PR #23945 完整报告

sgl-project/sglang

docs: enable MiMo V2.5 MTP cookbook path

合并时间：2026-04-29 01:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23945

执行摘要

本 PR 为 Xiaomi MiMo-V2.5 模型（base 变体）在官方 cookbook 中启用了 EAGLE MTP 规范解码支持。核心改动是移除了部署命令生成器中针对 base 变体的强制禁用逻辑，并修正了命令行参数名，同时更新了文档描述与基准测试结果。变更仅涉及前端代码片段和文档，无运行时风险，但 review 曾指出多模态 benchmark 数据无效，已在合并前替换。

功能与动机

在此之前，MiMo-V2.5 cookbook 仅允许 Pro 变体选择 EAGLE MTP 选项；base 变体（310B / 15B 活跃参数）虽然 checkpoint 包含 MTP 权重，但命令生成器通过 computeConstraints 强制禁用了该选项。用户只能手动推断所需参数。PR body 明确目标："Enable EAGLE MTP for MiMo-V2.5 in the cookbook command generator. Update the MiMo-V2.5 deployment notes to describe the checkpoint MTP path and the required Hopper flags."

实现拆解

1. 移除禁用约束（mimo-v25-deployment.jsx）：删除 if (!isPro) { c.eagleMtp = { force: "disabled" }; } 块，使 base 变体的 EAGLE 选项不再被固定。
2. 统一启用条件：useMtp 计算从 isPro && eagleMtp === "enabled" 改为 eagleMtp === "enabled"，允许 base 变体生效。
3. 修正 CLI 参数名：--speculative-algo → --speculative-algorithm，匹配框架实际参数。
4. 更新注释与 UI 文案：移除 "Pro only" 标注，subtitle 改为 "EAGLE"。
5. 同步文档（MiMo-V2.5.mdx）：重写 MTP 描述，明确 "Both variants support EAGLE speculative decoding with MTP weights"；补充 DP attention 配置说明；替换速度基准为 MTP 模式下的新数据。
6. 基准数据替换：使用 H200 8 GPU 实际运行 EAGLE MTP 的延迟和吞吐量结果替代旧数据。

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

部署命令生成器核心脚本。移除了 base 变体禁用 EAGLE MTP 的强制约束，修正了参数名，是本次功能启用的关键文件。

```
// 从 computeConstraints 中删除了针对非 Pro 的强制禁用块，
// 使 base 变体也可自由选择 EAGLE MTP。

const computeConstraints = (options, variant, hwKey) => {
```

```
    ...
    if (!isPro) {
      // 之前此处有：c.eagleMtp = { force: "disabled", reason: "..." };
      // 现已删除，因此 base 变体的 EAGLE 选项不再被固定。
      if (spec && spec.dp > 1) {
        c.dpAttention = { force: "enabled", reason: "DP attention required for dp>1" };
      } else {
        c.dpAttention = { force: "disabled", reason: "DP attention not needed for dp=1" };
      }
    }
    ...
  };

  // 生成命令时，useMtp 不再要求 isPro，base 变体也可为 true
  const useMtp = eagleMtp === "enabled"; // 之前是 isPro && eagleMtp === "enabled"

  // 同时修正了参数名
  if (useMtp) {
    flags.push(" --speculative-algorithm EAGLE"); // 之前是 --speculative-algo
    flags.push(" --speculative-num-steps 3");
    flags.push(" --speculative-eagle-topk 1");
    flags.push(" --speculative-num-draft-tokens 4");
  }
```

评论区精华

- `--dp = TP / 4` 表达式可能被误解：reviewer 建议不要在文档中使用带等号和空格的 literal 写法，以避免用户复制时困惑。该评论未被回复，但 PR 仍被批准，推测团队认为不影响使用。
- 多模态 benchmark 数据异常：第 5.3.3 节的图像评测中多项指标为 0，reviewer 指出数据无效。作者已在后续提交中替换为有效数据，该问题已解决。

风险与影响

- 低风险：仅修改文档和前端片段，不影响运行时行为。主要潜在风险是用户遵循 cookbook 配置时若硬件不支持 MTP（如非 Hopper）可能启动失败，但组件本身有硬件校验，且注释已说明依赖 Hopper 架构。
- 影响：base 变体用户现在可一键生成 MTP 命令，减少了配置门槛。团队维护成本下降，文档一致性提升。

关联脉络

本 PR 是此前 #23808 "Xiaomi MiMo-V2.5-Pro day0 support" 的后续完善，将相同的能力通过文档带给 base 变体。结合近期其他文档 PR（如 #23943 DeepSeek-V4 低延迟方案），可见团队在重点关注 cookbook 的质量和覆盖度。