

PR #29253 完整报告

sgl-project/sglang

Add MiMo V2.5 Blackwell vision FA4 recipe

合并时间: 2026-06-25 13:47

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29253>

执行摘要

为 MiMo V2.5 多模态模型在 Blackwell GPU 上添加 `--mm-attention-backend fa4` 参数配置，解决 FA3 后端不支持的问题，并通过文档明确告知用户。

功能与动机

MiMo V2.5 多模态检查点的视觉编码器内部请求 FlashAttention-3 后端，但 SGLang 在 Blackwell GPU 上拒绝 FA3 后端并抛出 `ValueError`。Pro 版本的部署食谱已使用 `--attention-backend fa4`，但 base 版本 (V2.5) 未覆盖多模态注意力后端。需要在 Blackwell 上为视觉编码器添加 `--mm-attention-backend fa4` 以使用受支持的 FA4 后端。

实现拆解

1. 修改 JSX 部署代码片段: 在 `docs_new/src/snippets/autoregressive/mimo-v25-deployment.jsx` 中，当模型为 MiMo V2.5 (非 Pro) 且硬件为 B200/GB300 时，向命令行标志数组中添加 `--mm-attention-backend fa4`。同时更新注释以反映这一变更。
2. 更新 MDX 文档: 在 `docs_new/cookbook/autoregressive/Xiaomi/MiMo-V2.5.mdx` 中添加一条说明，指导用户在 Blackwell 上部署 MiMo V2.5 时必须传递 `--mm-attention-backend fa4`，并解释原因。

`docs_new/src/snippets/autoregressive/mimo-v25-deployment.jsx`

核心变更文件，添加了 Blackwell 下 `--mm-attention-backend fa4` 标志的逻辑分支。

```
// ... 在 buildFlags 函数内的 Blackwell 分支中，针对非 Pro 模型添加视觉注意力后端
} else {
  // 当模型为 MiMo V2.5 (非 Pro) 且硬件为 Blackwell 时，
  // 必须传递 --mm-attention-backend fa4，因为视觉编码器需要 FA4 后端
  if (blackwell) flags.push(" --mm-attention-backend fa4");
  flags.push(" --mem-fraction-static 0.65");
  flags.push(" --chunked-prefill-size 16384");
}
// ... 其余标志
```

评论区精华

无实质讨论，仅有审核者批准。

风险与影响

风险极低：仅影响配置生成和文档，不影响核心运行时逻辑。条件判断已限定仅在 Blackwell 硬件的非 Pro 模型上添加参数。影响范围小，仅确保 MiMo V2.5 在 Blackwell GPU 上能正确启动。

关联脉络

- 与 #28347 类似，均涉及 Blackwell 上的注意力后端配置。
- 与 #29252 并行，后者调整了 Gemma4 的 Blackwell 内存比例。