PR #51780 完整报告

vllm-project/vllm

[Model] Enable tower and connector LoRA for Keye

合并时间：2026-08-11 18:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/51780

执行摘要

- 一句话：Keye 增加 tower 与 connector 的 LoRA token 数辅助方法
- 推荐动作：值得快速阅读：这是一个小而清晰的模式，展示了如何在多模态模型中桥接语言模型 token 计数与塔 / 连接器层输入长度，适合作为实现 #31479 同类功能的参考。如果读者不关心 Keye 或多模态 LoRA，可不必精读。

功能与动机

关联 Issue #31479 指出：多模态语言模型中的 token 数不一定与 vision tower 或 connector 中线性层所需的输入长度一致，而 lora_mapping 需要在激活之前知道精确的 token 数，因此这些辅助函数是必要的，用于弥合差异并计算出正确长度。本 PR 为 Keye 模型补齐这两个函数，属于该 issue 的后续落地。

实现拆解

1. 实现入口：在 vllm/model_executor/models/keye.py 的 BaseKeyeModule 基类中新增两个方法。
2. get_num_mm_encoder_tokens(num_image_tokens)：读取 self.config.vision_config.spatial_merge_size，返回 num_image_tokens * merge_size**2，把语言模型看到的合并后图像 token 数放大回 vision tower 所需的未合并 patch 序列长度。
3. get_num_mm_connector_tokens(num_vision_tokens)：返回 num_vision_tokens // merge_size**2，使 mlp_AR projector 侧 LoRA 层能按合并块数量对齐激活长度。
4. 继承设计：方法定义在基类上，KeyeForConditionalGeneration 与 KeyeVL1_5ForConditionalGeneration 自然继承，避免重复实现。
5. 测试配套：初版提交包含 tests/models/multimodal/processing/test_keye.py（31 行），reviewer 要求移除，作者遵从；最终没有自动化测试落地。

关键文件：

- vllm/model_executor/models/keye.py（模块 模型层；类别 source；类型 data-contract；符号 get_num_mm_encoder_tokens, get_num_mm_connector_tokens）：唯一变更文件，在 BaseKeyeModule 上新增两个 token 数转换方法，是 Keye tower/connector LoRA 支持的核心契约。

关键符号：get_num_mm_encoder_tokens, get_num_mm_connector_tokens

关键源码片段

vllm/model_executor/models/keye.py

唯一变更文件，在 BaseKeyeModule 上新增两个 token 数转换方法，是 Keye tower/connector LoRA 支持的核心契约。

```
class BaseKeyeModule(nn.Module):
    # 基类：KeyeForConditionalGeneration 与 KeyeVL1_5ForConditionalGeneration
    # 都继承这两个 token 数转换方法，供 LoRA 激活映射使用。
    def get_num_mm_encoder_tokens(self, num_image_tokens: int) -> int:
        # 语言模型看到的是 spatial merge 之后的图像 token 数；
        # 视觉 tower (.visual.) 工作在未合并的 patch 序列上，
        # 编码器侧输入长度需按 merge_size**2 值放大回去。
        merge_size = self.config.vision_config.spatial_merge_size
        return num_image_tokens * merge_size**2

    def get_num_mm_connector_tokens(self, num_vision_tokens: int) -> int:
        # mlp_AR 投影在每个 spatial_merge_size × spatial_merge_size
        # 的块上做线性变换，连接器侧长度是 vision token 数
        # 除以块内 patch 数的整数结果。
        merge_size = self.config.vision_config.spatial_merge_size
        return num_vision_tokens // merge_size**2
```

评论区精华

评审线程集中在新增测试文件上：reviewer jeejeelee 评论“please remove this test”，作者 liushujia122 回复“Removed the standalone test as requested. The implementation is unchanged, and the relevant pre-commit hooks and git diff check pass.”。该测试最终从提交中移除，jeejeelee 随后以“LGTM”批准合并。Claude bot 因 PR 来自 fork 而跳过自动审查。

- 移除初版新增的 test_keye.py 测试 (testing)：测试文件从最终提交中移除，PR 仅保留 keye.py 的 8 行新增源码。

风险与影响

- 风险：风险点集中在三处：
  - 缺少自动化测试：初版临时测试被移除后，最终 PR 未保留任何测试，后续 Keye 的视觉配置若调整（如 spatial_merge_size 语义变化），这两个换算函数可能被悄悄破坏。
  - 整除截断依赖：num_vision_tokens // merge_size**2 依赖输入恰好是 merge_size**2 的倍数，否则会静默截断而非报错。
  - 配置字段依赖：方法读取 self.config.vision_config.spatial_merge_size，如果未来出现缺失该字段的 Keye 变体，会抛出 AttributeError。 由于这两个方法目前未在现有执行路径中被调用，回归风险较低，但契约需要调用方正确遵守。

- 影响：对用户而言，一旦 Keye 的 tower/connector LoRA 完整接入，将可为视觉塔和投影器分别挂载可训练适配器；对系统没有性能、显存或部署影响，改动是纯增量方法。对团队而言，本 PR 为多模态 LoRA 支持补齐了 Keye 这一模型的契约，后续实现 Keye LoRA 的配置与激活映射时可以直接复用这两个入口。
- 风险标记：缺少自动化测试，整除截断依赖，配置字段依赖

关联脉络

- PR #26674 [Feature] Enable tower and connector LoRA for Qwen VL series and idefics3: #31479 中的基础实现，定义了 tower 与 connector 的 LoRA 支持以及这两个辅助函数框架；本 PR 沿用同一模式落地到 Keye。