PR #24144 完整报告

sgl-project/sglang

[BugFix][EPD] adapt for qwen3.5-mtp & del duplicated logs

合并时间：2026-05-23 15:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24144

执行摘要

- 一句话：适配 Qwen3.5-MTP 模型，删除冗余日志
- 推荐动作：变更简单清晰，建议合并。值得关注的设计决策是：为 MTP 草稿模型注册多模态处理器的方式是否足够通用，是否需要考虑更多模型变体。

功能与动机

当使用 --speculative-algo NEXTN 运行 Qwen3.5-VL 模型时，模型架构名称会从 Qwen3_5ForConditionalGeneration 转换为 Qwen3_5ForCausalLMMTP（多 Token 预测草稿模型）。然而 Qwen3_5ForCausalLMMTP 并未在 QwenVLImageProcessor.models 中注册，导致多模态处理器查找失败并报错 ValueError: No processor registered for architecture: ['Qwen3_5ForCausalLMMTP']。另外，get_mm_data 中存在重复的 debug 日志（Get embedding slice for {modality}），增加了不必要的噪音。

实现拆解

1. 处理器注册：在 python/sglang/srt/multimodal/processors/qwen_vl.py 的导入部分增加 from sglang.srt.models.qwen3_5_mtp import Qwen3_5ForCausalLMMTP，并在 QwenVLImageProcessor 类的 models 列表中添加 Qwen3_5ForCausalLMMTP。这样当遇到该架构时，多模态处理器可以正确查找到 QwenVLImageProcessor。
2. 日志清理：删除 get_mm_data 方法中的 logger.info(f"Get embedding slice for {modality}, num_tokens={num_tokens}") 行。该日志在每次切片嵌入时打印，属调试信息，对生产环境无意义，移除后减少日志噪音。
3. 代码合并：第二个 commit 将 main 分支合并到特性分支，确保与主线兼容。

关键文件：

- python/sglang/srt/multimodal/processors/qwen_vl.py（模块 多模态；类别 source；类型 dependency-wiring）：唯一变更文件，包含所有修改：添加导入、注册新模型、删除冗余日志。

关键符号：未识别

关键源码片段

python/sglang/srt/multimodal/processors/qwen_vl.py

唯一变更文件，包含所有修改：添加导入、注册新模型、删除冗余日志。

```
# python/sglang/srt/multimodal/processors/qwen_vl.py

from sglang.srt.models.qwen3_5_mtp import Qwen3_5ForCausalLMMTP # 新增导入

class QwenVLImageProcessor(SGLangBaseProcessor):
    models = [
        Qwen2VLForConditionalGeneration,
        Qwen2_5_VLForConditionalGeneration,
        Qwen3VLForConditionalGeneration,
        Qwen3VLMoeForConditionalGeneration,
        Qwen3_5ForConditionalGeneration,
        Qwen3_5MoeForConditionalGeneration,
        Qwen3_5ForCausalLMMTP, # 新增注册，解决 MTP 草稿模型无处理器问题
        InternS2PreviewForConditionalGeneration,
        Qwen3OmniMoeForConditionalGeneration,
    ]

    def get_mm_data(self, prompt, embeddings, **kwargs):
        # ... 省略上下文 ...
        for modality, num_tokens in self.mm_tokens.items():
            # logger.info(f"Get embedding slice for {modality}, num_tokens={num_tokens}") #
            已删除，避免日志噪音
            mm_items.append(...)
```

评论区精华

无实质性讨论。gemini-code-assist[bot] 自动进行了评论但未提出修改建议，liusy58 和 ShangmingCai 均直接批准了 PR，无未解决疑虑。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅包含一行导入、一行列表添加和一行日志删除，不涉及核心逻辑改动。但有两点可注意：
  - 若未来 Qwen3_5ForCausalLMMTP 的配置与其他 Qwen3.5 变体有差异，当前的简单注册可能不满足要求，但当前场景下该草稿模型共享相同的视觉配置。
  - 缺少直接针对该场景的自动化测试，回归风险存在于推测解码与 Qwen3.5-VL 的集成路径中。
  - 影响：用户影响：修复了用户在使用 --speculative-algo NEXTN 运行 Qwen3.5-VL 模型时遇到的 ValueError，使该功能可用。同时日志噪音减少，生产环境日志更干净。 系统影响：仅改动一个文件，对系统其他部分无影响。 团队影响：无。
- 风险标记：暂无

关联脉络

- PR #25843 Route concat MLA to JIT and remove unused downcast: 同样涉及 deepseek 和 JIT kernel，但与本 PR 无直接关联。
- PR #26126 [RL] [Spec v2] Use stop-aware seqlen for returned topk metadata: 同属 speculative decoding 修复系列，但关注点不同。