

PR #39406 完整报告

sgl-project/sglang

[AMD] GLM-5.2 MI355X MXFP4: bump image to 20260913, enable TOPK_V2

合并时间: 2026-09-14 21:50

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39406>

[AMD] GLM-5.2 MI355X MXFP4: 镜像升至 20260913 并启用 TOPK_V2

执行摘要

本 PR 是一次纯文档 / 配方同步: 把 GLM-5.2 MI355X MXFP4 的 Playground 配置与 cookbook 说明对齐到 SemiAnalysisAI/InferenceX#3081, 改动只有 2 个文件、+7/-7。核心动作有两个——`docs/src/snippets/configs/zai-org/glm-5.2.jsx` 中 `dockerImages["mi355xlmxfp4"]` 由 `v0.5.19-rocm720-mi35x-20260910` 升到 `-20260913`, 以及四个 MXFP4 cell 的 `env` 注入 `SGLANG_OPT_USE_TOPK_V2=true`; MDX 侧同步了 MTP 验证说明。运行时、内核、测试零改动, 风险不在回归而在文档准确性: 镜像与 flag 之间是隐式耦合, 而该 PR 没有 AMD 实机验证背书。

功能与动机

PR body 写明: 「This PR syncs the GLM-5.2 MI355X MXFP4 cookbook recipes with SemiAnalysisAI/InferenceX#3081」。关联 Issue #3081 给出两条动机:

- 更新上游 SGLang ROCm 镜像, 从 `lmsysorg/sglang-rocm:v0.5.19-rocm720-mi35x-20260910` 到 `-20260913`。
- 为 ROCm 上的 GLM-5.x 启用 `SGLANG_OPT_USE_TOPK_V2=true`, 因为「上游 PR `sgl-project/sglang#36684` 和 `#36851` 已使 v2 fused top-k kernel 可用, 并为该模型系列在 ROCm 上启用此路径」。

换句话说, 这不是修 bug, 而是「上游内核能力具备后, 配方跟进去打开开关」, 且刻意把 InferenceX 侧 harness 语义 (并发点集重分配) 排除在外。维护者 sogalin 在合并前确认了收益: 「LGTM, `SGLANG_OPT_USE_TOPK_V2=1` can help improve the performance for GLM.」

实现拆解

1. 变更入口与范围界定

落地文件只有两个: `docs/src/snippets/configs/zai-org/glm-5.2.jsx` (+6/-6, 核心) 与 `docs/cookbook/autoregressive/GLM/GLM-5.2.mdx` (+1/-1, 同步说明)。`python/sglang` 下的运行时、内核与测试目录完全没有被触碰, 因此本 PR 不改变任何实际执行行为, 只改变用户按文档生成启动命令时看到的镜像与环境变量。

2. 镜像分桶更新

`dockerImages` 以 `hw` 与 `hwIquant` 为 key 分桶，本次只动 "mi355xlmxfp4" 一项。边界的正确性由分桶本身保证：

桶位	值	本次是否变化
mi355x (默认 /FP8/BF16)	v0.5.13.post1-rocm720-mi35x-20260618	否
mi355xlmxfp4	v0.5.19-rocm720-mi35x-20260913	是 (原 -20260910)
mi325x / mi300x	v0.5.13.post1-rocm700-mi30x-20260616	否

这样 MDX 中 `gfx950 block-FP8` 精度修复说明所引用的 20260618 镜像保持不变，FP8/BF16 档位不受影响。

3. 环境变量注入：四个 MXFP4 cell

`env: []` → `env: ["SGLANG_OPT_USE_TOPK_V2=true"]`，覆盖 `strategy` 为 low-latency (TP8/EP1, 5-step MTP)、balanced (TP4)、high-throughput (TP4/EP4, 5-step MTP)、mtp-314 (TP4, 3-step MTP) 四个 cell。关键点是 `flags` 数组一字未改：TP/EP 拓扑、`--kv-cache-dtype fp8_e4m3`、Triton DSA 前后端、draft token 数与 `--mem-fraction-static` 全部原样保留，因此后续任何性能变化只能归因于 top-k 内核路径本身，而不是并行或显存参数。

