PR #18467 完整报告

sgl-project/sglang

VLM: support passing --mm-process-config for all models

合并时间：2026-04-12 17:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/18467

执行摘要

该 PR 修复了多模态语言模型中 --mm-process-config 参数仅对 Qwen VL 有效的 bug，通过按模态分离配置并使用 HuggingFace 的 kwargs 路由，避免了参数冲突。影响范围覆盖所有 VLM，提升了配置灵活性和系统健壮性。

功能与动机

动机源于 issue #14672，用户报告 --mm-process-config 参数设置后未生效。PR body 指出，之前尝试（PR #14968）因使用 kwargs.update() 导致参数冲突，本 PR 旨在彻底解决此问题，确保配置正确传递给所有模型。

实现拆解

- 配置验证：在 server_args.py 添加 _handle_multimodal() 方法，验证 mm_process_config 结构。
- 基类处理：在 base_processor.py 提取 image_config、video_config、audio_config，并使用 setdefault().update() 注入为 images_kwargs 等。
- 特定处理器：更新 ernie45_vl.py 和 midashenglm.py 以保持一致性，移除 qwen_vl.py 重复代码。
- 测试与文档：新增测试文件 test_mm_process_config.py，更新文档说明配置传递机制。

评论区精华

- gemini-code-assist[bot] 建议代码优化：" 使用局部变量提高可读性 "，被采纳。
- yuan-luo 提问：" 为何 qwen_vl 无需改动？" edwingao28 解释：" 因继承基类，而 ernie45_vl 覆盖方法需显式添加。"
- mickqian 建议：" 提取验证逻辑到单独函数 "，edwingao28 实施。

风险与影响

风险低：回归风险通过测试覆盖缓解；性能开销可忽略；兼容性改变是预期修复。影响正面：用户可灵活配置多模态处理；系统更可靠；团队代码更易维护。

关联脉络

直接关联 issue #14672；历史 PR 中，PR #22361（Whisper 批量编码）同为多模态优化，体现该领域持续改进趋势。