

PR #27668 完整报告

sgl-project/sglang

Fix MiMo-V2.5-Pro DP-attention dp size in cookbook deployment snippet

合并时间: 2026-06-10 22:08

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27668>

执行摘要

PR#27668 修复了 MiMo-V2.5-Pro 部署代码片段生成器中 DP-attention 大小计算错误的问题。原代码在 Pro 变体上错误地将 dp 设置为 tp，导致 `attn_tp` 不等于检查点要求的 8，模型加载失败。本次修改统一了 Pro 和 base 变体的处理逻辑，通过为 Pro Hopper 配置添加 `dp: 2` 并修正 `dpSize` 计算，确保所有组合下 `attn_tp == 8`。

功能与动机

MiMo-V2.5-Pro 的 FP8 检查点使用 TP=8 交错融合的 `qkv_proj`，因此模型加载器 `mimo_v2.py` 要求运行时每个 DP 组的 attention-TP 必须为 8。原代码在 `dpSize` 计算中错误地使用了 `tp`（如 `dp = tp = 16`），导致 `attn_tp = 1`，从而抛出异常：

```
MiMoV2 fused qkv_proj checkpoint is TP=8-interleaved; got attention tp_size=<n>
```

该问题已在 [#23808](#) 的模型作者确认：Pro 模型仅支持 TP8，对于 `tp=16` 的配方，`dp` 应设为 2。

实现拆解

1. 扩展 `HW_VARIANT_SPEC` 定义：为 Pro 的 Hopper 行（`prolh200`、`prolh100`，`tp=16`）添加 `dp: 2`，使得 `dp = tp / 8 = 2`；Blackwell 行（`prolb200`、`prolgb300`，`tp=8`）不设置 `dp` 字段，表示单一 attention 组、DP-attention off。
2. 统一 `computeConstraints` 中的 DP-attention 约束：将原本仅对 base 变体生效的 DP-attention 强制逻辑扩展到 Pro 变体。当 `spec.dp > 1` 时（如 Pro Hopper `dp=2`），强制启用 DP-attention；否则强制禁用。同时增加了非 JAX 条件检查，避免影响 TPU 路径。
3. 修正 `dpSize` 计算：将原来的条件表达式 `!isPro ? spec.dp : tp` 改为统一读取 `spec.dp`（若未定义则默认为 1），确保 Pro 变体不再错误地使用 `tp` 作为 `dp` 大小。这样生成的命令中 `--dp` 值正确，且与 [#23808](#) 验证的 `day0` 配方一致。
4. 更新注释：将原有的仅描述 base TP=4 约束的注释，更新为同时描述 Pro 的 TP=8 交错约束与 DP-attention 推导逻辑，帮助读者理解不同变体的配置规则。

[docs_new/src/snippets/autoregressive/mimo-v25-deployment.jsx](#)

唯一修改的文件，包含所有逻辑变更：`HW_VARIANT_SPEC`扩展、`computeConstraints`统一、`dpSize` 修正及注释更新。

```
// 硬件变体规格定义：每个条目包含模型 slug、tp、多节点标志等
// Pro 版检查点为 TP=8 交错融合，base 版为 TP=4 交错融合
```

```

// dp 字段表示 DP-attention 所需的 DP 度数 (tp / factor)
const HW_VARIANT_SPEC = {
  "prolh200": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 16, multinode: true, nnodes: 2, blackwell:
    false, jax: false, dp: 2 }, // tp/factor = 16/8 = 2
  "prolh100": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 16, multinode: true, nnodes: 2, blackwell:
    false, jax: false, dp: 2 },
  "prolb200": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 8, multinode: false, blackwell: true, jax:
    false }, // tp=8 == factor, 无需 dp
  "prolgb300": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 8, multinode: true, nnodes: 2, blackwell:
    true, jax: false },
  // ... 其他配置行
};

// 计算配置约束：根据当前变体和硬件强制某些选项的状态
const computeConstraints = (variant, hardware) => {
  const isPro = variant === "pro";
  const spec = HW_VARIANT_SPEC[`${variant}|${hardware}`];
  const blackwell = spec ? spec.blackwell : false;
  const jax = spec ? spec.jax : false;
  const c = {};
  // 对非 JAX 路径统一处理：两种检查点都是 TP 交错融合的
  if (spec && !jax) {
    const factor = isPro ? 8 : 4; // Pro 因子 8, base 因子 4
    if (spec.dp > 1) {
      // dp>1 时强制启用 DP-attention
      c.dpAttention = { force: "enabled", reason: `Checkpoint is TP=${factor}-interleaved; DP-
        attention is required (--dp = tp/${factor} = ${spec.dp}).` };
    } else {
      // 否则强制禁用（单一 attention 组）
      c.dpAttention = { force: "disabled", reason: `Single attention group on this hardware (tp=
        ${factor}, no dp-attention).` };
    }
  }
  // ... DeepEP / JAX 等其他约束
  return c;
};

// 在生成命令时：dpSize 直接从 spec 获取，替代原来的条件分支
const dpSize = spec?.dp ?? 1; // 使用回退值 1 防止 undefined 生成非法命令

```

评论区精华

review 中 [gemini-code-assist](#) 提出了两条改进建议：

关于 `spec.dp > 1` 的隐式转换：当 `spec.dp` 未定义时（如 Pro Blackwell 无 `dp` 字段），依赖隐式转换不够健壮。建议改为 `if (spec.dp && spec.dp > 1)`。关于状态过渡期 `dpSize` 可能为 `undefined`：在切换硬件的中间渲染阶段，`values.dpAttention` 可能暂为 `"enabled"` 而 `spec.dp` 尚未更新，导致 `dpSize` 为 `undefined`，生成 `--dp undefined`。建议使用回退值 1，如 `const dpSize = spec?.dp ?? 1`。

两条建议均未被采纳，但当前实现的实际行为与建议等价，风险较低。

风险与影响

风险：仅修改部署文档生成逻辑（.jsx 文件），未触及任何运行时代码，因此无回归风险。存在两个低风险点：

- 未来新增硬件配置行时若忘记添加 dp 字段，spec.dp 为 undefined，spec.dp > 1 返回 false，DP-attention 将被强制禁用。对于 tp=16 的 Hopper 配置，这将生成不带 --enable-dp-attention 的命令，导致 loader 拒绝启动。
- 在极端时序下（硬件切换瞬间），dpSize 可能短暂为 undefined，但 useEffect 会立即修正。

影响：

- 正面：MiMo-V2.5-Pro 用户使用 Hopper 双节点部署时，生成的命令从无效的 --tp 16 --dp 16 --enable-dp-attention 变为有效的 --tp 16 --dp 2 --enable-dp-attention，与模型作者确认的配方一致。
- 无影响：base 变体行为不变；Blackwell 和 TPU 路径不变。

关联脉络

该 PR 直接关联到 [#23808](#) 的 MiMo-V2.5 模型支持，其中模型作者确认了 Pro 模型的 TP8 限制和正确的 dp 大小。本次修复确保了文档生成器与模型实际加载约束一致，是 MiMo-V2.5 模型部署支持的重要补充。