4. 文档同步：MDX 的 MTP 验证说明

`GLM-5.2.mdx` 的“MTP / EAGLE speculative decoding on AMD”段落把固定镜像从 20260910 改为 20260913，并新增一句说明该镜像启用 `SGLANG_OPT_USE_TOPK_V2=true`，附 #36684 / #36851 链接。作用是避免 `snippet` 与 `cookbook` 两个文件对镜像 tag 与 `env` 的描述出现漂移。

5. 范围收敛：HiCache 双 arm 说明被自行回退

commit `4ea02a0` 曾为 MI355X MXFP4 加入 InferenceX 双 arm 的说明 (TP4/EP4 + HiCache DRAM offload 仅并发 ≥ 4；TP8/EP1 无 KV offloading 跑低并发 ≤ 2)，commit `17eca5b` 以 Revert 整体撤销。PR body 的“What is NOT changed”给出理由：并发点集重分配属于 benchmark harness 变更，不属于 cookbook serving 命令。最终 HiCache knob 维持「host-specific、故意省略、可在 Playground 的 Hierarchical KV Cache 卡片切换」的既有表述。

6. 测试与验证配套

没有新增或修改任何测试。验证项集中在 InferenceX 侧 (Docker Hub digest 存在性、`bash -n`、PyYAML 解析、矩阵生成核对)；本仓库 CI 读数为“Latest PR Test (Base)”通过、“PR Test (Extra)”失败、“AMD ROCm 10”无对应运行，四个 cell 仍为 `verified: false`。

`docs/src/snippets/configs/zai-org/glm-5.2.jsx`

本次变更的核心文件：同时承担镜像分桶升级（dockerImages["mi355xlmxfp4"] 由 20260910 到 20260913）与四个 MXFP4 cell 的 env 注入（SGLANG_OPT_USE_TOPK_V2=true），是 Playground/cookbook 生成启动命令的唯一事实来源。

```
// AMD MI355X 的镜像按量化形态分桶：FP8/BF16 档沿用较旧的 20260618 镜像，
// 只有 `mi355xlmxfp4` 这一档跟随最新日构建（本次 20260910 → 20260913）。
// 注意：默认 mi355x 档未同步升级，因此 TOPK_V2 只在 MXFP4 四个 cell 上被打开。
```

```
dockerImages: {
  h200: "lmsysorg/sglang:latest",
  b200: "lmsysorg/sglang:latest",
  gb300: "lmsysorg/sglang:latest",
  b300: "lmsysorg/sglang:latest",
  mi355x: "lmsysorg/sglang-rocm:v0.5.13.post1-rocm720-mi35x-20260618",
  // 本次唯一被 bump 的条目：该镜像内含 GLM-5.2 MXFP4 所需的 ROCm 侧改动，
  // 也是启用 SGLANG_OPT_USE_TOPK_V2 的前置条件（旧镜像不要开这个 flag）。
  "mi355xlmxfp4": "lmsysorg/sglang-rocm:v0.5.19-rocm720-mi35x-20260913",
  mi325x: "lmsysorg/sglang-rocm:v0.5.13.post1-rocm700-mi30x-20260616",
  mi300x: "lmsysorg/sglang-rocm:v0.5.13.post1-rocm700-mi30x-20260616",
},
```

```
// MI355X + MXFP4 一共四个 cell（low-latency / balanced / high-throughput / mtp-314），
// 本次统一把 env 从 [] 改为注入 SGLANG_OPT_USE_TOPK_V2=true，从而在 ROCm 上走
// v2 fused top-k 的 MoE 路由内核（上游 sgl-project/sglang#36684、#36851 落地，
// 取代此前 gfx950 上强制 v1 的 workaround）。
```

```
// flags 保持原样：TP/EP 拓扑、--kv-cache-dtype fp8_e4m3、Triton DSA 前后端、
// MTP 步数与 --mem-fraction-static 都未改，性能差异只能归因于 top-k 内核本身。
```

```
{
  match: { hw: "mi355x", variant: "default", quant: "mxfp4", strategy: "low-latency", nodes:
    "single" },
  verified: false,
  env: ["SGLANG_OPT_USE_TOPK_V2=true"], // ← 本次新增；balanced / high-throughput / mtp-
    314 同样处理
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp 8",
    "--ep-size 1",
    "--kv-cache-dtype fp8_e4m3",
    "--dsa-prefill-backend triton",
    "--dsa-decode-backend triton",
    "--speculative-algorithm EAGLE",
    "--speculative-num-steps 5",
    "--speculative-eagle-topk 1",
    "--speculative-num-draft-tokens 6",
    "--chunked-prefill-size 131072",
    "--mem-fraction-static 0.80",
    "--watchdog-timeout 1200",
    "--host {{HOST_IP}}",
```

```
"--port {{PORT}}",  
],  
},
```

