PR #37112 完整报告

sgl-project/sglang

[Diffusion] Fuse FLUX.2 gated residual normalization on Blackwell

合并时间：2026-09-01 08:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/37112

执行摘要

- 一句话：FLUX.2 门控残差归一化融合为单 CUDA 内核，去噪步长提速约 0.78%
- 推荐动作：值得精读。三个设计决策值得学习：(1) 跨 block 延迟物化——把 FFN 残差更新推迟到下一个 LayerNorm，用

功能与动机

FLUX.2 的 transformer block 采用 parallel 结构，注意力输出经 residual_gate_add 更新残差后，下一个 block 的 LayerNorm 与 adaLN 调制又各自启动独立内核。profile 显示每个 denoise step 有 80 次 gated residual 与 80 次 norm/modulate 调用、共 160 次相关 kernel launch。PR body 明确该优化来自 Baseten 的 "Agentic Kernels in Production" 博客，目标是减少 launch 开销并压缩 denoise step 延迟。融合后相关调用变为 77 次融合加边界 3+3 次，每 step 净减 77 次 GPU 启动，且以 bit-exact 作为硬性验收标准，保证图像质量零回退。

实现拆解

变更入口是 python/sglang/multimodal_gen/runtime/models/dits/flux_2.py 的 FLUX.2 transformer 前向链，配套新增 JIT 调度模块与 CUDA 内核。

1. CUDA 内核层：新增 python/sglang/kernels/jit/csrc/diffusion/flux2_gated_resnorm.cuh，D=6144 专用、128 线程（4 warp）、每线程 4 元素 128 位向量化。第一遍读 residual/update/gate，按 BF16 舍入语义做乘积与累加并写回更新后残差，同时累积 Welford 统计；第二遍复现 aten 的四 warp 归约树（先 (0,2)、(1,3)，再 (0,1)）计算 mean/rstd；第三遍重读更新后残差完成归一化与 adaLN。三遍结构刻意重读残差以压低寄存器压力、保持占用率。
2. JIT 调度层：新增 python/sglang/kernels/ops/diffusion/norm/flux2_gated_resnorm_jit.py。can_defer_flux2_gated_residual 集中检查 torch.compile 状态、BF16、CUDA 设备、batch 1、hidden 6144、非空、连续、32 字节对齐、Blackwell+ 以及 gate 行形态；can_use_flux2_gated_resnorm 进一步要求 scale/shift 可折叠为 [6144] 行；flux2_gated_resnorm_raw 通过 load_jit 加载内核并执行。任何前置条件不满足都返回 False，由上层走 eager 路径。
3. 模型主路径：修改 flux_2.py，引入 PendingGatedResidual 三元组类型，Flux2ParallelSelfAttention.forward 与 Flux2TransformerBlock.forward 的入参 / 返回值

从 torch.Tensor 扩展为 Tensor | PendingGatedResidual。_defer_gated_residual 在满足条件时返回三元组（延迟物化），_flux2_gated_resnorm 在下一个 LayerNorm 处调用融合内核；在 split_hidden_states、双流 / 单流边界、单 block 推理尾部以及模型输出处用 _materialize_gated_residual 强制落地，FP16 clip 只在已物化 Tensor 上执行。

4. 导出与测试配套：python/sglang/kernels/ops/diffusion/init.py 登记 can_defer_flux2_gated_residual、can_use_flux2_gated_resnorm、flux2_gated_resnorm_raw 三个导出符号；新增 test/registered/jit/test_flux2_gated_resnorm.py，注册到 base-b-kernel-unit 阶段的 4-gpu-b200 runner，断言融合输出与 eager 位级一致，并覆盖 FP16、batch 2、torch.compile 模式的降级行为。PR 还验证了 --enable-torch-compile 与 --dit-layerwise-offload true 两种生产模式的完整图像生成。

