PR #25904 完整报告

sgl-project/sglang

:memo: docs(diffusion): add MXFP4 quantization docs

合并时间：2026-05-25 15:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25904

执行摘要

本 PR 为 Diffusion 模块新增 MXFP4 量化文档，同步更新了量化方法表格和 MXFP4 在线 / 离线使用说明，并更新了 Ascend NPU 量化支持矩阵。文档基于已合并的 MXFP4 Diffusion PR（#22338）编写，是跨 PR 的文档配套。

功能与动机

MXFP4 量化已通过 PR #22338 实现，但缺乏文档指导用户使用。PR body 明确说明这是对 #22338 的文档跟进，并遵循 #24918（MXFP8 文档）的更新模式。

实现拆解

1. 更新量化方法表格：在 docs_new/docs/sglang-diffusion/quantization.mdx 的 msmodelslim 支持列表中补充 W4A4_MXFP4 / W4A4_MXFP4_DUALSCALE 和 mxfp4_npu。
2. 新增 MXFP4 Online 量化小节：说明使用 --quantization mxfp4_npu 加载原 FP16/BF16 模型即可在线量化权重，激活值在推理时逐 token 量化。解释了 mxfp4_npu 名称由来（mxfp4 为 ROCm/aiter 保留）。提及双级 block scales：L1=32, L0=512。
3. 新增 MXFP4 Offline 量化小节：描述如何加载 msmodelslim 预量化权重，包括 wan_repack.py 转换、双级 scale（weight_scale, weight_dual_scale）和 mul_scale 的加载逻辑。
4. 更新 Ascend NPU 硬件支持表：在 docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_quantization.mdx 的矩阵中为 Diffusion 增加 MXFP4 Linear 支持行，明确 A5+ 可用。

无需展示，纯文档变更。

评论区精华

无实质 review 讨论，仅有两个 Gemini 机器人的配额警告。最终由 sglang-npu-bot 自动合并。

风险与影响

风险：无代码变更，风险极低。需确保文档中的命令、硬件要求与实际实现一致。 影响：影响 Diffusion 用户和 Ascend NPU 用户，降低 MXFP4 量化使用门槛。

关联脉络

- 关联 PR #22338（MXFP4 实现），本 PR 为其文档配套。
- 遵循 PR #24918（MXFP8 文档）的更新模式，保持文档风格一致。