PR #30278 完整报告

sgl-project/sglang

Fix LTX2 RoPE JIT kernel CI

合并时间：2026-07-07 10:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30278

执行摘要

- 一句话：修复 LTX2 RoPE JIT 内核测试精度问题
- 推荐动作：该 PR 属于小型修复，对于关注 CI 稳定性和精度模型的开发者有一定参考价值。值得注意的决策是使用 F.rms_norm 在 fp32 域直接计算，避免了 nn.RMSNorm 在 autocast 下的隐式类型转换，这是一种通用的测试精度修复模式。

功能与动机

CI 测试失败（参见 PR body 中的 CI 日志链接），原因是 JIT 内核测试未能通过精度校验。根因是测试参考实现使用了 torch.nn.RMSNorm 配合 autocast，导致归一化在 bf16 下舍入，与 CUDA 内核的 fp32 归一化行为不匹配，产生精度误差。

实现拆解

1. 修正测试参考实现（test/registered/jit/diffusion/test_ltx2_qknorm_split_rope.py）：将 _reference 函数中的 torch.nn.RMSNorm + autocast 替换为 F.rms_norm，并在 fp32 域执行计算，避免额外的类型转换和舍入。添加了导入 F。
2. 更新 CUDA 内核注释（python/sglang/jit_kernel/csrc/diffusion/ltx2_qknorm_split_rope.cuh）：修改注释以准确描述精度模型——RMSNorm 和 split RoPE 都在 fp32 下运行，仅在最终 attention 输入前舍入到 bf16。

关键文件：

- test/registered/jit/diffusion/test_ltx2_qknorm_split_rope.py（模块 测试；类别 test；类型 test-coverage；符号 _reference）：测试参考实现的核心修改：用 F.rms_norm 替换 nn.RMSNorm + autocast，使精度模型与 CUDA 内核一致。
- python/sglang/jit_kernel/csrc/diffusion/ltx2_qknorm_split_rope.cuh（模块 JIT 内核；类别 other；类型 core-logic）：更新 CUDA 内核注释以准确描述精度模型，反映修复后的测试预期。

关键符号：_reference

关键源码片段

test/registered/jit/diffusion/test_ltx2_qknorm_split_rope.py

测试参考实现的核心修改：用 F.rms_norm 替换 nn.RMSNorm + autocast，使精度模型与 CUDA 内核一致。

```
def _reference(
    q: torch.Tensor,
    k: torch.Tensor,
    q_cos: torch.Tensor,
    q_sin: torch.Tensor,
    k_cos: torch.Tensor,
    k_sin: torch.Tensor,
    q_weight: torch.Tensor,
    k_weight: torch.Tensor,
    eps: float,
) -> tuple[torch.Tensor, torch.Tensor]:
    # rms_norm isn't autocast fp32-preserving, so feed fp32 inputs directly
    # to keep the normalized value unrounded until the final RoPE output.
    q_norm = F.rms_norm(q.float(), (q.shape[-1],), q_weight.float(), eps)
    k_norm = F.rms_norm(k.float(), (k.shape[-1],), k_weight.float(), eps)
    q_ref = _apply_split_rotary_ref(q_norm, q_cos, q_sin)
    k_ref = _apply_split_rotary_ref(k_norm, k_cos, k_sin)
    return q_ref.to(dtype=torch.bfloat16), k_ref.to(dtype=torch.bfloat16)
```

评论区精华

PR 没有 review 评论或讨论。但参考 PR body 的 CI 失败链接和提交记录可以推断，问题在于测试参考实现与 CUDA 内核的精度模型不一致。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅限于测试文件和注释，不影响任何生产逻辑。所有 5 个测试用例均通过 CI。
- 影响：影响范围仅限于 LTX2 QKNorm split-RoPE JIT 内核的 CI 测试。修复后 CI 测试将不再因精度问题失败，保障了该内核的持续集成可靠性。
- 风险标记：暂无

关联脉络

- PR #30117 Support Cutedsl BF16 GEMM JIT kernel: 同为 JIT 内核相关 PR，且该 PR 的 CI 失败可能与此处修复的测试有关（PR body 引用了 30117 的 CI 失败链接）。
- PR #29690 （未在历史 PR 列表中，但 PR body 引用了其 CI 失败链接）：PR body 中第一个 CI 失败链接来自该 PR，可能是最初引入 LTX2 JIT 内核的 PR。