

PR #35068 完整报告

sgl-project/sglang

[Docs] Feature MiniMax-H3 in the popular-models banner

合并时间: 2026-08-17 10:09

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35068>

执行摘要

该 PR 是纯文档变更: 将 MiniMax-H3 加入 docs 首页与 Cookbook 首页的 featured-models 轮播, 并移除 GLM-5.2 以保持四项配额。这是轮播中首次出现扩散 (diffusion) 模型, 标志 diffusion 文档在官网入口获得与自回归模型同等的展示位。改动仅一个文件, 文案均以 MiniMax-H3 cookbook 页面为事实来源, 无运行时代码影响。

功能与动机

PR body 明确说明当前轮播只展示自回归 LLM, 而 MiniMax-H3 已有完整 cookbook 页面与厂商卡片, 却没有 banner 位:

The featured-models carousel on the docs landing page (hero banner) and the Cookbook home (compact strip) currently lists only autoregressive LLMs. MiniMax-H3 has a complete cookbook page and vendor card, but no banner slot — this adds one so the diffusion side of the cookbook is represented in the rotation.

本次改动因此有三个目标:

1. 让扩散文档在首页获得可见性;
2. 保持列表规模为四项 (移除 GLM-5.2, 但不删除其 cookbook 页面与导航);
3. 所有文案严格限定为链接页面可验证的事实 (平台数、精度、生成时长、任务类型等)。

实现拆解

1. 定位数据源: 唯一改动文件为 docs/src/snippets/configs/popular-models.jsx, 其中 export const popularModels 数组是被 <PopularModels> 组件同时消费的单一数据源, 分别驱动 docs 首页 hero 轮播与 Cookbook 首页紧凑条。
2. 新增 MiniMax-H3 条目: 在数组第二位插入完整对象, 含 name、vendor、href、logo、badge、tags 与 hero 块。href 指向已存在的 diffusion cookbook 页面, logo 复用 /cards/logos/minimax.png, 避免新增图片资源。
3. 移除 GLM-5.2 条目: 删除 GLM-5.2 对象, 使列表保持四个模型; PR body 明确说明其 cookbook 页面与导航入口不受影响, 只是不再出现在轮播位。
4. 文案约束: 文件头部注释规定了事实来源——平台数与精度取自对应模型配置的 supportedHardware 与 quantizations, blurb 复述页面开头信息。MiniMax-H3 条目遵循该规则: 7 platforms 与 BF16 / FP8 来自 configs/MiniMaxAI/minimax-h3.jsx, 4–15 s

at 24 fps 对应页面 §4 与 §8, blurb 概括了 t2va / fl2va / ref2va 三种任务、Ulysses xRing 序列并行与硬件范围。

5. 验证：运行 `node docs/scripts/check_cookbook_configs.mjs` 与 `mint validate` 均通过；纯文档改动，无单元测试配套。

关键源码片段

`docs/src/snippets/configs/popular-models.jsx`

本次唯一改动文件，是 docs 首页 hero 轮播与 Cookbook 首页紧凑条共用的数据源，直接决定两个页面展示模型列表。

```
// popularModels 是 docs 首页与 Cookbook 首页轮播的单一数据源，
// <PopularModels> 在页面渲染时遍历该数组，新增条目会同时出现在两个页面。
// 下面是本次新增的 MiniMax-H3 条目（其余条目省略）。
{
  name: "MiniMax-H3", // 模型名，紧凑条（compact strip）展示
  vendor: "MiniMax", // 厂商名
  // 指向已存在的 diffusion cookbook 页面：/cookbook/diffusion/MiniMax/MiniMax-H3
  href: "/cookbook/diffusion/MiniMax/MiniMax-H3",
  logo: "/cards/logos/minimax.png", // 复用已有 logo，不新增图片资源
  badge: "New",
  // 平台数与精度来自 minimax-h3.jsx 配置的 supportedHardware 与 quant overlay
  tags: ["7 platforms", "Video + audio", "BF16 / FP8"],
  hero: {
    eyebrow: "Featured model • New",
    headline: "Meet MiniMax-H3 on SGLang",
    // blurb 复述 cookbook 页面开头：视频 + 同步立体声音轨、三种任务、
    // Ulysses x Ring 序列并行与硬件范围
    blurb:
      "MiniMax's video-and-audio diffusion model — one request returns an MP4 carrying 24 fps video and a synchronized stereo audio track. SGLang Diffusion serves all three task profiles — text, first/last frame, and image / video / audio reference conditioning — with Ulysses x Ring sequence parallelism and recipes across B200, B300, H200, H100, AMD MI300X / MI355X, and 2x RTX 5090.",
    tags: ["Video + synced audio", "4–15 s at 24 fps", "8x B200 → 2x RTX 5090"],
    cta: "Open the MiniMax-H3 cookbook",
    caption: "MiniMax-H3 deployment guide",
  },
}
```

评论区精华

该 PR 没有产生实质 review 评论线程。唯一审核者 [wisclmy0611](#) 直接 APPROVED，未留下文字意见，说明这次文档内容变更在评审环节没有争议。PR body 中关于平台数、量化精度、时长等描述的准确性论证是主要的“自我审查”，例如对照 `supportedHardware` 列出 B200 / B300 / H200 / H100 / MI300X / MI355X / RTX 5090，并说明 `BF16 / FP8` 的出处。

风险与影响

- 共享数据契约变更：popular-models.jsx 同时驱动 docs 首页 hero 轮播与 Cookbook 首页紧凑条，任一字段错误都会影响两处展示；本次改动属于数组结构的内容调整，无运行时逻辑，但后续维护时需同步更新两份页面。
- 文案过期风险：轮播文案是静态快照，若 MiniMax-H3 页面后续更新平台数、精度或性能数据，而本文件未同步，可能出现不一致。
- 可见度影响：移除 GLM-5.2 的 banner 位会降低该模型在官网首页的曝光，虽然其 cookbook 页面与导航仍然存在；这是有意的营销 / 文档策略决策。
- 用户与团队影响：docs 用户会看到新的轮播内容与扩散模型入口；文档维护团队需要保持本文件与各模型配置页一致。无性能、安全、兼容性影响。

关联脉络

本次变更处在 diffusion 文档与功能协同演进的脉络中。近期多个 PR 持续增强 SGLang 的扩散模型能力：[#34929](#) 为 LTX-2.3 开启 breakable CUDA graphs，[#34928](#) 为 Sana BCG 提速，[#34932](#) 加速 Cosmos3 T2I QKNorm+RoPE，[#35016](#) 收紧 diffusion 性能基线并恢复 Hunyuan3D 全管线覆盖。本文档 PR 是在入口侧让这些能力被更多用户看见的第一步，也响应了此前 diffusion 相关 cookbook 页面已就绪但缺少首页曝光位的问题。未来不排除更多 diffusion 模型进入轮播，或该列表演化为按任务类型自动推荐的配置结构。