PR #34350 完整报告

sgl-project/sglang

[Diffusion] Avoid slow cuBLASLt GELU epilogue on SM120

合并时间：2026-08-12 12:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34350

执行摘要

- 一句话：SM120 禁用 cuBLASLt GELU 融合，RTX 5090 提速约 20.4%
- 推荐动作：值得精读。PR 虽小但展示了“按架构限制 kernel 选择 + 位精确回退优先”的决策模式，并有完整的精度与性能验证方法。对于维护 diffusion 内核选择逻辑或关注 SM120/Blackwell 性能的工程师，本 PR 的 guard 写法（get_jit_cuda_arch 判架构）可以直接复用。

功能与动机

PR body 明确指出：FLUX 融合 Linear+GELU 路径使用的 cuBLASLt GELU epilogue 在 H100/H200 上是有益的，但 "the current PyTorch/CUDA stack selects an SM120 implementation that is slower than GEMM followed by native tanh-GELU for the production up-projection shape"，且 "The existing runtime guard therefore chooses a slower path on RTX 5090"。因此需要修正运行时 guard，让 RTX 5090 走更快的 eager 路径，同时保留 SM90 的融合加速。

实现拆解

1. 定位运行时 guard：python/sglang/kernels/ops/diffusion/fused_linear_gelu.py 中的 can_fuse_linear_gelu(linear, x) 是 FLUX 融合 Linear+GELU 路径的准入判断函数，此前只检查设备、dtype 与权重 dtype 一致性。
2. 新增架构判断：引入 from sglang.kernels.jit.utils import get_jit_cuda_arch，在函数早退路径中加入 arch.major * 10 + arch.minor >= 120 即返回 False 的分支，将 SM120 及以上架构排除在 cuBLASLt 融合之外。
3. 行为拆分与精度取舍：SM120 走已有 eager Linear + 原生 tanh-GELU 实现，保留参考结果（避免 cuBLASLt epilogue 接近但不位精确的结果）；SM90 融合路径不变。
4. 验证与基准：RTX 5090 上 test_fused_linear_gelu.py 5 passed（含 torch.compile fullgraph 用例）；H200 变更路径套件 1265 passed；H100 diffusion 套件 3227 passed、22 skipped。FLUX 生产形状 (1, 512, 3072) -> 12288 下，RTX 5090 runtime dispatch 从 295.96 us 降至 235.67 us（约 -20.4%），H200 无回归。

关键文件：

- python/sglang/kernels/ops/diffusion/fused_linear_gelu.py（模块 内核选择；类别 source；类型 core-logic；符号 can_fuse_linear_gelu）：核心修改文件，在

can_fuse_linear_gelu 中加入 SM120+ 架构 guard，拒绝慢速 cuBLASLt GELU epilogue，回退到 eager 路径。

关键符号：can_fuse_linear_gelu

关键源码片段

python/sglang/kernels/ops/diffusion/fused_linear_gelu.py

核心修改文件，在 can_fuse_linear_gelu 中加入 SM120+ 架构 guard，拒绝慢速 cuBLASLt GELU epilogue，回退到 eager 路径。

```
# python/sglang/kernels/ops/diffusion/fused_linear_gelu.py

import torch
import torch.nn as nn

# 新增导入：获取 JIT CUDA 架构的工具函数，用于运行时判断
from sglang.kernels.jit.utils import get_jit_cuda_arch
from sglang.kernels.ops.diffusion.quality_gate import QualityGatedFusion
from sglang.srt.utils.custom_op import register_custom_op


def can_fuse_linear_gelu(linear: Any, x: torch.Tensor) -> bool:
    """判断 ``gelu(linear(x))`` 当前是否可以使用融合的 cuBLASLt epilogue。"""
    # 非 CUDA 输入或非 bf16/fp16 精度直接拒绝融合
    if not (x.is_cuda and x.dtype in (torch.bfloat16, torch.float16)):
        return False

    # SM120（如 RTX 5090）：当前 PyTorch/CUDA 栈选中的 cuBLASLt GELU
    # epilogue 实现比 GEMM + 原生 GELU kernel 更慢，实测 FLUX 生产
    # up-projection 形状 (1, 512, 3072) -> 12288 下整体放缓约 20%，
    # 因此在此显式拒绝融合，走 eager Linear + 原生 tanh-GELU 路径。
    arch = get_jit_cuda_arch()
    if arch.major * 10 + arch.minor >= 120:
        return False

    # 权重缺失或 dtype 不一致时也无法融合
    if getattr(linear, "weight", None) is None or x.dtype != linear.weight.dtype:
        return False

    # SM90（H100/H200）等架构继续走原有静态判断，保持融合加速
    return can_fuse_linear_gelu_static(linear)
```

评论区精华

本 PR 没有实质性的 review 讨论线程。唯一一条 Issue 评论是作者 BBuf 贴出的 CI run 链接（Run #31513290921），用于指向验证产物。可以从 PR body 提炼的核心决策是：SM120 回退到已有 eager 实现以保留位精确的参考结果，而非继续使用接近但不位精确的 cuBLASLt epilogue。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 影响范围：*can_fuse_linear_gelu* 是通用 *guard*，不只是 FLUX。所有依赖该 *guard* 的 diffusion Linear+GELU 融合在 SM120+ 上都会回退到 *eager* 路径，其他模型的形状性能未验证。
  2. 性能风险：基准仅覆盖 FLUX 生产形状 (1, 512, 3072) -> 12288 的 BF16 测量；其他形状或 *dtype* 下 *cuBLASLt epilogue* 可能更快，架构级一刀切可能偏保守。
  3. 架构判断来源：*guard* 依赖 *get_jit_cuda_arch()* 的返回值；在非 JIT 环境或 AMD/NPU 等平台上，该工具函数的兼容性需要额外确认。
  4. 测量不确定性：PR *body* 提到 H100 时钟在隔离进程间漂移，SM90 开销通过同进程内归一化估算，存在亚微秒级噪声。
- 影响：
  1. 用户影响：RTX 5090（SM120）上的 diffusion/FLUX 用户获得约 20% 的 up-projection 阶段速度提升；H100/H200 用户行为不变。
  2. 系统影响：仅修改一个运行时 *guard* 函数，无 API、配置或部署变更，回归面小。
  3. 团队影响：为未来按架构指导 *kernel* 选择的模式提供了基准方法（架构 *guard* + 分架构性能验证），可复用到其他融合算子。 - 风险标记：SM120 架构级全局禁用融合 *epilogue*，基准仅覆盖 FLUX 生产形状，依赖 *get_jit_cuda_arch* 运行时判断

关联脉络

- PR #34349 [Diffusion] Tune QK head LayerNorm for SM120: 同批次针对 SM120（RTX 5090）的 diffusion kernel 性能调优，共享目标架构与验证方法论。
- PR #34347 [Diffusion][MiniMax H3] Fix SM120 QKNorm+RoPE rounding: 同批次 SM120 kernel 修复，均涉及 JIT/ 融合 kernel 在 SM120 上的行为修正。
- PR #34412 [Diffusion] Improve bit-exact fusion fallback diagnostics: 同为 diffusion 融合回退 / 诊断路径，涉及质量门控基础设施，与本 PR 的 *guard* 回退互补。