

PR #39631 完整报告

sgl-project/sglang

[AMD] Prefer HIP Top-K for GLM-5.x on ROCm

合并时间: 2026-09-16 11:55

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39631>

执行摘要

- 一句话: ROCm 默认切 HIP Top-K, GLM-5.x 交互性提升
- 推荐动作: 值得快速阅读 (约 5 分钟): 单文件小改动, 但包含一个可复用的设计模式——用 env 的 `is_set()` 判断区分“用户显式选择”与“平台默认值”, 避免覆盖用户配置; 同时以 A/B 数据驱动平台默认 kernel 切换。关注 AMD 平台支持、参数解析或环境变量契约的工程师建议阅读; 对纯 CUDA 使用者影响很小。

功能与动机

PR #36851 曾为 GLM-5.x 在 ROCm 上强制启用 Top-K v2, 但该 fused top-k v2 内核 (`topk_transform_paged_v2`) 是 CUDA/Hopper 专属 JIT 路径, 依赖与 `cg::this_cluster()` 线程块簇特性, ROCm `gfx9xx` 无法 JIT 编译; 而 GLM override 会无条件替换 DSA 家族已有的 HIP Top-K 选择, 公共 fallback 还会覆盖用户的显式设置。AgentX A/B 验证显示 HIP Top-K 在 P90 交互性上更优 ($96.97 \rightarrow 106.69$ tok/s/user, +10.0%) 且吞吐基本不变, 因此本 PR 以数据为依据调整默认值并修复覆盖用户设置的问题。

实现拆解

1. 移除 GLM 专属 override: 删除 `python/sglang/srt/arg_groups/model_hook.py` 中 `GlmMoeDsaForCausalLM` 分支下的 `envs.SGLANG_OPT_USE_TOPK_V2.set(True)` 及相关解释注释 (约 10 行), 使 GLM-5.x 回到 DSA 家族统一的 Top-K 决策路径, 不再被特殊对待。
2. 收紧 DSA 默认值写入条件: 将 `is_deepseek_dsa(hf_config)` 改为 `is_deepseek_dsa(hf_config) and not envs.SGLANG_OPT_USE_TOPK_V2.is_set()`, 只有当环境变量尚未被显式设置时才写入默认值 `False`, 避免覆盖用户的显式选择。
3. 更新兼容性注释: 把原先关于“CUDA/Hopper 专属 JIT 内核在 ROCm 无法编译”的说明改写为“默认优先 HIP Top-K 同时尊重显式选择”, 使注释与新行为一致。
4. 验证与配套: 通过 `py_compile` 和 `git diff --check`; MI355X 上 AgentX A/B (并发 8、1200 秒) 显示 P90 交互性从 96.97 提升到 106.69 tok/s/user (+10.0%), 每 GPU 吞吐 $9553.2 \rightarrow 9539.0$ tok/s (-0.15%); 3200 秒长时验证当时仍 pending。本 PR 未附带测试文件, 也未改动配置 schema。

关键文件:

- python/sglang/srt/arg_groups/model_hook.py (模块 模型钩子; 类别 source; 类型 core-logic; 符号 handle_model_specific_adjustments): 唯一变更文件, 承载 Top-K 路径默认值决策的核心控制流调整