关键文件：

- python/sglang/multimodal_gen/runtime/models/dits/flux_2.py（模块 sglang/multimodal_gen；类别 source；类型 data-contract；符号 _materialize_gated_residual, _defer_gated_residual, _flux2_gated_resnorm）：源码主路径；涉及符号 _materialize_gated_residual, _defer_gated_residual, _flux2_gated_resnorm；包含 控制流调整、配置键调整、符号定义调整；+107/-31
- test/registered/jit/test_flux2_gated_resnorm.py（模块 flux2/gated/resnorm；类别 test；类型 test-coverage；符号 _is_blackwell, TestFlux2GatedResnorm, test_gated_residual_norm_modulate_is_bit_exact, test_gated_residual_defer_rejects_unsupported_inputs）：测试配套；涉及符号 _is_blackwell, TestFlux2GatedResnorm, test_gated_residual_norm_modulate_is_bit_exact, test_gated_residual_defer_rejects_unsupported_inputs；包含 测试覆盖调整、导入关系调整、控制流调整；+68/-0
- python/sglang/kernels/ops/diffusion/norm/flux2_gated_resnorm_jit.py（模块 sglang/kernels；类别 infra；类型 infrastructure；符号 _blackwell_or_newer, _aligned, _row_bf16, can_defer_flux2_gated_residual）：部署 / 基础设施；涉及符号 _blackwell_or_newer, _aligned, _row_bf16, can_defer_flux2_gated_residual；包含 导入关系调整、控制流调整、配置键调整；+119/-0
- python/sglang/kernels/jit/csrc/diffusion/flux2_gated_resnorm.cuh（模块 sglang/kernels；类别 other；类型 dependency-wiring）：辅助文件；包含 导入关系调整、控制流调整、配置键调整；+231/-0
- python/sglang/kernels/ops/diffusion/__init__.py（模块 sglang/kernels；类别 infra；类型 infrastructure）：部署 / 基础设施；包含 配置键调整；+3/-0

关键符号：_materialize_gated_residual, _defer_gated_residual, _flux2_gated_resnorm, _is_blackwell, TestFlux2GatedResnorm, test_gated_residual_norm_modulate_is_bit_exact, test_gated_residual_defer_rejects_unsupported_inputs, _blackwell_or_newer, _aligned, _row_bf16

评论区精华

本 PR 没有人工 review 评论（review_comments_count=0），Issue 评论区仅含作者本人的 5 条 CI 操作命令（/tag-run-ci-label extra 与 4 次 /rerun-failed-ci），评审以自动验证为主。

真正有分量的论证集中在 PR body：以 profile 数据说明 77 次 launch 减少，以双 SHA256一致、SSIM 1.0、PSNR infinity、LPIPS 0.0 断言位级等价，把 *eager* 数值契约当作硬性验收标准——任何一位输出不一致都视为失败。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 数据契约变更：Flux2TransformerBlock.forward 与 Flux2ParallelSelfAttention.forward 的返回类型从 torch.Tensor 变为 torch.Tensor 或 PendingGatedResidual 联合类型，任何新增调用点若未处理 tuple 分支都会出现类型错误或错误结果；当前靠边界强制物化兜底，但这是需要维护者长期理解的隐式约定。
  - 延迟物化语义风险：*deferred* 三元组是跨 block 的临时状态，若未来在中间环节插入读取 hidden_states 的操作（如 logging、profiling、sparsity 钩子），可能绕过物化点导致错误；FP16 clip 只作用于已物化 Tensor，tuple 形态不会 clip。
  - 数值等价依赖内核实现：bit-exact 只对测试覆盖的 [1, 64, 6144] 形状直接验证，其他生产形状（如 [1, 512, 6144]、[1, 4096, 6144]）依赖内核通用实现；PR 微基准显示这些形状 max abs diff 为 0，但自动化测试未覆盖。
  - 平台与编译链耦合：快速路径绑定 Blackwell（SM 10+）与 BF16；新增 .cuh 依赖 load_jit 编译链，sgl_kernel 头文件 API 变更可能影响编译，与同系列融合 PR 共享该风险。
  - CI 覆盖依赖 B200 硬件：测试注册在 4-gpu-b200 runner，非 Blackwell 环境自动 skip，缺少对其他架构的回归保护。
- 影响：
  - 用户影响：FLUX.2 BF16 在 Blackwell（GB300/B200）上生成延迟小幅下降（*denoise/step* -0.78%、*warmed* 端到端 -0.71%，内核微基准 1.4-2.2 倍加速），输出图像与 *eager* 路径完全一致；非 Blackwell、FP16、batch >1 等场景自动回退，行为不变。
  - 系统影响：每 *denoise step* GPU kernel 启动数从 1121 降至 1044（-6.87%），相关内核时间 -21.85%，对 launch-bound 的 diffusion 部署有利。
  - 团队影响：建立跨 block 延迟物化 + 位精确融合内核的可复用模式，与 Qwen-Image 系列融合 PR 形成体系；新增 B200 CI 测试挂载点，后续融合类改动可沿用。
  - 风险标记：暂无

关联脉络

- 暂无明显关联 PR