PR #25412 完整报告

sgl-project/sglang

[Doc] DSV4 cookbook: clean up env vars, add MegaMoE toggle, unify docker image

合并时间：2026-05-17 02:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25412

执行摘要

此 PR 清理了 DeepSeek-V4 部署 cookbook 中已过时的环境变量，并新增 MegaMoE 精度切换选项，简化了用户配置。Docker 镜像统一部分已回退给专用 PR。变更后部署命令更简洁，同时支持新后端。

功能与动机

DeepSeek-V4 部署文档中的环境变量逐渐过时（已默认启用或不再需要），且新增的 MegaMoE 后端需要在前端提供切换选项。同时希望统一 Docker 镜像以减少维护成本。这些改进便于用户正确生成部署命令并体验新功能。

实现拆解

1. 新增 MegaMoE 配置：在命令生成器配置对象 options 中新增 megamoe 选项，提供 Disabled、W4A8、W4A4 三种选择。
2. 清理过时环境变量：从 low-latency、balanced、max-throughput 等多个部署模式中移除不再需要的环境变量（如 SGLANG_JIT_DEEPGEMM_PRECOMPILE、SGLANG_OPT_SWA_SPLIT_LEAF_ON_INSERT 等），这些变量已默认启用或已删除。
3. 更新命令生成逻辑：在 generateCommand 函数中解构参数时新增 megamoe，为后续根据选择生成对应参数做准备。
4. 回退 Docker 镜像变更：因 reviewer 指出另一 PR #25410 已处理，回退本 PR 中的 Docker 镜像统一修改。

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

唯一变更文件，包含 MegaMoE 选项和环境变量清理。

```
// 在 options 对象中新增 MegaMoE 切换配置（位于 HiCache 之后）
megamoe: {
  name: "megamoe",
  title: "MegaMoE",
  items: [
    { id: "disabled", label: "Disabled", default: true },
    { id: "w4a8", label: "W4A8", default: false },
    { id: "w4a4", label: "W4A4", default: false, subtitle: "FP4 acts" },
  ],
},
```

```
// generateCommand 函数解构增加 megamoe
const generateCommand = () => {
  const { hardware: rawHardware, modelSize, recipe, reasoningParser, toolcall, hicache,
  megamoe } = values;
  // 此后会根据 megamoe 的值决定是否追加 --moe-runner-backend megamoe 等参数
  // 同时删除了之前大量冗余的 env vars，例如 SGLANG_JIT_DEEPGEMM_PRECOMPILE 等
  // B200/B300 Pro 分支现在不再需要任何额外 env vars
}
```

评论区精华

Reviewer Fridge003 在评论中指出 Docker 镜像统一变更应交给专用 PR #25410 处理，建议移除。作者采纳并及时回退了相关改动，使 PR 聚焦于核心清理和选项添加。

风险与影响

风险较低。被移除的环境变量均以默认启用或已废弃，不会影响现有用户。MegaMoE 选项依赖 #25406，文档已注明依赖关系，在 #25406 合并前不可用。影响限于 DeepSeek-V4 部署文档读者，命令生成将更简洁。

关联脉络

此 PR 是 DeepSeek-V4 部署工具链持续演进的一部分。其依赖的 #25406 将 MegaMoE 解耦为独立后端，是本次选项添加的技术基础。Docker 镜像统一工作由 #25410 单独推进，体现了团队并行协作的开发模式。