PR #26094 完整报告

sgl-project/sglang

[VLM] fix: fix only the grids from last split mm item is collected for qwen-vl

合并时间：2026-05-25 09:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26094

执行摘要

- 一句话：修复 Qwen-VL 多模态 grid 收集只取最后一项
- 推荐动作：该 PR 属于重要的 bug 修复，特别是对多图或视频帧场景。建议仔细审阅 _concat_mm_item_grid 中 _as_grid_batch 的处理逻辑，并确认 CI Extra 失败是否与此变更相关。另外，Review 中关于维度一致性的建议值得参考，虽未采纳，但可在后续测试中关注。

功能与动机

在 Qwen-VL 多模态输入被 split 成多个 MultimodalDataItem 时，原有的 compute_mrope_positions 方法只检查最后一个 item 的 model_specific_data，导致非最后一项的 image/video grid 丢失，影响后续 MRoPE 位置编码计算的正确性。PR 通过引入统一的 grid 收集方法来彻底修复此问题。

实现拆解

1. 新增 _concat_mm_item_grid 类方法：遍历 mm_items 列表，根据 modality 类型过滤出对应 model_specific_data 中的 grid 值，利用已有的 _as_grid_batch 确保维度统一，最后拼接所有 grid 为 batch 返回。
2. 重构 compute_mrope_positions：移除原有的逐 item 手动检查赋值，直接调用 _concat_mm_item_grid 分别获取 image_grid_thw 和 video_grid_thw，使逻辑简洁且正确。
3. 新增 _get_grid_from_output_or_items 类方法：提供一个回退链：优先从 processor output 获取，失败则从 mm_items 拼接，最后可从原始 input_data 兜底。
4. 清理 process_mm_data_async：移除重复的 grid 获取代码，改为调用 _get_grid_from_output_or_items 统一处理；同时将视频预处理条件从 base_output.videos 改为显式检查非 dict 类型，避免误处理预计算元数据。另外，将两处 getattr 改为 _get_processor_output_value 以保证函数式风格和 None 安全性。

关键文件：

- python/sglang/srt/multimodal/processors/qwen_vl.py（模块 多模态；类别 source；类型 core-logic；符号 _concat_mm_item_grid, _get_grid_from_output_or_items）：核心变更文件，修复 grid 收集 bug，新增两个辅助方法，并清理相关逻辑。

关键符号：_concat_mm_item_grid, _get_grid_from_output_or_items, compute_mrope_positions

关键源码片段

python/sglang/srt/multimodal/processors/qwen_vl.py

核心变更文件，修复 grid 收集 bug，新增两个辅助方法，并清理相关逻辑。

```
    def compute_mrope_positions(self, input_ids, mm_items):
        # 旧实现只取最后一个 item 的 grid，现在改为遍历所有 split item 并拼接
        image_grid_thw = self._concat_mm_item_grid(
            mm_items, "image_grid_thw", Modality.IMAGE
        )
        video_grid_thw = self._concat_mm_item_grid(
            mm_items, "video_grid_thw", Modality.VIDEO
        )

        input_ids_tensor = torch.tensor(input_ids, dtype=torch.long).unsqueeze(0)
        mrope_positions, mrope_position_delta = MRotaryEmbedding.get_rope_index(
            spatial_merge_size=self.hf_config.vision_config.spatial_merge_size,
            image_token_id=self.mm_tokens.image_token_id,
            video_token_id=self.mm_tokens.video_token_id,
            vision_start_token_id=self.vision_start_token_id,
            model_type=self.model_type,
            tokens_per_second=getattr(
                self.hf_config.vision_config, "tokens_per_second", None
            ),
            input_ids=input_ids_tensor,
            image_grid_thw=image_grid_thw,
            video_grid_thw=video_grid_thw,
        )
        return mrope_positions.squeeze(1), mrope_position_delta

    @classmethod
    def _concat_mm_item_grid(cls, mm_items: list[MultimodalDataItem], key, modality):
        """Collect grids of a specific modality from all mm_items and concat."""
        grids = []
        for item in mm_items:
            if not item.is_modality(modality):
                continue
            # 利用现有 _as_grid_batch 保证形状为 (N, 3) 等
            grid = cls._as_grid_batch(item.model_specific_data.get(key))
            if grid is not None:
                grids.append(grid)
        if not grids:
            return None
        if len(grids) == 1:
            return grids[0]
        return torch.cat(grids, dim=0)

    @classmethod
    def _get_grid_from_output_or_items(
```

```
        cls, ret, mm_items, key, modality, input_data=None
    ):
        """Fallback chain: output -> mm_items -> raw input_data."""
        grid = cls._get_processor_output_value(ret, key)
        if grid is None:
            grid = cls._concat_mm_item_grid(mm_items, key, modality)
        if grid is None and input_data and isinstance(input_data[0], dict):
            grid = input_data[0].get(key)
        # 注意：review 建议可在此处加 _as_grid_batch 以统一维度，但当前未采用
        return grid
```

评论区精华

Review 中出现一条值得关注的讨论：gemini-code-assist[bot] 指出 _get_grid_from_output_or_items 方法在从 ret 或 input_data[0] 返回值时可能返回 1D tensor，而下游 MRotaryEmbedding.get_rope_index 期望 batch 维度，建议用 _as_grid_batch 包装最终结果以确保维度一致。该建议未被采纳，推测是因为 ret 返回的值本身已经过 processor 处理带有 batch 维，且 _get_processor_output_value 和 _as_grid_batch 已在内部保证。

- 返回维度一致性 (correctness): 未采纳，推测因上游 processor output 已保证 batch 维度，或 _as_grid_batch 在调用点已处理。

风险与影响

- 风险：核心风险来自 grid 维度不匹配：如果 _concat_mm_item_grid 中 _as_grid_batch 处理不当或遇到空输入，可能返回 None 导致下游 torch.cat 失败。但新增的逻辑已通过 if not grids: return None 做保护，且 get_rope_index 内部已有 None 处理，风险较低。另外，视频预处理条件放宽可能影响部分预计算路径，但已通过 isinstance 检查区分，不会意外跳过。
- 影响：影响范围局限于 Qwen-VL 系列模型的多模态推理路径。修复后，涉及多个图像 / 视频帧的请求（如视频理解、多图输入）将正确生成所有 grid，确保 MRoPE 位置编码正确，从而提升多模态任务的精度。对单图像 / 视频请求无影响。测试覆盖：当前 PR 未包含专门的测试用例，但 CI 中 Extra 测试失败，可能与变更有关。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #26101 [VLM] accept precomputed multimodal metadata: 同样涉及 Qwen-VL 预处理元数据提取，本 PR 的 _get_grid_from_output_or_items 可视为对其的补充。
- PR #26149 [VLM] feat: accept grid_thws from preprocessed metadata for kimi: 类似思路：从预计算元数据中提取 grid。