PR #21356 完整报告

sgl-project/sglang

[diffusion] doc: update quantization.md

合并时间：2026-03-25 14:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21356

执行摘要

本次 PR 更新了 SGLang-Diffusion 模块的量化文档，新增了 quantization.md 提供详细使用指南和量化家族表，并同步更新了 CLI 文档和索引，移除过时旧文档，确保用户能正确使用量化 transformer 功能。

功能与动机

动机是更新量化文档以反映最新的量化功能支持，例如从历史 PR 20137 中引入的 NVFP4 支持。作者在 issue 评论中提及“pure doc change”，表明这是一个纯文档维护变更，旨在提供更准确、易用的使用指南，减少用户配置错误。

实现拆解

主要变更文件包括：

- docs/diffusion/quantization.md：新增文档，包含快速参考、量化家族（如 FP8、NVFP4、Nunchaku-SVDQ）和具体示例。例如，表格列出了不同量化家族的检查点形式、CLI 用法和平台说明。
- docs/diffusion/api/cli.md：更新 CLI 文档，添加 --transformer-path 和 --transformer-weights-path 等量化参数的说明，强调推荐使用方式。
- docs/diffusion/index.md 和 docs/index.rst：在索引中添加量化文档链接，提升文档可发现性。
- python/sglang/multimodal_gen/docs/quantization.md：删除旧文档，避免信息冗余和潜在误导。

评论区精华

由于没有 review 评论，讨论部分为空。作者在 issue 评论中简要说明这是一个文档变更，并跳过了常规 review 流程，表明团队对纯文档更新采用简化处理。

风险与影响

风险较低：文档内容准确性是关键，但由维护者更新，风险可控；删除旧文档可能影响链接，但新文档已替代，且变更范围小。影响正面：提升用户体验，提供更清晰的使用指南；对系统无代码影响，不引入性能或安全风险。

关联脉络

此 PR 与历史 PR 20137（“[diffusion] Support nvfp4 for Flux.2”）密切相关，后者添加了 NVFP4 量化支持，文档更新可能反映了此功能，显示了量化功能的持续演进。从近期历史 PR 看，quant 和 diffusion 标签常一起出现，表明该模块在积极扩展量化能力，文档更新是功能完善的自然延伸。