PR #25152 完整报告

sgl-project/sglang

docs: prepend SGLANG_JIT_DEEPGEMM_PRECOMPILE=0 for H200 FP8 Flash max-throughput

合并时间：2026-05-13 15:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25152

1. 执行摘要

本次 PR 在 DeepSeek-V4 cookbook 交互式命令生成器中，为 H200 FP8 + Flash (small) + max-throughput 配置添加 SGLANG_JIT_DEEPGEMM_PRECOMPILE=0 环境变量，使用户直接获得跳过 DeepGEMM JIT 预编译的启动命令。仅 3 行新增代码，范围精确可控，无风险。

2. 功能与动机

PR body 明确说明：在 H200 FP8 Flash (small) 的 max-throughput 部署场景中，DeepGEMM 的 JIT 预编译不是必经步骤，跳过它可以减少用户启动服务的等待时间。原有生成命令未包含该环境变量，用户需要手动添加，现在由生成器自动输出。

3. 实现拆解

仅修改 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 一个文件：

1. 进入 max-throughput recipe 分支。
2. 在 hardware === "h200" 的分支内，新增 if (!isBig) 条件判断（isBig 表示 Pro 模型）。
3. 当满足条件时，向 recipeEnv 数组推入 SGLANG_JIT_DEEPGEMM_PRECOMPILE=0。

改动前后对比清晰：原来 H200 分支只设置 SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK，现在对 Flash 模型额外设置 DeepGEMM 跳过标志。

关键源码片段

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

唯一变更文件；在 DeepSeek-V4 部署命令生成的 React 组件中，为 H200 FP8 Flash max-throughput 组合添加环境变量。

```
} else if (recipe === "max-throughput") {
  if (hardware === "h200") {
    if (!isBig) {
      // Flash (small) 场景不需要 DeepGEMM JIT 预编译，跳过以加速启动
      recipeEnv.push("SGLANG_JIT_DEEPGEMM_PRECOMPILE=0");
    }
    recipeEnv.push(isBig
      ? "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=128"
      : "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=256");
```

```
} else if (isBig && hardware === "b200") {
  // B200/B300 Pro 的精度验证环境变量（不变）
  // ...（省略已有逻辑）
}
```

5. 评论区精华

该 PR 没有实质性 review 讨论。gemini-code-assist[bot] 自动评论表示无反馈，wisclmy0611 直接批准。

6. 风险与影响

无风险。3 行新增代码，仅影响 JSX 中命令生成逻辑，不改动任何运行时代码。

- 影响范围：仅 DeepSeek-V4 cookbook 页面上选择 H200 + Flash (small) + max-throughput 的用户，会看到命令中自动包含 SGLANG_JIT_DEEPGEMM_PRECOMPILE=0。
- 无回归风险：其他硬件、模型大小、recipe 组合均未受影响。

7. 关联脉络

本 PR 是对 DeepSeek-V4 cookbook 文档的持续优化，与 PR#25115、#25134 属于同一文档生成器维护线。此前已有 B200 等硬件跳过 DeepGEMM 预编译的先例，本 PR 补充了 H200 Flash 场景的遗漏。