PR #22269 完整报告

sgl-project/sglang

[EPD][VLM] Support Kimi K25 EPD

合并时间：2026-04-10 10:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22269

执行摘要

本 PR 为 Kimi K2.5 视觉语言模型添加了 Encoder-Prefill-Decode (EPD) 解耦支持，通过修改编码器服务器、模型加载器和处理器逻辑，解决了该模型在分布式推理中的适配问题，扩展了 sglang 的多模态生态，提升了 EPD 架构的通用性。

功能与动机

EPD disaggregation 允许视觉编码和语言处理在分离的实例上运行，以优化资源利用。Kimi K2.5 模型此前未完全集成到此路径中，导致编码器无法加载检查点子集、网格元数据不匹配、令牌计数错误等问题。根据 PR body 描述，本 PR 旨在“填补这些空白”，使 Kimi K2.5 能端到端运行在 --encoder-only / --language-only 模式下，与现有 VLM 保持一致的 EPD 流程。

实现拆解

1. 编码器服务器 (encode_server.py)：
   - 添加 grid_thws 到网格属性列表，并为 kimi_k25 模型优先处理此属性。
   - 新增 _kimi_k25_tokens_from_patch_grid 函数，根据 merge_kernel_size 计算令牌数。
   - 传递 model_type 参数到 _get_mm_grid_dim 以支持模型特定逻辑。
2. 模型实现 (models/kimi_k25.py)：
   - 支持 encoder_only 和 language_only 模式，条件初始化 language_model。
   - 修复 get_image_feature 中的设备 / 数据类型对齐问题。
   - 强化属性访问（如 start_layer）以处理语言模型缺失情况。
3. 处理器 (multimodal/processors/kimi_k25.py)：
   - 新增 get_mm_data 函数，扩展图像占位符并附加预计算嵌入。
4. 服务器参数 (server_args.py)：
   - 扩展 EPD 验证，将 KimiVLForConditionalGeneration 和 KimiK25ForConditionalGeneration 加入允许列表。

评论区精华

review 中，gemini-code-assist[bot] 提出了关键设计疑虑：

- 关于硬编码逻辑：“The logic for selecting grid attributes is hardcoded for 'kimi_k25'. If other models require similar custom attribute ordering, this will become

unmaintainable.”

- 关于属性访问：“Using nested getattr calls for language_model is fragile. It is cleaner to check hasattr(self, 'language_model').” 这些评论被标记为中等优先级，但 PR 在后续提交中通过代码简化部分改进，最终 CI 通过并获得批准，未解决的设计问题留作未来优化。

风险与影响

- 技术风险：硬编码模型逻辑可能增加维护成本；设备对齐错误可能引发运行时异常；令牌计数依赖模型配置，变化可能影响准确性。
- 影响评估：用户现在可以使用 Kimi K2.5 进行 EPD 推理，提升分布式效率；系统扩展了 VLM 支持，但需关注模型特定适配的代码复杂度；团队需确保新增逻辑的测试覆盖和文档更新。

关联脉络

从近期历史 PR 看，本 PR 是 sglang 多模态和 EPD 架构持续演进的一部分。例如，PR 22089 为 Qwen3-ASR 添加流式 ASR 功能，同样涉及多模态扩展；PR 22329 为 AMD 平台添加 EPD 相关环境变量。这些 PR 共同推动着分布式推理和模型生态的完善，本 PR 通过支持 Kimi 模型进一步丰富了 EPD 的应用场景。