PR #24052 完整报告

sgl-project/sglang

[Docs] quick fix delete --enable-dp-attention in sgl-jax

合并时间：2026-04-30 15:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24052

PR 分析报告：PR #24052

执行摘要

该 PR 修复了 MiMoV2.5-Pro TPU 部署文档中的一处错误：在 sgl-jax (TPU) 分支中，--enable-dp-attention 参数不存在（DP attention 默认启用），导致用户复制命令后启动服务器会因参数解析失败。变更仅移除该标志并添加注释说明。

功能与动机

PR body 明确指出："In sgl-jax, DP attention is enabled by default and the --enable-dp-attention flag does not exist. The MiMo-V2.5-Pro TPU launch command in the cookbook included this flag, which would cause the server to fail at argument parsing." 这是一次纯文档 bugfix。

实现拆解

1. 新增注释：在 docs_new/src/snippets/autoregressive/mimo-v25-deployment.jsx 的 sgl-jax 分支顶部添加多行注释，解释 sgl-jax 中 --tp-size 的含义、DP 实例 TP 的自动推导机制，以及 --enable-dp-attention 参数不存在的原因。
2. 移除无效参数：将 flags 构建中的 if (useDpAttn) flags.push(" --dp-size ${dpSize}", "--enable-dp-attention"); 改为 if (useDpAttn) flags.push(" --dp-size ${dpSize}");，仅删除 --enable-dp-attention，保留 --dp-size。CUDA (GPU) 分支保持不变。

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

唯一变更文件：移除 TPU 分支中的无效参数 --enable-dp-attention，并添加注释说明 sgl-jax 的 DP 约定。

```
// 在 generateCommand 函数中，sgl-jax (TPU) 分支部分（约第 229 行）
if (jax) {
  // sgl-jax 惯例：
  // - `--tp-size` 总是总 JAX 设备数；每个 DP 实例的 TP 由 tp/dp 自动推导。
  // - 没有 `--enable-dp-attention` 标志 —— DP attention 是默认行为
  //  （FFN 层自动选择 EP 拆分用于 MoE，attn-TP 拆分用于 dense）。
  const isV7x = hardware === "tpu-v7x";
  const useEp = expertParallelism === "enabled";
  const useDpAttn = dpAttention === "enabled";
  const dpSize = isV7x ? 4 : 8;
```

```
const flags = [];
flags.push(`  --model-path ${slug}`);
flags.push("  --trust-remote-code");
flags.push(`  --tp-size ${tp}`);
if (useEp) flags.push(`  --ep-size ${tp}`);
if (useDpAttn) flags.push(`  --dp-size ${dpSize}`); // 之前还包含 " --enable-dp-
attention"，现已移除
// … 其余标志
```

评论区精华

审核人 JustinTong0323 提出："then don't add --dp-size ? Without --enable-dp-attention the total chips needed is tp_size * dp_size"。PR 作者 JamesBrianD 解释 sgl-jax 中 tp-size 等于可用设备数，--dp-size 仍然需要用于指示 DP 并行度，而 --enable-dp-attention 参数不存在。最终达成一致，仅移除无效参数。

风险与影响

风险：极低。仅修改文档字符串，不涉及任何运行时逻辑。影响：仅影响阅读 MiMoV2.5-Pro TPU部署文档的用户，CUDA路径不受影响。合入后新文档将避免用户因参数错误而启动失败。

关联脉络

无直接关联历史 PR。这是一次独立的文档修正。