PR #26149 完整报告

sgl-project/sglang

[VLM] feat: accept grid_thws from preprocessed metadata for kimi

合并时间：2026-05-25 09:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26149

执行摘要

- 一句话：支持 Kimi 图像预计算 grid_thws 元数据
- 推荐动作：该 PR 改动很小，属于对已有预处理管道的适配。建议关注后续是否有统一的多模态元数据方案。

功能与动机

允许上游预处理模块将已计算好的 grid_thws 直接注入元数据，减少重复计算。PR body 提到："Allow Kimi-K2.5 image embedding to read grid_thws from preprocessed multimodal metadata."

实现拆解

仅修改 python/sglang/srt/models/kimi_k25.py 中 get_image_feature 方法的 grid_thws 获取逻辑：

1. 将原来的直接访问 item.image_grid_thw 改为优先从 item.model_specific_data 字典中获取 image_grid_thw。
2. 若 model_specific_data["image_grid_thw"] 为 None，则回退使用 model_specific_data["grid_thws"] 作为备用字段。
3. 将收集到的 tensor 列表拼接后移至设备。
4. 后续流程（vision_tower 调用、mm_projector 投影）不变。

关键文件：

- python/sglang/srt/models/kimi_k25.py（模块 多模态；类别 source；类型 data-contract；符号 get_image_feature）：核心模型文件，修改了 grid_thws 的获取逻辑，支持从预处理元数据读取。

关键符号：get_image_feature

关键源码片段

python/sglang/srt/models/kimi_k25.py

核心模型文件，修改了 grid_thws 的获取逻辑，支持从预处理元数据读取。

```
# python/sglang/srt/models/kimi_k25.py
```

```
# 修改后的 get_image_feature 方法片段
# 优先从 item.model_specific_data 中读取 image_grid_thw，
# 若不存在则回退读取 grid_thws 作为兼容
image_grid_thws = []
for item in items:
    grid_thw = item.model_specific_data.get("image_grid_thw")
    if grid_thw is None:
        grid_thw = item.model_specific_data["grid_thws"] # 备用字段，确保上游已设置
    image_grid_thws.append(grid_thw)
grid_thws = torch.concat(image_grid_thws, dim=0).to(device)
```

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 数据一致性：如果 model_specific_data 中同时存在 image_grid_thw 和 grid_thws 但值不同，当前逻辑会优先用 image_grid_thw，可能掩盖错误。
  2. 向后兼容：原有的 item.image_grid_thw 路径被替代为字典访问，若调用方未填充 model_specific_data，get 返回 None 会走 grid_thws 备用字段；只有当 model_specific_data 不包含任何相关键时才会抛出 KeyError。
  3. 性能：新增 python 循环，对微小批次来说可忽略。 - 影响：仅影响 Kimi-K2.5（kimi_k25.py）模型加载图像特征时的 grid_thws 来源路径。对其他模型无影响。对用户透明，但需要确保上游预处理模块正确设置 model_specific_data。 - 风险标记：数据契约变更，缺少测试覆盖

关联脉络

- PR #26101 [VLM] accept precomputed multimodal metadata: 前序 PR，建立了预处理元数据通道，本 PR 是 Kimi 模型对该通道的适配。