PR #21230 完整报告

sgl-project/sglang

Add LFM2-VL (Liquid Foundation Model 2 Vision-Language) support

合并时间：2026-04-04 16:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21230

执行摘要

本 PR 为 SGLang 添加了 LFM2-VL 视觉语言模型支持，通过集成 SigLip2 视觉编码器的 NaFlex可变分辨率特性和LFM2混合语言模型，扩展了多模态能力。变更包括新增配置、模型、处理器文件，并修改现有代码以确保兼容性。测试显示模型在 ROUGE-L 和 logprob divergence 上表现一致，建议工程师关注混合缓存设计和处理器重构以学习集成模式。

功能与动机

添加 LFM2-VL 模型支持，旨在丰富 SGLang 的多模态模型生态系统。PR body 中明确说明：'This PR adds support for the LFM2-VL vision-language architecture, combining a SigLip2 vision encoder (NaFlex variable-resolution) with the LFM2 hybrid language model.'示例模型为LFM2.5-VL-1.6B，支持多图像输入，以提升视觉语言推理任务的覆盖范围。

实现拆解

- 配置层：新增 python/sglang/srt/configs/lfm2_vl.py，定义混合缓存属性，如 full_attention_layer_ids 和 linear_layer_ids，用于 KV 和卷积状态管理。
- 模型层：新增 python/sglang/srt/models/lfm2_vl.py 实现多模态投影器（Lfm2VlMultiModalProjector）和整体模型；新增 python/sglang/srt/models/siglip2.py 实现 SigLip2 视觉编码器，支持 NaFlex 可变分辨率。
- 处理器层：新增 python/sglang/srt/multimodal/processors/lfm2_vl.py，经重构使用基类 SGLangBaseProcessor，简化图像数据处理。
- 集成层：修改 python/sglang/srt/models/lfm2.py 重命名 inputs_embeds 为 input_embeds 并添加 get_input_embeddings() 方法；更新 python/sglang/srt/model_executor/model_runner.py 以检测 LFM2-VL 为混合模型；注册配置和更新文档。

评论区精华

- 性能优化：mickqian 在 lfm2_vl.py 中建议：'nit: we could bring this line ahead to achieve better perf'，关注代码顺序对性能的影响。
- 设计简化：mickqian 在 multimodal/processors/lfm2_vl.py 中指出：'we are deprecating these functions... Could you look at qwen_vl.py for example?'，推动使用基类模式，作者响应并重构代码。

风险与影响

- 风险：新模型集成可能引入回归，特别是 lfm2_vl.py 中的混合缓存逻辑；siglip2.py 的 NaFlex 实现复杂性可能影响稳定性；处理器重构需确保与现有多模态模型兼容。
- 影响：用户可访问新模型，提升多模态任务能力；系统增加新组件，需额外维护；团队需确保设计一致性，避免架构分裂。

关联脉络

与近期 PR 如 22038（VLM 优化）和 20707（diffusion 新特性）相关，显示 SGLang 在多模态和模型支持上的持续演进。此 PR 进一步扩展了视觉语言模型支持，与历史 PR 中的 multimodal 和 feature 标签一致，揭示生态系统扩展趋势。