件仍仅按 `hw` 查找，导致 GLM-5.2 等使用量化专用镜像的模型在 Playground 中显示错误的 Docker 命令。本次改动将 Playground 的镜像解析方式与 Deployment 对齐，确保两处文档输出一致。

功能与动机

PR #29380 更新了 Deployment 引擎的镜像解析器，使其能够按 `hwlquant` → `hw` 的顺序查找 Docker 镜像。但 Playground 引擎拥有自己的镜像解析器副本，仍仅按 `hw` 键查找。结果对于配置了量化专用镜像的模型，Playground 中显示的 Docker 命令使用了错误的镜像。例如 GLM-5.2 NVFP4 在 Playground 中渲染为 `lmsysorg/sglang:latest`，而 Deployment 组件正确显示为 `lmsysorg/sglang:dev-glm52-nvfp4`。

实现拆解

1. 提取配置对象— 在 `_playground.jsx` 的 `renderCommandLines` 函数中，使用 `const di = config.dockerImages || {}`；安全地获取 `dockerImages` 配置对象，避免空值导致错误。
2. 多级镜像查找— 将原来的单级查找 `config.dockerImages[sel.hw] || "lmsysorg/sglang:dev"` 改为三级查找：
 - 首选 `di[${sel.hw}]${sel.quant}`（硬件和量化组合键）
 - 次选 `di[sel.hw]`（仅硬件键）
 - 兜底 `"lmsysorg/sglang:dev"`（默认镜像）

此逻辑与 `_deployment.jsx` 中的实现完全一致。

1. 仅修改 Playground 组件— 变更仅影响 `docs_new/src/snippets/_playground.jsx`，Deployment 组件已在 PR #29380 中更新。

`docs_new/src/snippets/_playground.jsx`

更新了 Docker 镜像选择逻辑，使 Playground 组件支持按 `hwlquant` 组合键解析镜像，与 Deployment 组件保持一致。

```
let cmd;
if (mode === "docker") {
  // 按 `hwlquant`（最具体）→ `hw` → 默认 `:dev` 依次解析镜像，与 _deployment.jsx 对齐。
```

```

const di = config.dockerImages || {};
const image = di[`${sel.hw}|${sel.quant}`] || di[sel.hw] || "lmsysorg/sglang:dev";
const portFlag = f.find((x) => x.split(/\s=/)[0] === "--port");
const servePort = portFlag ? portFlag.slice("--port".length).trim() : "{{PORT}}";
const dockerLines = [
  "docker run --gpus all",
  "  --shm-size 32g",
  (multinode || pdMode) ? "  --network host" : `  -p ${servePort}:${servePort}`,
  "  -v ~/.cache/huggingface:/root/.cache/huggingface",
  `  --env "HF_TOKEN={{HF_TOKEN}}"`,
  ...cellEnv.map((e) => `  --env ${e}`),
  "  --ipc=host",
  `  ${image}`,
  "  sglang serve",
  ...f.map((x) => "    " + x),
];
cmd = dockerLines.join(" \\n");
} else {
  // Python 模式保持不变
  const flagBlock = f.map((x) => "    " + x).join(" \\n");
  const envBlock = cellEnv.length ? cellEnv.join(" \\n") + " \\n" : "";
  cmd = `${envBlock}sglang serve \\n${flagBlock}`;
}

```

评论区精华

无实质性讨论。PR 获得 Fridge003 批准，无 review 评论。

风险与影响

- 风险: 极低。变更严格向下兼容，当配置中不存在 hwlquant 键时行为与旧代码完全一致。目前仅 GLM-5.2 使用了 hwlquant 键，因此对其他所有 cookbook 的渲染输出无影响。mint validate 验证通过。
- 影响: 提升文档准确性，确保 Playground 和 Deployment 两处展示一致的 Docker 命令。对运行时无影响。

关联脉络

- PR #29380: 本 PR 的前置工作，为 Deployment 组件添加了 hwlquant 镜像解析能力。
- 本 PR 是对其的补充修复，确保 Playground 组件也获得相同能力，保持两处文档生成逻辑的一致性。