

PR #36828 完整报告

sgl-project/sglang

[AMD] Update v4 amd cookbook 0828

合并时间: 2026-08-29 08:20

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36828>

执行摘要

- 一句话: 更新 AMD DeepSeek-V4 cookbook 镜像与 MoRI EP 配置
- 推荐动作: 该 PR 值得快速浏览, 因为它体现了文档如何随功能演进同步更新, 特别是为 AMD 用户提供了 MoRI 后端的实用配置。无需精读代码, 但值得关注新增环境变量的实际效果。

功能与动机

随着 DeepSeek-V4 在 AMD 平台上的支持和 MoRI 后端的完善, 现有 cookbook 文档需要同步更新。PR body 明确指出更新 mori-ep 部分, 引入两个新环境变量 (关联 PR#36130 和 PR#36119), 并提升镜像版本以匹配最新的每日构建。

实现拆解

1. 首先在 docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx 中更新了 dockerImages 配置, 将 MI300X 和 MI355X 的镜像 tag 从 v0.5.17-rocm720-*-20260822 更新为 v0.5.18-rocm720-*-20260828, 确保用户拉取到最新每日构建。
2. 在 docs/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 中, 同步更新了文档中的镜像版本示例和 docker pull 命令, 保持与配置片段一致。
3. 在 DeepSeek-V4.mdx 中新增了 'MoRI EP (AMD expert parallelism)' 小节, 说明 AMD 上专家并行使用 MoRI all-to-all 后端, 并给出推荐配置命令: `--moe-a2a-backend mori`, 同时解释了两个可选环境变量的用法 (FP4 需同时启用两个变量, FP8 仅需 `SGLANG_MORI_RECV_BOUND=1`)。
4. 提交历史显示进行了多次迭代: 先更新 EP 部分, 再分两次调整镜像版本以修正到正确版本, 最后将命令中的模型路径改为 {MODEL_PATH} 占位符。

关键文件:

- docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx (模块 配置片段; 类别 source; 类型 configuration): 核心配置片段, 更新了 AMD 镜像版本 tag, 是用户获取最新镜像的关键入口。
- docs/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx (模块 用户文档; 类别 other; 类型 documentation-update): 主要文档变更, 包含镜像版本更新和新增 MoRI EP 配置说明。

关键符号：未识别

关键源码片段

[docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx](#)

核心配置片段，更新了 AMD 镜像版本 tag，是用户获取最新镜像的关键入口。

```
// docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx (片段)
dockerImages: {
  // ... 其他平台使用 latest
  // AMD 每日更新的 sglang-rocm 镜像，每次重新验证时需更新日期 tag
  mi300x: "lmsysorg/sglang-rocm:v0.5.18-rocm720-mi30x-20260828",
  mi355x: "lmsysorg/sglang-rocm:v0.5.18-rocm720-mi35x-20260828",
},
```

[docs/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx](#)

主要文档变更，包含镜像版本更新和新增 MoRI EP 配置说明。

```
# 从 DeepSeek-V4.mdx 的 MoRI EP 小节摘录
# 在 AMD 上启用 MoRI EP，推荐配置：
export SGLANG_MORI_DISPATCH_DTYPE=mxfp8 # 可选，FP8 类型
```

```
export SGLANG_MORI_RECV_BOUND=1 # 可选，用于接收端限流
sglang serve \
  --model-path {MODEL_PATH} \
  --tp 8 --dp 8 --enable-dp-attention \
  --ep-size 8 --moe-a2a-backend mori --deepseek-mode normal \
  <other args from the Deploy panel>
```

```
# FP4 场景建议同时启用两个变量；FP8 只需 SGLANG_MORI_RECV_BOUND=1
```

评论区精华

无 review 评论或讨论。仅有一位审批者 HaiShaw 直接批准，未提出任何修改意见。

- 暂无高价值评论线程

风险与影响

- 风险：本 PR 为纯文档变更，不涉及代码逻辑，风险极低。但需注意：镜像版本 tag 的更新依赖 sglang-rocm:v0.5.18-rocm720-*-20260828 在 Docker Hub 上已实际发布，若镜像不存在，用户按文档操作会失败；建议确认镜像已发布。另外，新增的 MoRI 环境变量配置是否正确需要与相关代码 PR（#36130、#36119）保持一致。
- 影响：影响范围限于 AMD 用户部署 DeepSeek-V4 时的文档指导。更新后用户可遵循新配置启用 MoRI EP 以获得更好的专家并行性能。对现有功能无影响，团队维护成本低。
- 风险标记：文档与镜像版本一致性依赖，无测试覆盖

关联脉络

- PR #36130 : PR body 中提及, 为 SGLANG_MORI_RECV_BOUND=1 环境变量引入的关联 PR。
- PR #36119 : PR body 中提及, 为 SGLANG_MORI_DISPATCH_DTYPE=mxfp8 环境变量引入的关联 PR。