PR #22064 完整报告

sgl-project/sglang

[Diffusion] Fix weight scale swizzle and add large-M kernel config for FLUX.2-dev-NVFP4

合并时间：2026-04-04 11:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22064

执行摘要

本 PR 修复了扩散模型 FLUX.2-dev-NVFP4 路径中的权重缩放交织缺失和核配置浪费问题，优化了 CUTLASS 内核性能，并移除了过时的 comfy-kitchen 依赖。变更影响扩散模型的正确性和效率，建议相关工程师关注实现细节以提升量化实践。

功能与动机

PR 旨在解决两个关键 bug：首先，ModelOptFp4LinearMethod.process_weights_after_loading() 方法缺少 CUTLASS TMA 内核所需的块交织（swizzle），导致降级路径产生错误结果（cosine 相似度下降约 5%）；其次，默认核配置 KernelConfigDefault（集群 4x4）在 FLUX.2 模型 M≈4352 时浪费约 25% 计算资源。同时，移除 comfy-kitchen 第三方依赖以简化代码库。

实现拆解

变更涉及三个核心文件：

- python/sglang/jit_kernel/csrc/gemm/nvfp4/nvfp4_scaled_mm_sm100.cuh：添加 KernelConfigLargeM 结构体用于 M>1024（集群 1x4），并调整 KernelConfigDefault 的集群从 4x4 到 2x4 以优化中等 M 范围。代码片段：
- python/sglang/multimodal_gen/runtime/layers/quantization/modelopt_quant.py：在 process_weights_after_loading 方法中添加缺失的 reshape 和 permute 操作实现 swizzle，并移除 ComfyUIFp4LinearMethod 及相关代码。
- python/sglang/multimodal_gen/runtime/platforms/cuda.py：删除与 comfy-kitchen 相关的平台检测方法，如 has_modelopt_fp4_best_performance_kit。

评论区精华

Review 过程简单，仅由 mickqian 批准，未引发技术讨论。这表明变更被认为风险低且已通过测试，无需深入交锋。

风险与影响

风险：权重 swizzle 修复可能波及其他量化模型路径，需确保回归测试覆盖；新核配置针对 FLUX.2 优化，在其他大 M 场景下性能未验证；依赖移除可能影响向后兼容性，但测试显示无回归。 影响：修复提升扩散模型生成结果的正确性，优化 NVFP4 量化性能（CUTLASS 路径

比 cuDNN 快约 10%），并简化代码维护。

关联脉络

本 PR 是 NVFP4 量化支持演进的一部分：相关 PR 如 #22047 回滚了 NVFP4 Marlin 降级，显示 Blackwell GPU 上量化策略的调整；#21766 涉及 JIT 内核优化，与本 PR 的核配置技术相关；#20707 则扩展了扩散模型功能，表明扩散模块的持续增强。整体看，仓库正聚焦于高性能量化内核和扩散模型的集成优化。