PR #32945 完整报告

sgl-project/sglang

Qwen3.5 NVFP4 V2

合并时间：2026-08-08 06:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32945

执行摘要

该 PR 将 Qwen3.5-397B 在 Blackwell 上的 NVFP4 checkpoint 引用从 nvidia/Qwen3.5-397B-A17B-NVFP4 升级为 nvidia/Qwen3.5-397B-A17B-NVFP4-V2，涉及文档站部署命令生成器与 cookbook 两处共 4 行改动。变更不涉及推理运行时，对系统无行为影响，但会改变用户按文档拉取的模型版本。

功能与动机

PR body 指出需要基于 SemiAnalysisAI/InferenceX 的 PR#2238 更新 Qwen3.5 FP4 Blackwell checkpoint。仓库近期仍在修复 NVFP4 相关后端（如 PR#33543），说明该量化链路仍在演进；文档及时指向上游推荐的 V2 版本，可以避免用户继续使用旧 checkpoint。

实现拆解

1. 部署命令生成器：docs/src/snippets/autoregressive/qwen35-deployment.jsx 中 generateCommand 函数生成 sglang serve 命令。本次将 modelName 的 FP4 分支中 Blackwell 路径改为 nvidia/Qwen3.5-397B-A17B-NVFP4-V2，并同步注释；AMD MI355X 的 MXFP4 分支保持不变。
2. cookbook 指南：docs/cookbook/autoregressive/Qwen/Qwen3.5.mdx 在模型对照表与 FP4 说明段落两处替换 NVFP4 链接为 V2；AMD MXFP4 链接未改动。
3. 配套改动：未新增测试或 CI 配置；改动通过文档站构建与 cookbook 渲染生效，属于纯展示层同步，风险面小。

关键源码片段

docs/src/snippets/autoregressive/qwen35-deployment.jsx

核心改动文件。文档站通过该 JSX 动态生成 Qwen3.5 的 sglang serve 部署命令，modelName 的选择直接决定用户拉取的模型路径；本次将 Blackwell NVFP4 分支指向 V2 checkpoint。

```
// generateCommand 是文档站生成 sglang serve 命令的核心函数。
// 这里聚焦展示 modelName 的 FP4 分支选择逻辑。
let modelName;
if (quantization === 'fp4') {
  // AMD MI355X 使用 MXFP4 checkpoint；Blackwell 使用 NVFP4-V2。
  // 本次更新将旧路径 `nvidia/Qwen3.5-397B-A17B-NVFP4` 改为 `...-NVFP4-V2`，
```

```
    // 对齐上游 InferenceX PR#2238 推荐的发布版本。
    modelName = hardware === 'mi355x'
      ? 'amd/Qwen3.5-397B-A17B-MXFP4'
      : 'nvidia/Qwen3.5-397B-A17B-NVFP4-V2';
  } else {
    // 非 FP4 的 BF16 / FP8 继续按模型系列拼接 Hugging Face 模型名。
    const suffix = MODEL_SUFFIX[model];
    const quantSuffix = quantization === 'fp8' ? '-FP8' : '';
    modelName = `Qwen/Qwen3.5-${suffix}${quantSuffix}`;
  }
```

评论区精华

没有技术性 review 评论，唯一的交互是：

> 提交者 faradawn：Hi @JustinTong0323 @zijiexia can you help comment/approve?

最终由维护者 ishandhanani 直接批准；gemini-code-assist[bot] 的评论仅声明其代码审查服务已停运，无技术价值。

风险与影响

- 一致性风险：本次只覆盖 2 个文件，仓库其他位置可能仍引用旧 NVFP4 名称，建议全仓搜索确认。
- 外部依赖：新 checkpoint 来自 NVIDIA HF 仓库，若上游调整或下架，按文档执行的用户会拉取失败。
- 测试缺口：generateCommand 没有输出快照测试，未来模型名逻辑改动不易被察觉。
- 影响评估：仅影响文档读者，不改变 SGLang 运行时，团队维护成本低。

关联脉络

该 PR 与 PR#33543（Fix Nemotron W4A16 NVFP4 MoE backend）同属 NVFP4 量化支持链路：一个修复后端正确性，一个同步官方 checkpoint 引用，反映仓库对 NVIDIA NVFP4 生态的跟进。上游 InferenceX 的 PR#2238 提供了 V2 的发布原因，可作为后续文档补充说明的参考。