

# PR #31957 完整报告

sgl-project/sglang

fix(vlm): reject Moss vision metadata mismatches

合并时间: 2026-07-29 07:09

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31957>

## 执行摘要

- 一句话: 拒绝 Moss-VL 视觉元数据与 token 计数不匹配
- 推荐动作: 此 PR 值得精读, 是典型「静默失败 → 快速失败」的改进实践, 展示了 fail-closed 的设计决策。测试覆盖了异常的多种边界, 可作同类修复的参考。

## 功能与动机

`_compute_vision_position_ids` 之前使用 `min` 取元数据和 token 计数的小值, 并在任一方为零时提前返回, 静默构建部分或空的 vision 位置元数据, 掩盖了上游数据不匹配的问题, 可能导致推理结果异常。PR body 明确要求 'require Moss-VL vision frame metadata to map one-to-one to image tokens'。

## 实现拆解

1. 在 `python/sglang/srt/multimodal/processors/moss_vl.py` 的 `_compute_vision_position_ids` 方法中, 将原 `if len(flat_eff_h) == 0 or len(image_token_indices) == 0` 的提前返回逻辑改为先计算 `frame_count` 和 `image_token_count`, 当两者不等时直接抛出 `ValueError` (含具体计数信息)。
2. 移除 `num_matches = min(...)` 及后续切片, 改为直接使用完整的 `flat_eff_h`、`flat_eff_w`、`flat_vis_starts` 和 `image_token_indices`, 因为现在已保证计数一致。
3. 新增 `test/registered/unit/models/test_moss_vl_processor.py` 测试文件, 通过参数化测试覆盖 `frame_count` 多于、少于、等于 token 数以及正整数匹配的场景, 验证正确行为和拒绝行为。

关键文件:

- `python/sglang/srt/multimodal/processors/moss_vl.py` (模块 多模态处理器; 类别 source; 类型 core-logic; 符号 `_compute_vision_position_ids`): 核心修改文件, 修复了静默截断的 bug, 改为严格校验并抛出异常。
- `test/registered/unit/models/test_moss_vl_processor.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `_vision_info`, `_processor`, `test_moss_vl_rejects_vision_metadata_token_mismatch`, `test_moss_vl_accepts_matching_vision_metadata_and_tokens`): 新增测试文件, 覆盖了各种不匹配场景和匹配场景, 确保行为正确。

关键符号: `_compute_vision_position_ids`

## 评论区精华

无 review 评论，仅作者提交后触发了 CI 并合并。

- 暂无高价值评论线程

## 风险与影响

- 风险：核心风险在于调用方未处理 `_compute_vision_position_ids` 可能抛出的 `ValueError`，导致整个请求处理中断。但鉴于上游应该是内部调用，该异常会沿调用栈传播到请求处理入口，返回 500 错误，比静默错误更安全。此外，`if frame_count == 0` 的提前返回路径保留，仅当两者同为 0 时跳过计算，逻辑正确。无性能退化，只是将 `min+` 切片 替换为简单比较。
- 影响：影响范围限定在 Moss-VL 模型视觉处理路径。对于正常的匹配数据，行为不变。对于不匹配的请求，之前可能构建错误位置编码，现在会直接报错拒绝请求，让用户发现数据准备问题。对系统整体稳定性和可调试性有正面影响。
- 风险标记：核心路径变更，错误处理变更，测试覆盖新增

## 关联脉络

- PR #28370 Fix invalid escape warnings in tool parsers: 同样是对静默问题的修复，体现了项目中类似 fail-closed 的改进趋势。
- PR #31417 Return 400 instead of 500 for unfetchable or unparseable multimodal inputs: 也在多模态处理中改进了错误处理策略，使错误更明确。