评论区精华

本 PR 讨论极简：2 条 APPROVED、0 条行内 review 评论、1 条 issue 评论，没有技术分歧被记录。

sogalin (APPROVED) : 「LGTM, SGLANG_OPT_USE_TOPK_V2=1 can help improve the performance for GLM.」

这是全 PR 唯一的价值论证，来自维护者经验而非本 PR 的实测数据。

1am9trash (APPROVED) : 「LGTM, cc @Jacob0226, @EricKing626」

纯批准并转交相关同学，未附加条件或疑问。

值得留意的三个细节：其一，无人追问「旧镜像 + TOPK_V2」组合是否安全；其二，HiCache 双 arm 说明不是被 review 逼退的，而是作者在第三个 commit 自行 Revert，属主动的范围收敛；其三，PR body 标记了“PR Test (Extra)”失败，但讨论中没有人解释原因，也没有把它与本次改动建立或排除关联。

风险与影响

风险

1. 镜像与 flag 的隐式耦合：SGLANG_OPT_USE_TOPK_V2=true 只在 "mi355xlmxfp4" 分桶下注入，其正确性依赖 20260913 镜像提供 v2 fused top-k 内核；但默认 mi355x 档仍停在 20260618，MDX 也没有写出「旧镜像请勿开启」的告警。用户混用旧镜像 + 手动补 flag，就会走到未验证路径。
2. 缺少实机与性能证据：AMD ROCm 10 无 CI 运行、PR Test (Extra) 失败且未解释、四个 cell 仍为 verified: false；性能收益目前只有 reviewer 的经验判断。
3. 外部同步源未闭环：依据的 InferenceX#3081 仍是 open，且含本 PR 明确不做的并发点集重分配，上游再动一次就会带来新一轮对齐成本。
4. 回退带来的信息缺口：双 arm 说明被 Revert 后，cookbook 读者无法从文档了解 TP4/EP4 在低并发下已被移出 sweep 的背景；若 InferenceX 侧已承载该解释则可接受。

影响

- 用户：使用 GLM-5.2 MI355X MXFP4 配方的用户会切到新镜像并启用新内核，预期 MoE top-k 段开销下降；FP8/BF16 及其他硬件档位完全不受影响。
- 系统：零运行时改动，无兼容性、显存或部署影响；回滚只需回退文档。
- 团队：固化了一次「AMD 日构建镜像 + env flag 同步」的操作范式，并用一次 Revert 划清了 cookbook 与 benchmark harness 的职责边界。

关联脉络

- 上游依赖：#36684 与 #36851（在 PR body 与 MDX 中被引用）让 v2 fused top-k 内核在 ROCm 上对 GLM-5.x 可用，取代 gfx950 上强制 v1 的 workaround；本 PR 是能力落地

后的配方跟进。

- 同模式先例：#39358 把 Qwen3.5 MI355X cookbook 对齐到 AttnFP8-V2 与 HiCache 直连布局，#39396 为 DeepSeek-V4 MI355X 补齐 PD 配方，均改动 docs/src/snippets 下的配置文件，说明「AMD 新特性 → 文档配方同步」已是一条稳定的工作流。
- 平台特色：#37413 在 gfx950 上通过开关启用 fp8 two-pool unified_kv，与本 PR 在 MXFP4 上打开 TOPK_V2 属于同一类「条件具备后打开 AMD 特有开关」的变更。
- 流程面：#39324 / #39346 的 sgl-kernel 版本 bump 与本次镜像 tag 升级都涉及外部制品版本同步，但前者影响构建产物、后者只影响文档里的容器 tag，风险等级不可同日而语。

结论：这是一个低风险、低复杂度但需要留意细节一致性的文档 PR，值得借鉴的是「用 Revert 收敛范围」与「镜像 /env 单一事实来源」两点，而不是它的代码量。