PR #22099 完整报告

sgl-project/sglang

Align diffusion nightly presets and broaden skill discovery

合并时间：2026-04-04 21:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22099

执行摘要

本 PR 对齐了扩散模型的 nightly 基准测试预设和技能文档，以 LTX-2 双阶段模型为中心，扩展了融合 QK norm + RoPE 等优化机会指导。通过更新脚本、文档和删除冗余文件，提升了开发效率，但需注意基准测试逻辑中 denoise 延迟计算的正确性风险。

功能与动机

主要动机是统一扩散模型的 nightly 比较配置和文档，确保优化工作优先利用现有快速路径。PR body 强调围绕 LTX-2 案例对齐配置，并添加缺失机会指导（如 fused QK norm + RoPE 和 Nunchaku fused GELU MLP），以避免重复造轮子并提升性能优化效率。

实现拆解

- 测试组织：重命名 Z-Image 融合调制测试文件到 jit_kernel/tests/diffusion/，提高模块化。
- 技能文档更新：修改多个 SKILL.md 文件（如 sglang-diffusion-benchmark-profile 和 sglang-diffusion-ako4all-kernel），集成 LTX-2 案例、简化技能流并移除未使用的 Triton/CUDA 内核技能，聚焦于基准测试和机会发现。
- 基准测试脚本：在 bench_diffusion_denoise.py 中添加 LTX-2 预设，更新 denoise 延迟计算逻辑以聚合多个阶段（如 LTX2AVDenoisingStage 和 LTX2RefinementStage）。
- CI 集成：扩展 run_comparison.py 的 payload 处理，支持 FPS 和 negative_prompt 等新参数，确保 nightly 比较兼容性。
- 性能分析：修改 profiler.py，添加 _resolve_profiler_log_dir 函数，统一日志目录解析。
- 清理工作：删除未使用的脚本和文档文件（如 nsight-profiler.md 和 bench_diffusion_rmsnorm.py），减少维护负担。

评论区精华

review 中仅有一个评论，聚焦于 bench_diffusion_denoise.py 的 denoise 延迟计算逻辑：

> "The current logic only sets denoise_latency_s if denoise_stage_total_ms is strictly greater than 0.0. While unlikely, if matching stages were found but their durations were exactly 0.0, the script would fall back to denoise_steps_ms. It might be more robust to track if any matching stages were found at all, regardless of their duration, to avoid unnecessary fallback when the primary metric source is present."
> 此讨论点出了潜在的正确性问题，建议改进逻辑以更鲁棒地处理边界情况，但目前尚未解决。

风险与影响

- 技术风险：*denoise* 延迟计算逻辑可能在阶段持续时间为 0 时出错，导致性能指标不准确；文档更新若与代码脱节，可能误导开发者；删除技能文件可能中断依赖这些文件的开发工作流。
- 影响范围：影响扩散模型性能优化团队，通过标准化基准测试和文档，提升协作效率；系统层面，确保 nightly 比较的准确性，促进持续集成；技能流简化可能减少维护开销，但需确保过渡平滑。

关联脉络

本 PR 是扩散模型优化生态的一部分，与历史 PR 如 #21080（FA4 推测解码支持）、#22038（VLM 编码优化）等协同，共同推进 SGLang 在多模态和性能方面的发展。它强化了文档和配置对齐，为后续内核优化提供坚实基础，并反映了团队在提升开发体验和性能可观测性方面的持续投入。