PR #23501 完整报告

sgl-project/sglang

[VLM] Fix Kimi-K2.5 CPU path: rename grid_thws -> image_grid_thw

合并时间：2026-04-28 04:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23501

执行摘要

- 一句话：修复 Kimi-K2.5 CPU 路径 grid 键名不一致
- 推荐动作：可快速合并。对于涉及多路径（CPU/GPU）的键名重构，建议统一检查所有路径避免遗漏。

功能与动机

GPU 路径已在 #22858 中统一为 SGL 标准键名 image_grid_thw，但 CPU 路径 (_cpu_call) 直接返回 HF processor 的原始输出（键名为 grid_thws），导致下游 KimiK25ForConditionalGeneration.get_image_feature 读取 item.image_grid_thw 时抛出 AttributeError。该问题在使用 gpu_image_decode=False（例如为保持训练 / 推理像素一致性）时触发 VLM server warmup 失败。

实现拆解

1. 在 python/sglang/srt/multimodal/processors/kimi_k25.py 的 KimiGPUProcessorWrapper._cpu_call 方法中，将原本直接返回 HF processor 结果的语句改为先保存结果到局部变量 out。
2. 从 out 中 pop 出 grid_thws 键（如果存在），并将其值赋给 image_grid_thw 键。
3. 返回修改后的 out 字典。

关键文件：

- python/sglang/srt/multimodal/processors/kimi_k25.py（模块 多模态处理器；类别 source；类型 core-logic）：对 Kimi-K2.5 多模态处理器的 _cpu_call 方法进行键名映射修复，使其返回 SGL 标准键名 image_grid_thw，与 GPU 路径一致。

关键符号：未识别

关键源码片段

python/sglang/srt/multimodal/processors/kimi_k25.py

对 Kimi-K2.5 多模态处理器的 _cpu_call 方法进行键名映射修复，使其返回 SGL 标准键名 image_grid_thw，与 GPU 路径一致。

```
# python/sglang/srt/multimodal/processors/kimi_k25.py
```

```
# ... 之前代码 ...

def _cpu_call(self, text, images, **kwargs):
    """Fallback: token expansion + medias kwarg -> original HF processor."""
    input_text = text[0] if isinstance(text, list) else text

    if images:
        # Token expansion via media_tokens_calculator
        parts = input_text.split(self._image_token)
        result = [parts[0]]
        for image, part in zip(images, parts[1:]):
            num_tokens = self._hf_processor.media_processor.media_tokens_calculator(
                {"type": "image", "image": image}
            )
            result.append(self._image_token * num_tokens + part)
        input_text = "".join(result)

        # Convert to medias format for Kimi's HF processor
        kwargs["medias"] = [{"type": "image", "image": img} for img in images]

    out = self._hf_processor(text=[input_text], **kwargs)
    # 从 HF processor 原生输出中取出 grid_thws（可能的键名），
    # 重新映射为 SGL 标准键名 image_grid_thw，与 GPU 路径保持一致。
    grid_thws = out.pop("grid_thws", None)
    if grid_thws is not None:
        out["image_grid_thw"] = grid_thws
    return out

# ... 后续代码 ...
```

评论区精华

无实质性 review 讨论，仅有一个 bot 配额警告和两个 CI 触发命令，以及一个 LGTM 评论。两位 reviewer 均直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：极低风险。仅修改一个键名映射，且 GPU 路径已使用相同映射稳定运行；修改仅影响 gpu_image_decode=False 的 CPU 路径。无需测试配套，因为行为不变（只是键名对齐）。
- 影响：影响范围窄：仅修复 Kimi-K2.5 模型在使用 CPU 图像解码时的启动崩溃。对 GPU 路径无影响。
- 风险标记：遗漏同步修改

关联脉络

- PR #22858 [VLM] Kimi K2.5 GPU path grid key rename: 本 PR 是对 #22858 GPU 路径改名的 CPU 路径补全，修复了遗漏的键名不一致。