PR #24918 完整报告

sgl-project/sglang

:memo: docs(diffusion): add MXFP8 quantization docs for Wan2.2 on Ascend NPU

合并时间：2026-05-11 13:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24918

执行摘要

该 PR 为 SGLang Diffusion 模块的 Wan2.2 模型在 Ascend NPU 上的 MXFP8 量化补充了用户文档，将在线 / 离线两种 MXFP8 模式从“进行中”标记为“已支持”，并提供了详细的用法示例和硬件要求。不涉及任何运行时代码变更，风险极低。

功能与动机

动机源自已合并的功能 PR #20922，该 PR 实现了 Wan2.2 在 Ascend NPU 上的 MXFP8 量化支持。当前文档 PR 补充了对应的用户文档，帮助用户了解和使用这些新特性。

实现拆解

1. 更新 quantization.mdx：
   - 在 msmodelslim 量化家族表中加入 mxfp8 条目。
   - 将 wan_repack.py 的使用方式从多步骤工作流简化为一步命令，并增加支持的 --model-type 参数说明。
   - 在可用方法列表中，将 W8A8_MXFP8（离线）和 mxfp8（在线）标记为已完成。
   - 新增“MXFP8 Online Quantization”和“MXFP8 Offline Quantization”两小节，分别介绍在线和离线量化的用法、硬件要求及示例命令。
2. 更新 ascend_npu_quantization.mdx：
   - 将 Ascend A5 系列和 Diffusion 对应的 MXFP8 状态从“WIP”改为“✓”，表明已支持。
3. 测试与部署：无相关修改。

无可用关键源码片段（文档仅包含文本和 Markdown 表格）。

评论区精华

Review 中 gemini-code-assist[bot] 提出了三个涉及 wan_repack.py 和 fp8.py 的代码问题：

> load_sharded_safetensors 不支持分片 checkpoint，会直接报错。 shutil.copytree 缺少 dirs_exist_ok=True 参数，重跑会崩溃。 使用 parameter.data 直接赋值应替换为 copy_()。

作者 TallMessiWu 均回复“No longer applicable”，表明这些代码问题在当前文档变更中已被规避或已由其他 PR 修复，不适用于此。

风险与影响

风险：无。文档变更不产生任何回归风险。

影响：

- 用户：Ascend NPU 用户可直接参考文档使用 MXFP8 量化部署 Wan2.2。
- 系统：无影响。
- 团队：减少对同一问题的用户咨询。

关联脉络

- 关联 PR #20922：实现了 Wan2.2 MXFP8 量化功能，当前 PR 为其补充文档，属于同一功能线的配套更新。