PR #22091 完整报告

sgl-project/sglang

[diffusion] Default NVFP4 to CUTLASS and add all-model shape benchmarks

合并时间：2026-04-04 16:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22091

执行摘要

本 PR 将扩散模型 NVFP4 量化后端的默认实现从 comfy-kitchen 切换为 CUTLASS，并添加全模型形状基准测试，旨在基于数据驱动优化 Blackwell GPU 上的推理性能，影响扩散模块用户和性能调优流程。

功能与动机

为解决 Blackwell GPU 上 comfy-kitchen 后端非最佳默认的问题，PR body 引用基准测试结果：'On Blackwell, comfy-kitchen is not the best default once we look beyond FLUX.' 在 265 个形状案例中，CUTLASS 获胜 252 个，FLOP 加权时间比仅为 1.013x，优于 comfy 的 1.095x，因此切换默认以提升整体性能。

实现拆解

- 基准测试层：新增 bench_diffusion_nvfp4_scaled_mm.py 脚本和 diffusion_nvfp4_shapes.json 形状库，支持多后端性能对比。
- 平台接口层：修改 cuda.py 的 get_modelopt_fp4_gemm_op 函数，默认返回 CUTLASS 后端；新增 get_modelopt_flashinfer_fp4_backend 函数，通过环境变量控制 FlashInfer 后端。
- 配置与兼容性：在 envs.py 添加 SGLANG_DIFFUSION_FLASHINFER_FP4_GEMM_BACKEND 环境变量；更新 transformer_load_utils.py 优化 FLUX.2 NVFP4 处理；简化 interface.py 移除不再使用的方法。

评论区精华

reviewer mickqian 提出建议：

> may consider adding more nvfp4 test to CI later

这提示未来测试覆盖可增强，但当前 PR 无技术争议或深度讨论。

风险与影响

- 风险：默认后端变更可能在少数形状（12.7/265）导致性能回退，但可通过环境变量 SGLANG_DIFFUSION_NVFP4_LINEAR_BACKEND=comfy 覆盖；新增代码未经过大规模 CI 测试，可能存在隐藏 bug。

- 影响：提升扩散模型在 Blackwell 上的推理速度约 8.2%（基于 FLOP 加权比），为用户带来直接性能收益；基准测试基础设施增强，支持团队后续优化决策。

关联脉络

与历史 PR #22064 紧密相关，后者修复 NVFP4 权重缩放并添加大 M 核配置，本 PR 已对齐其变更。此外，PR #18762 等扩散优化工作显示仓库持续关注 JIT 内核和性能调优，本 PR 延续了这一趋势，强调数据驱动的后端选择策略。