PR #21373 完整报告

sgl-project/sglang

[diffusion] doc: consolidate documentation

合并时间：2026-03-25 16:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21373

执行摘要

本 PR 对 SGLang Diffusion 文档进行大规模重构，整合目录结构，新增使用和参考章节，旨在提升文档可读性和用户导航效率，特别强调 sgl-kernel 在优化推理中的作用，属于常规文档维护工作。

功能与动机

动机源于优化文档组织，PR body 中明确表示要 'consolidate the directory structure' 并 'emphasize the usage of sgl-kernel'，以解决原有文档松散、用户查找困难的问题，通过前后目录结构对比图展示了简化效果。

实现拆解

- 主页重构：docs/diffusion/index.md 重写，精简特性列表，突出 sgl-kernel 和 JIT 内核，移除冗余平台详情。
- CLI 文档简化：docs/diffusion/api/cli.md 移除详细参数列表，改为强调使用 sglang generate --help 作为权威来源，提升实用性。
- 量化文档更新：docs/diffusion/quantization.md 调整示例命令，移除过时参数如 --attention-backend torch_sdpa，以更清晰指南支持量化工作流。
- 新增章节：添加 usage.md 组织日常使用文档，reference.md 收集环境变量等参考内容，形成结构化导航。
- 性能文档调整：docs/diffusion/performance/index.md 简化内容，突出缓存和注意力后端，并更新子章节链接。
- 索引同步：docs/index.rst 更新主索引，移除旧条目，反映新文档结构，确保整体一致性。

评论区精华

无实质性技术讨论，仅 Issue 评论中触发 CI 测试的自动化命令（如 /tag-and-rerun-ci），表明变更已通过自动化检查，并由作者在评论中说明 'pure doc change, bypassing'，确认了文档变更的低风险性质。

风险与影响

风险：文档内部链接可能断裂（如旧路径引用），需在合并后验证；内容准确性需与代码同步，避免示例过时。 影响：显著改善用户体验，帮助新用户快速上手扩散模块，但对系统功能、性能或安全性无直接影响；团队需更新文档引用习惯，以适应新结构。

关联脉络

与 PR 21356（更新 quantization.md）直接相关，两者都修改了同一文件，显示扩散模块文档的持续优化；类似 PR 20846（Ascend 文档更新）反映仓库对文档维护的重视，表明这是一个跨模块的文档重构趋势，旨在提升整体项目可维护性。