

PR #34347 完整报告

sgl-project/sglang

[Diffusion][MiniMax H3] Fix SM120 QKNorm+RoPE rounding

合并时间: 2026-08-12 09:18

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34347>

执行摘要

- 一句话: SM120 融合 QKNorm+RoPE 舍入修复, 回归通过
- 推荐动作: 值得精读。这是一个典型的“编译器收缩破坏 bit-exact 语义”案例: 单个头文件内用模板 `constexpr if` + 架构宏 + `asm volatile` 组合, 为 SM120 强制 `round-to-nearest` 舍入顺序, 同时为 pre-SM120 保留逐字节不变的旧算术。可重点关注: PTX `mul.rn / add.rn / sub.rn` 的位模式处理方式、`bf16_t` 与 `__half` 的统一分支写法, 以及用架构 `guard` 控制风险面的设计。建议后续为 SM120 增加更细的独立回归用例, 覆盖两种 `rotary` 布局与 `dtype` 组合。

功能与动机

PR body 说明: The diffusion fused QKNorm+RoPE kernel is expected to preserve the split BF16/FP16 rounding boundary: both RoPE products round to the activation dtype before the final add or subtract. On SM120, nvcc may contract the packed local expression, so the fused result differs from the split reference for a small number of BF16 values and fails the exact-rounding regression test. 即融合 kernel 的精度契约要求两个 RoPE 乘积先各自舍入到 BF16/FP16 再做加减; SM120 上 nvcc 将紧凑表达式收缩后, 少量 BF16 值不再满足该契约, 导致 exact-rounding 回归测试失败。修复目标是恢复位级一致, 而不是放宽测试。

实现拆解

1. 定位问题: 在 `python/sglang/kernels/jit/csrc/diffusion/qknorm_rope.cuh` 的融合 kernel `fused_qknorm_rope_warp` 中, NeoX 与 interleaved 两种 rotary 布局分别使用 `values[i] * cos + partner_values[i] * sin`、`x * cos - y * sin` 等紧凑表达式。SM120 上编译器可能将其收缩, 导致两个乘积不再先分别舍入到 BF16/FP16, 从而与 split 参考 kernel 产生少量位级差异。
2. 新增舍入 helper: 新增模板函数 `rotary_mul_rn`、`rotary_add`、`rotary_sub`。当 `__CUDA_ARCH__ >= 1200` 时, 通过 `__bfloat16_as_ushort / __half_as_ushort` 取位模式, 再用 `asm volatile` 发射 `mul.rn.bf16`、`mul.rn.f16` 以及 `add.rn`、`sub.rn` 等 PTX 指令, 强制每一步都 `round-to-nearest` 到激活 dtype; pre-SM120 走回退分支, 保持原 C++ 算术逐字节不变。
3. 替换调用点: NeoX layout 的 `lane_id < kHalfRotaryLanes` 分支改为 `rotary_sub / rotary_add`, interleaved layout 的 `x / y` 分支同步替换; 同一套 helper 通过

`std::is_same_v` 同时覆盖 `bf16_t` 与 `__half` 两种类型。

4. 测试与性能配套：未新增测试文件，依赖既有 `exact-rounding` 回归与全量 QKNorm+RoPE 套件。PR body 给出 RTX 5090 `exact-rounding` 回归 1 `passed`, 1250 `deselected`、H100 全量 1251 `passed`；RTX 5090 五个 production shape 的 fused 延迟基本不变（71.680 us ~ 94.592 us 量级），H100 差异在 -0.45% ~ +0.43%。

关键文件：

- `python/sglang/kernels/jit/csrc/diffusion/qknorm_rope.cuh`（模块 JIT 内核；类别 `source`；类型 `core-logic`；符号 `rotary_mul_rn`, `rotary_add`, `rotary_sub`, `fused_qknorm_rope_warp`）：唯一变更文件，是融合 QKNorm+RoPE 的 JIT kernel 头文件。本 PR 的核心修正在此：为 SM120 新增显式 `round-to-nearest` 的 PTX helper，并替换 NeoX 与 `interleaved` 两种 `rotary` 布局的调用点。

关键符号：`rotary_mul_rn`, `rotary_add`, `rotary_sub`, `fused_qknorm_rope_warp`

评论区精华

本 PR 没有公开 review 评论线程，技术要点集中在 PR body 的约束声明与验证数据：

“On SM120, `nvcc` may contract the packed local expression, so the fused result differs from the split reference for a small number of BF16 values and fails the `exact-rounding regression test`.”

“Preserve the existing C++ arithmetic byte-for-byte for pre-SM120 targets.”

“RTX 5090 fused latency stayed effectively unchanged on five production shapes.”

核心决策是：用显式 PTX `mul.rn` / `add.rn` / `sub.rn` 指令代替对编译器的信任，同时用 `__CUDA_ARCH__` 宏把风险隔离在 SM120 之内，保证 SM90/H100 路径行为不变。

- 暂无高价值评论线程

风险与影响

- 风险：技术风险
- 内联汇编依赖：`asm volatile` 使用 `"=h" / "h"` 约束，依赖 16-bit 寄存器映射和 `__bfloat16_as_ushort` 等位转换的端序语义；更换编译器版本或目标架构时需重新验证。
- 平台条件编译：`guard` 只针对 `__CUDA_ARCH__ >= 1200`，若未来新架构（如 SM130）同样出现 `contraction`，需要同步扩展；当前没有自动测试覆盖新架构的行为。
- 测试覆盖偏薄：变更未新增独立测试文件，RTX 5090 上 `exact-rounding` 回归仅 1 `passed`, 1250 `deselected`；若回归用例未覆盖全部 `rotary` 布局或 `dtype` 组合，仍有漏网风险。
- 精度契约敏感：该 kernel 属于 `diffusion` 精度关键路径，后续任何重构都可能重新引入 SM120 上的舍入差异。
- 影响：用户与系统：仅影响 SM120（RTX 50 系列）上 `diffusion` 融合 QKNorm+RoPE 路径，输出与 `split` 参考 `bit-exact` 对齐，消除 `exact-rounding` 回归失败；SM90/H100 等 pre-SM120 平台的代码路径完全未动，行为不变。

性能: RTX 5090 五个 production shape 的 fused 延迟几乎不变, H100 差异在 -0.45% ~ +0.43%, 属于测量噪声量级。

团队与 CI: 去掉一个 SM120 上的确定性回归失败; 为后续 diffusion 内核优化提供了可依赖的位级精度基线。

- 风险标记: 平台条件编译 (SM120 专属路径), 内联 PTX 汇编, 精度位级敏感

关联脉络

- PR #34314 [diffusion] Ideogram-4: fuse Qwen3-style RoPE and SwiGLU silu-mul (denoise -5.1% H100 / -4.7% H200, bit-exact): 同属 diffusion JIT 内核的 RoPE/SwiGLU 融合工作, 且同样强调 bit-exact 验证; 本 PR 修复的 qknorm_rope.cuh 正是融合内核路径的精度保障。
- PR #32921 [diffusion][model] Add native SANA-Video T2V support: diffusion 原生内核方向的模型接入 PR, 与本 PR 共享融合内核与精确舍入回归测试的上下文。
- PR #34423 [diffusion] fix: nightly diffusion benchmark passes the retired --warmup flag: diffusion 回归 / 基准 CI 的维护 PR, 与本 PR 的 exact-rounding 回归测试同属 diffusion 质量保障链。