PR #36608 完整报告

sgl-project/sglang

[AMD] Add GLM-5.3-Flash recipes for MI300X, MI325X, and MI355X

合并时间：2026-08-27 13:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36608

执行摘要

本 PR 为 GLM-5.3-Flash cookbook 新增 MI300X、MI325X、MI355X 三个 AMD ROCm 平台的部署配方，同时把未在 AMD 验证的 MTP 投机解码、FP8 + TRT-LLM DSA 等组合在配置面板中显式禁用。文档记录了 MI300X 与 MI355X 的 GSM8K 准确度验证结果（97.35% / 97.65%），并刻意不给 MI325X 标注任何数据。纯文档与配置变更，无运行时代码改动，但依赖未合入的引擎 PR #36607。

功能与动机

PR body 明确动机：为 zai-org/GLM-5.3-Flash 提供可复制粘贴的 AMD ROCm 部署配方，与现有 NVIDIA 配方并列。AMD 命令依赖 #36607 的引擎改动（该 PR stacked on 模型支持 PR #36507）。作者明确表示只做 accuracy/correctness 验证，不发布吞吐 / 延迟声明；MI325X 虽然与 MI300X 同为 gfx942，但没有直接测量，因此保持 unverified 标注。

实现拆解

1. 扩展硬件矩阵：docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx 的 supportedHardware 新增 mi300x、mi325x、mi355x；isRecommendedSelection 将三者与 h100、h200 一样映射到 bf16-tilelang 推荐配对。
2. 禁用未验证组合：同一文件中，对 AMD 禁用 low-latency（Adaptive MTP）、fp8-trtllm（FP8 KV + TRT-LLM DSA）与 deep_gemm MoE runner；新增 ROCm 7.2 镜像映射与 TP8 非投机操作点命令。
3. 记录准确度基准：docs/src/snippets/configs/zai-org/glm-5.3-flash-benchmarks.jsx 新增 MI300X（97.35%）与 MI355X（97.65%）两个 accuracy-only 条目，记录模型修订、引擎 head、source manifest SHA256、镜像与请求统计；MI325X 仅有占位匹配行。
4. 更新文档说明：docs/cookbook/autoregressive/GLM/GLM-5.3-Flash.mdx 更新 description、安装指引、策略说明与精度表格，注明 AMD 仅提供 High Throughput 配方且依赖 PR #36607。
5. 配套验证：通过 check_cookbook_configs.mjs、pre-commit、mintlify broken-links 与 mintlify validate；无测试文件改动，CI 的 AMD ROCm 7.2 作业未在本次 commit 上运行。

docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx

本次变更的核心配置文件：扩展硬件支持矩阵，为 AMD 禁用未验证的 MTP、FP8 + TRT-LLM 与 deep_gemm 组合，并新增 ROCm 7.2 镜像与 TP8 操作点命令，所有 AMD 配方逻辑都集中在此。

```
// 以下为整理后的 AMD 门控逻辑片段；原文件中的这些判断以内联数组形式出现
const AMD_HW = ["mi300x", "mi325x", "mi355x"];

// 1) low-latency（Adaptive MTP 投机解码）在 AMD 上未验证，直接禁用，
// 引导用户使用非投机的高吞吐配方
{
  id: "low-latency",
  label: "Low Latency",
  disabled: (s) => AMD_HW.includes(s.hw),
  disableReason:
    "MTP speculative decoding has not been validated for GLM-5.3-Flash on AMD ROCm; " +
    "use the non-speculative High Throughput recipe.",
},

// 2) 推荐配对随硬件切换：AMD 与 Hopper 一样使用 BF16 KV + TileLang DSA
const pairing = ["h100", "h200", ...AMD_HW].includes(s.hw)
  ? "bf16-tilelang"
  : "fp8-trtllm";

// 3) FP8 KV + TRT-LLM DSA 在 AMD 上同样禁用，避免未经验证的组合
{
  id: "fp8-trtllm",
  label: "FP8 + TRT-LLM",
  disabled: (s) => ["h100", "h200", ...AMD_HW].includes(s.hw),
  disableReason: "This recipe uses BF16 KV cache with TileLang DSA on Hopper and AMD
  ROCm GPUs.",
},

// 4) AMD 使用专门构建的 ROCm 7.2 镜像；引擎改动未进镜像时需挂载 PR #36607 源码。
// mi325x 与 mi300x 共用 gfx942 镜像，但未直接验证
const dockerImages = {
  mi300x: "lmsysorg/sglang:v0.5.18-rocm720-mi30x",
  mi325x: "lmsysorg/sglang:v0.5.18-rocm720-mi30x",
  mi355x: "lmsysorg/sglang:v0.5.18-rocm720-mi35x",
};
```

docs/src/snippets/configs/zai-org/glm-5.3-flash-benchmarks.jsx

新增 MI300X 与 MI355X 的 accuracy-only 基准记录，并刻意让 MI325X 只保留占位匹配行，体现验证与推断的严格区分。

```
// 新增 AMD 记录只包含准确度数据，刻意不写吞吐或延迟，避免无基准支撑的性能声明
{
  match: { hw: "mi300x", strategy: "high-throughput" },
  sglang_version: "9e692c9216",
  accuracy: { gsm8k_pct: 97.35 },
  notes:
    "Accuracy-only validation on 8x MI300X (gfx942, TP8) with model revision " +
    "3f1971b7b5f7a528c9c4ef6212c8785298a8c24a, PR #36607 head 9e692c9216, " +
    "image lmsysorg/sglang:v0.5.18-rocm720-mi30x. Full GSM8K 1,284/1,319, " +
```

```
      "100% stop rate, zero errors. No throughput or latency benchmark was run.",
  },
  // MI325X 只有占位匹配行：共享 gfx942 但未直接测量，因此不附任何数据
  { match: { hw: "mi325x", strategy: "high-throughput" } },
```

评论区精华

本 PR 没有任何 review 评论，审核者 zijiexia 直接 APPROVED。公开讨论集中在 PR body 的验证纪律上：不发布未测量的性能数字、MI325X 显式标注为未验证、未验证的 MTP/FP8 组合直接在面板中禁用。

风险与影响

- 配方依赖未合入的 PR #36607，用户在公共镜像上直接复制命令可能因引擎缺失而失败；文档已注明需挂载 PR 源码分支，但仍存在误导风险。
- MI325X 与 MI300X 同为 gfx942 但未直接测量，显存带宽与容量差异可能导致实际 KV 池与 batch 行为不同，文档以占位条目处理。
- 配置面板强制禁用 AMD 上的 MTP 与 FP8 + TRT-LLM，可能阻断社区在 AMD 上的自发验证，属于有意的保守选择。
- CI 未覆盖该 commit 的 AMD ROCm 7.2 作业，验证依赖人工执行，存在回归盲区。
- 影响范围：仅文档与 cookbook 面板配置，无运行时风险；对 AMD 用户可开箱获取验证过的部署命令，NVIDIA 用户不受影响。

关联脉络

- PR #36607：本 PR 依赖的 ROCm 引擎改动，配方命令只有在它合入后才能直接运行。
- PR #36507：GLM-5.3-Flash 模型支持 PR，是 #36607 的 stack 基础。
- PR #36660：同为 GLM-5.3-Flash cookbook 的修正 PR（修正 speculative flag 并记录 NVIDIA GSM8K），与本次 AMD 条目构成同一模型的文档完整度迭代。
- 近期 AMD 系列 PR（如 #35611、#36636、#36543）显示项目在系统化地为 AMD 平台补齐验证、CI 门控与基准记录，本 PR 是这一趋势在文档侧的落地。