PR #21899 完整报告

sgl-project/sglang

[VLM] Enable per-image MM splitting by default and remove MULTI_IMAGES modality

合并时间：2026-04-03 11:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21899

执行摘要

- 一句话：默认启用多模态图像分裂，移除 MULTI_IMAGES 模态，提升缓存命中率。
- 推荐动作：建议技术管理者和工程师精读此 PR，重点关注 _try_simple_split 函数的实现和处理器层的变更，以理解多模态缓存优化设计；同时注意向后兼容性风险和后续 ViT 优化方向，可作为多模态性能调优的参考案例。

功能与动机

PR body 中明确指出，动机是移除 SGLANG_ENABLE_MM_SPLITTING 环境变量，使多模态输入分裂默认开启；移除 MULTI_IMAGES 枚举；每个图像现在作为独立的 MultimodalDataItem，具有独立的哈希和 pad_value，从而实现每图像 RadixAttention 缓存。这改善了缓存重用，在部分共享图像前缀场景中缓存命中率从 0.0% 提升到 41.5%，多轮图像场景从 0.1% 提升到 78.1%。

实现拆解

实现方案按模块拆解：1) 环境配置层：在 environ.py 中移除 SGLANG_ENABLE_MM_SPLITTING 环境变量；2) 数据结构层：在 schedule_batch.py 中移除 Modality.MULTI_IMAGES 枚举，更新 MultimodalDataItem 类定义；3) 多模态管理器：在 mm_utils.py 中添加 _try_simple_split 函数作为回退逻辑，修改 pad_input_tokens 以处理每图像项；4) 处理器层：在 base_processor.py 中集成 get_new_expanded_mm_items 调用，并更新 llava.py、internvl.py、minicpm.py 等处理器直接创建每图像项；5) 模型层：调整 llava.py 和 minicpmv.py 的模型逻辑以适应每图像处理，如新增 _infer_image_aspect_ratio 函数。

关键文件：

- python/sglang/srt/environ.py（模块 环境配置）：移除了 SGLANG_ENABLE_MM_SPLITTING 环境变量，使多模态分裂行为默认开启，简化配置。
- python/sglang/srt/managers/mm_utils.py（模块 多模态管理器）：添加 _try_simple_split 回退函数并修改 pad_input_tokens 逻辑，是实现每图像分裂和缓存处理的 core 组件。
- python/sglang/srt/managers/schedule_batch.py（模块 调度批处理）：移除 Modality.MULTI_IMAGES 枚举，更新 MultimodalDataItem 类定义，影响整个多模态数据结构。

- python/sglang/srt/multimodal/processors/base_processor.py（模块 多模态处理器）：在 process_and_combine_mm_data 中集成 get_new_expanded_mm_items 调用，将分裂逻辑移至处理器层，是变更的关键枢纽。
- python/sglang/srt/models/llava.py（模块 模型）：更新 LLaVA 模型逻辑以处理每图像项，新增 _infer_image_aspect_ratio 函数，展示模型适配变更。

关键符号：pad_input_tokens, get_new_expanded_mm_items, _try_simple_split, _infer_image_aspect_ratio, process_and_combine_mm_data

评论区精华

无 review 评论，变更通过 6 个 commit 演进完成：从初始更新到启用分裂、移除枚举、修复问题、改进逻辑，表明迭代开发过程，但未涉及设计争议。

- 暂无高价值评论线程

风险与影响

- 风险：技术风险具体包括：1) 向后兼容性：移除 MULTI_IMAGES 枚举可能破坏依赖该枚举的旧代码，如模型检查或第三方集成；2) 分裂逻辑稳健性：_try_simple_split 函数依赖特征形状匹配，若模型特征格式异常可能导致分裂失败或缓存错误；3) 性能回归：新缓存逻辑在处理器层分裂可能增加内存开销，且未完全解决 ViT 冗余计算（如 PR body 所述），可能影响 TTFT；4) 测试覆盖：尽管修改了 test_mm_utils.py，但变更涉及多个处理器和模型文件，可能缺乏全面集成测试。
- 影响：影响范围：用户端：提高多模态请求的缓存命中率，理论上降低首次令牌时间（TTFT），改善响应性能；系统端：优化多模态处理效率，减少 KV 缓存冗余，但可能因每图像项增多而轻微增加内存管理复杂度；团队端：简化代码结构，移除过时逻辑，促进后续每图像 ViT 跳过优化（如 PR body 提到的跟进 PR）。影响程度：中等，主要限于多模态模块，但缓存改进可显著提升特定场景性能。
- 风险标记：向后兼容性破坏，分裂逻辑潜在 bug，缓存变更风险

关联脉络

- PR #19163 [Feature] Stronger transformers modeling backend with TP, PP, MoE, VLMs, and torch compile: 涉及多模态模型支持，与本 PR 的多模态优化主题相关，共同扩展 SGLang 的多模态能力。
- PR #21892 Skip broken AutoModel mapping entries when resolving Llava submodules: 涉及 Llava 多模态模型加载修复，与本 PR 中 Llava 处理器修改有功能关联。
- PR #21955 [diffusion] chore: fix stage profiler for multi-stage denoising: 涉及多模态生成性能分析，与本 PR 的缓存性能优化有间接主题关联。