

PR #45244 完整报告

vllm-project/vllm

minicpmv4_6: fix ImageSize (W,H) order for placeholder token calculation

合并时间: 2026-06-11 21:43

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45244>

执行摘要

- 一句话: 修复 MiniCPM-V 4.6 图片尺寸顺序错误
- 推荐动作: 值得合并。修复清晰, 风险低, 影响面小。

功能与动机

PR body 指出: `_compute_visual_tokens` 将 vLLM 的 `ImageSize(width, height)` 直接传递给期望 `(height, width)` 顺序的 transformers 处理器方法, 导致某些宽高比 (如 365x211) 下占位符 token 数量与视觉嵌入数量不匹配, 触发 `ValueError`。

实现拆解

1. 在 `minicpmv4_6.py` 的 `_compute_visual_tokens` 方法中, 将传入的 `image_size` 转换为 `(height, width)` 元组 `hf_image_size`, 然后传递给所有 HuggingFace 处理器方法 (`get_sliced_grid`、`find_best_resize`、`get_refine_size`)。
2. 在 `minicpmv.py` 中, 删除了版本 (4, 6) 的冗余分支代码, 因为现在所有 MiniCPM-V 4.6 的逻辑已由 `MiniCPMV4_6ProcessingInfo` 的覆盖方法处理。

关键文件:

- `vllm/model_executor/models/minicpmv4_6.py` (模块 模型实现; 类别 source; 类型 data-contract; 符号 `_compute_visual_tokens`): 核心修复文件, 修复了 `_compute_visual_tokens` 中图片尺寸顺序错误。
- `vllm/model_executor/models/minicpmv.py` (模块 模型实现; 类别 source; 类型 refactor; 符号 `get_slice_image_placeholder`, `get_num_image_tokens`): 删除了版本 (4, 6) 的冗余分支, 这些逻辑现在由 `MiniCPMV4_6ProcessingInfo` 覆盖。

关键符号: `_compute_visual_tokens`, `get_slice_image_placeholder`, `get_num_image_tokens`

关键源码片段

`vllm/model_executor/models/minicpmv4_6.py`

核心修复文件, 修复了 `_compute_visual_tokens` 中图片尺寸顺序错误。

```
# vllm/model_executor/models/minicpmv4_6.py (head)
```

```
def _compute_visual_tokens(
```

```

self,
image_size,
max_slice_nums: int | None = None,
downsample_mode: str | None = None,
) -> tuple[list[int], int, int]:
    """Compute grid, source_image_visual_tokens and patch_visual_tokens."""
    image_processor = self.get_image_processor()
    if max_slice_nums is None:
        max_slice_nums = image_processor.max_slice_nums

    patch_size = image_processor.patch_size
    scale_res = image_processor.scale_resolution
    downsample_mode = self._get_downsample_mode(downsample_mode)
    token_divisor = 4 if downsample_mode == "4x" else 16

    # vLLM ImageSize is (width, height); transformers expects (height, width)
    hf_image_size = (image_size.height, image_size.width)

    # transformers v5.7+ requires `scale_resolution` arg
    try:
        grids = image_processor.get_sliced_grid(
            hf_image_size,
            max_slice_nums,
            scale_res,
        )
    except TypeError:
        grids = image_processor.get_sliced_grid(
            hf_image_size,
            max_slice_nums,
        )

    if grids is None:
        best_size = image_processor.find_best_resize(
            hf_image_size,
            scale_res,
            patch_size,
            allow_upscale=True,
        )
        source_tokens = (
            best_size[0] * best_size[1] // (patch_size * patch_size * token_divisor)
        )
        return [0, 0], source_tokens, 0

    best_resize = image_processor.find_best_resize(
        hf_image_size,
        scale_res,
        patch_size,
    )
    source_tokens = (

```

```
        best_resize[0] * best_resize[1] // (patch_size * patch_size * token_divisor)
    )
    refine_size = image_processor.get_refine_size(
        hf_image_size,
        grids,
        scale_res,
        patch_size,
        allow_upscale=True,
    )
    patch_w = refine_size[0] // grids[0]
    patch_h = refine_size[1] // grids[1]
    patch_tokens = patch_w * patch_h // (patch_size * patch_size * token_divisor)
    return grids, source_tokens, patch_tokens
```

评论区精华

PR 只有一个评论，作者请求合并。没有实质性的讨论。

- 暂无高价值评论线程

风险与影响

- 风险：仅限于 MiniCPM-V 4.6 模型，影响面小。更改简单，将 ImageSize 转换为正确的顺序，没有引入新的逻辑。删除的代码是死代码（已被覆盖），风险低。
- 影响：影响 MiniCPM-V 4.6 模型用户，特别是使用非正方形图片的用户。修复了特定宽高比下的崩溃问题，对其他宽高比无影响。
- 风险标记：暂无

关联脉络

- PR #44980 Fix/minicpmv46 missing version: 同样修复 MiniCPM-V 4.6 模型相关问题。