PR #24876 完整报告

sgl-project/sglang

[Docs] Add MiniCPM-V 4.6 cookbook

合并时间：2026-05-11 12:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24876

执行摘要

该 PR 为 MiniCPM-V 4.6 多模态模型添加完整的 cookbook 使用指南，包含交互式部署命令生成器、模型特性介绍、安装步骤和高级用法。全部为文档新增，无运行时代码变更。

功能与动机

MiniCPM-V 4.6 是 OpenBMB 的新一代多模态模型，采用 Qwen3.5 风格的混合骨干网络和 NaViT 视觉编码器。PR body 说明这是 #24855 的文档预览，在模型权重公开前提供使用指南，帮助用户了解并部署模型。

实现拆解

1. 交互式组件：新建 docs_new/src/snippets/autoregressive/minicpm-v-4_6-deployment.jsx，提供硬件平台（A100/H100/H200/B200）、推理解析器、工具调用解析器和 Mamba 缓存等选项，动态生成 sglang serve 命令。组件支持暗黑模式、表单状态管理和剪贴板复制。
2. 主文档页面：新建 docs_new/cookbook/autoregressive/OpenBMB/MiniCPM-V-4_6.mdx，详细介绍模型架构（Gated Delta Net + 全注意力）、NaViT 视觉编码器、高分辨率切片、视频支持、推理模式和工具调用。包含 Docker 安装指南、命令生成器用法和性能调优建议。
3. 导航配置：修改 docs_new/docs.json，在导航树 InternVL 之后插入 OpenBMB 分组，指向新页面。
4. 首页索引：修改 docs_new/cookbook/autoregressive/intro.mdx，添加 OpenBMB 模型卡片，包含 logo 和链接。
5. 资源文件：新增 docs_new/cards/logos/openbmb.png 作为品牌标识。

以下为部署命令生成器组件的核心逻辑，展示了如何根据用户选择动态生成 sglang serve 命令，支持按硬件区分 attention backend 和参数调优：

```
// 摘自 minicpm-v-4_6-deployment.jsx

const options = {
  hardware: {
    name: 'hardware',
    title: 'Hardware Platform',
    items: [
      { id: 'a100', label: 'A100', default: false },
      { id: 'h100', label: 'H100', default: false },
```

```
      { id: 'h200', label: 'H200', default: true },
      { id: 'b200', label: 'B200', default: false },
    ],
  },
  // reasoning, toolcall, mambaCache 类似定义
};

const modelConfigs = {
  a100: { tp: 1, mem: 0.7 },
  h100: { tp: 1, mem: 0.7 },
  h200: { tp: 1, mem: 0.5 },
  b200: { tp: 1, mem: 0.4 },
};

const generateCommand = (values) => {
  const { hardware, reasoning, toolcall, mambaCache } = values;
  const hwConfig = modelConfigs[hardware];
  let cmd = `sglang serve --model-path openbmb/MiniCPM-V-4_6`;
  cmd += ` \
  --trust-remote-code`;
  cmd += ` \
  --dtype bfloat16`;
  if (hardware === 'b200') cmd += ` \
  --attention-backend trtllm_mha`;
  cmd += ` \
  --mem-fraction-static ${hwConfig.mem}`;
  if (reasoning === 'enabled') cmd += ` \
  --reasoning-parser qwen3`;
  if (toolcall === 'enabled') cmd += ` \
  --tool-call-parser qwen`;
  if (mambaCache === 'v2') cmd += ` \
  --mamba-scheduler-strategy extra_buffer`;
  cmd += ` \
  --host 0.0.0.0 --port 30000`;
  return cmd;
};
```

评论区精华

该 PR 没有 reviewer 评论，wisclmy0611 直接批准合并，未产生讨论。

风险与影响

- 风险：文档中的 tp 和 mem-fraction-static 参数基于待公开的模型参数量估计，实际部署可能需要调整。已在注释中标注“保守估计”并提示重调，风险较低。Docker 镜像标签可能随版本更新变化。
- 影响：对用户提供从安装到高级功能的完整指南，降低新模型使用门槛；对系统无影响；对团队完善文档生态。

关联脉络

该 PR 是 PR #24855（模型支持）的文档配套。在 #24855 合并后，本 PR 补全用户文档，形成从模型支持到使用指南的完整交付。后续可能随着模型权重公开发布更新参数和 license 信息。