PR #28471 完整报告

sgl-project/sglang

docs(cookbook): add AMD MI300X/MI325X/MI355X support for GLM-5.2

合并时间：2026-07-01 05:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28471

执行摘要

本 PR 为 GLM-5.2 Cookbook 新增 AMD MI300X / MI325X / MI355X 支持，基于已验证的 GLM-5.1 / DeepSeek-V3.2 架构的 ROCm 配方迁移，仅需替换模型路径。核心变更包括部署面板新增加硬件选项、指定 ROCm Docker 镜像、在 Playground 中禁用未验证的 MTP 和 Context Parallel 功能，并添加配置说明和基准测试数据。

功能与动机

扩大 GLM-5.2 的硬件覆盖范围，使 AMD 用户也能在 MI300X / MI325X / MI355X 上部署该模型。GLM-5.2 与 GLM-5.1 / DeepSeek-V3.2 共享 glm_moe_dsa 架构，因此验证过的 ROCm 配方可直接携带，仅需更换模型路径。

实现拆解

1. 部署面板配置 (glm-5.2.jsx)：在 supportedHardware 中添加 mi355x, mi325x, mi300x；添加对应硬件专用 ROCm Docker 镜像；在 Playground 的 Attention 卡片中将 AMD 硬件加入 CP 的禁用列表并更新禁用原因；在 Speculative 卡片中将 AMD 硬件加入 MTP 选项的禁用列表并添加禁用原因；在 cells 数组中新增 AMD 硬件的配置单元格（low-latency / balanced / high-throughput），使用 DSA tilelang 后端。
2. 基准数据 (glm-5.2-benchmarks.jsx)：新增 MI355X + FP8 在三种策略下的基准测试数据（TTFT、TPOT、吞吐量）。
3. 用户文档 (GLM-5.2.mdx)：修改 description 以包含 AMD GPU；新增 "AMD GPUs (MI300X / MI325X / MI355X)" 配置提示；添加 <Warning> 组件说明 gfx950 block-FP8 精度问题（已修复，但用于提醒用户使用指定镜像版本）。

docs_new/src/snippets/configs/zai-org/glm-5.2.jsx

核心配置文件，添加 AMD 硬件支持、Docker 镜像、禁用逻辑和配置单元格

```
// Playground 中 "Speculative Decoding" 卡片的配置
// 展示如何为特定硬件（AMD）禁用选项，并通过 disableReason 向用户解释
speculative: {
  options: [
    { id: "current", label: "Inherited from base" },
    { id: "off", label: "Off (greedy)" },
    // MTP 5-1-6 选项：在 NVIDIA 上启用，在 AMD 上禁用（原因说明）
    { id: "mtp-516", label: "EAGLE / MTP 5-1-6 (low-latency)",
      flags: ["--speculative-algorithm EAGLE", "--speculative-num-steps 5",
```

```
            "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 6"],
      disable: { hw: ["mi355x", "mi325x", "mi300x"] },
      disableReason: "MTP/EAGLE speculative decoding is not yet validated on AMD ROCm
      (MI300X/MI325X/MI355X): the gfx950 spec-decode draft kernel is not yet validated and at -
      -speculative-num-steps > 3 hits a separate build issue; the DSA nextn draft path is CUDA-
      only." },
    // MTP 1-1-2 选项：类似禁用逻辑
    { id: "mtp-112", label: "EAGLE / MTP 1-1-2 (balanced)",
      flags: ["--speculative-algorithm EAGLE", "--speculative-num-steps 1",
            "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 2"],
      disable: { hw: ["mi355x", "mi325x", "mi300x"] },
      disableReason: "MTP/EAGLE speculative decoding is not yet validated on AMD ROCm
      (MI300X/MI325X/MI355X): the gfx950 spec-decode draft kernel is not yet validated and at -
      -speculative-num-steps > 3 hits a separate build issue; the DSA nextn draft path is CUDA-
      only." },
  ],
},
```

评论区精华

> 审核者 zijiexia 要求禁用 AMD 的 MTP 选项，因为 AMD 暂不支持 MTP。作者响应并添加了 disable 逻辑和原因。 审核者询问 BF16 是否适合单节点 MI300X，作者确认内存不足后移除了对应配置。 作者发现 gfx950 block-FP8 精度 bug，并提交上游修复 PR #28685，文档固定正确镜像版本。

风险与影响

- 精度风险：如果用户使用早于 2026-06-18 的 ROCm 镜像，gfx950 上的 block-FP8 输出不可信。文档已固定正确镜像版本，但仍需关注。
- 兼容性风险：AMD 上的 MTP 和 CP 功能被禁用，如果用户强制启用可能失败。配置通过 disable 机制阻止，但后续版本更新需重新验证。
- 依赖风险：依赖特定 ROCm 镜像和内核版本，AMD 驱动更新可能导致配置失效。
- 用户影响：AMD 用户现在有明确的部署指南，可快速部署 GLM-5.2。

关联脉络

- 与 #28423 (DeepSeek-V4 AMD cookbook) 共享相同的迁移策略。
- 与 #28685 (gfx950 精度 bug) 形成 bug 发现 - 修复 - 文档警告的完整链路。
- 后续的 MTP 支持依赖于 #29373 等修复，体现了持续迭代的特征。