关键符号: handle_model_specific_adjustments

关键源码片段

python/sglang/srt/arg_groups/model_hook.py

唯一变更文件, 承载 Top-K 路径默认值决策的核心控制流调整

```
# python/sglang/srt/arg_groups/model_hook.py
# handle_model_specific_adjustments: 模型特定参数收尾入口。
# 以下为 ROCm (HIP) 平台上 DSA 家族 (DeepSeek-V3.2 / GLM-5.x)
# 的 Top-K 路径默认值决策逻辑, 也是本次 PR 的核心改动。

run_post_process_pass(server_args, _deepseek_moe_quant_resolution)
if get_platform().is_hip:
    if is_deepseek_dsa(hf_config) and not envs.SGLANG_OPT_USE_TOPK_V2.is_set():
        # 默认走预编译 HIP Top-K 路径。
        # topk_transform_paged_v2 是 CUDA/Hopper 专属 JIT 内核: 其源码
        # 依赖 <cooperative_groups.h> 与 cg::this_cluster() (线程块簇),
        # 在 ROCm gfx9xx 上无法 JIT 编译, CUDA graph 捕获会直接失败。
        # 与旧逻辑 (无条件 set + GLM override) 不同, 这里仅在用户没有
        # 显式设置 SGLANG_OPT_USE_TOPK_V2 时才写入默认值 False,
        # 显式选择 Top-K v2 或 HIP Top-K 都会被原样保留。
        envs.SGLANG_OPT_USE_TOPK_V2.set(False)
    if not resolved_view(server_args).enable_dp_attention and cfg.nnodes == 1:
        logger.info("Enable Aiter AllReduce Fusion for DeepseekV3ForCausalLM")
```

评论区精华

该 PR 没有实质性的 review 讨论: HaiShaw 直接给出 APPROVED (LGTM), 仅在关联 Issue 下留下“HIP specific”一条评论, 将变更定位于 HIP 平台。值得注意的权衡在 PR body 中: GLM 的 Top-K v2 支持 (#36851) 与 HIP 预编译 Top-K 路径在 ROCm 上存在目标冲突, 作者以 A/B 数据 (P90 +10%、吞吐 -0.15%) 为决策依据, 并选择用 `is_set()` 判断来同时满足“默认走 HIP Top-K”与“尊重显式配置”两个目标。

- 变更范围与平台定性 (other): 评审方认可 HIP 平台专属变更, 无需修改直接合并。

风险与影响

- 风险: 核心逻辑变更集中在 model_hook.py 的 HIP 分支, 影响 ROCm 上所有 `is_deepseek_dsa` 命中的模型 (DeepSeek-V3.2 / GLM-5.x), 默认 Top-K 实现从 JIT Top-K v2 切到预编译 HIP Top-K, A/B 显示吞吐微降 0.15% 且长时验证当时未完成, 存在极端负载下表现不确定的风险。另外, `env` 默认值语义依赖 `SGLANG_OPT_USE_TOPK_V2.is_set()` 的“显式设置”判断, 如果后续代码在其他路径上以编程方式 `set` 过该变量 (作为内部默认而非用户意图), 会被误判为显式选择而跳过平台默

认值，需要在后续演进中留意。最后，本次没有新增或修改测试文件，行为回归依赖 AMD CI 与人工验证。

- 影响：用户侧：AMD/ROCm 上运行 GLM-5.x、DeepSeek-V3.2 等 DSA 系模型的用户，默认推理路径变化，交互性 P90 提升约 10%，吞吐基本持平；显式设置过 `SGLANG_OPT_USE_TOPK_V2` 的用户行为不变。一致性侧：AMD 平台与 CUDA 平台在 Top-K 实现上的对齐策略发生调整——此前通过 GLM override 强制与 CUDA 行为对齐，现在 AMD 平台默认以本地预编译内核优先，跨平台输出可能产生微小数值差异。团队侧：该 PR 是 #36851 的 follow-up，代表 AMD 平台“先对齐、再按数据收敛默认值”的路径，为后续其他平台 kernel 选择提供了可复制的决策样例。
- 风险标记：AMD 平台默认行为变更，缺少测试覆盖，依赖 env 显式设置语义，长时验证未完成

关联脉络

- PR #36851 Enable Top-K v2 for GLM-5.x on ROCm（转述自本 PR body）：本 PR 的直接前身：PR body 说明 #36851 添加了 GLM 特定的 Top-K v2 override，本次将其移除并改为默认 HIP Top-K，是同一 Top-K 选择逻辑的演进。
- PR #39547 [AMD][bugfix] Fix DSV4 MTP crash: 同属 AMD HIP 平台注意力 / 解码路径的稳定性维护，与本次 Top-K 默认值调整共同反映 AMD 平台内核选择与修复的持续投入。