

PR #29381 完整报告

sgl-project/sglang

[NPU] Fix glm 4.6v

合并时间: 2026-07-01 15:13

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29381>

执行摘要

- 一句话: 修复 NPU GLM-4.6V 处理器参数名变更
- 推荐动作: 建议合并。确认上游 transformers 版本兼容性。

功能与动机

根据 PR body 中的 traceback, 在运行 GLM-4.6V 模型时, 由于 NPU 处理器 `npu_wrapper_glm46v_preprocess` 中函数参数名为 `interpolation`, 而上游 transformers 库已将其更新为 `resample`, 导致 `TypeError: missing 1 required positional argument: 'interpolation'`。修复此回归以确保 NPU 上 GLM-4.6V 预处理正常工作。

实现拆解

1. 修改 `_preprocess` 函数签名: 在 `python/sglang/srt/hardware_backend/npu/modules/glm46v_processor.py` 中, 将 `interpolation: Optional["tvF.InterpolationMode"]` 替换为 `resample: "PILImageResampling | tvF.InterpolationMode | int | None"`。
2. 更新 `self.resize` 调用: 将 `interpolation=interpolation` 改为 `resample=resample`。
3. 同步修改视频预处理 `_preprocess`: 在 `npu_wrapper_glm46v_video_preprocess` 中同样将参数名和调用处的 `interpolation` 更新为 `resample`。
4. 移除未使用导入: 删除了 `from typing import Optional`, 因为不再需要。
5. 修复 lint 问题: 第二次提交修正了第一次提交中可能存在的代码格式问题。

关键文件:

- `python/sglang/srt/hardware_backend/npu/modules/glm46v_processor.py` (模块 NPU 后端; 类别 source; 类型 dependency-wiring): 唯一修改的文件, 包含 NPU 后端的 GLM-4.6V 图像和视频预处理补丁。修复了因上游 transformers 参数名变更导致的崩溃。

关键符号: 未识别

关键源码片段

`python/sglang/srt/hardware_backend/npu/modules/glm46v_processor.py`

唯一修改的文件, 包含 NPU 后端的 GLM-4.6V 图像和视频预处理补丁。修复了因上游 transformers 参数名变更导致的崩溃。

python/sglang/srt/hardware_backend/npu/modules/glm46v_processor.py

修复前：参数名为 interpolation，来自旧版 transformers

修复后：参数名为 resample，对齐上游 transformers 库的 API 变更

```
def npu_wrapper_glm46v_preprocess(func):
```

```
    def _preprocess(
        self,
        images: list["torch.Tensor"],
        do_resize: bool,
        size: SizeDict,
        # 旧版: interpolation: Optional["tvF.InterpolationMode"]
        # 新版: resample 支持多种类型，移除 Optional 导入
        resample: "PILImageResampling | tvF.InterpolationMode | int | None",
        do_rescale: bool,
        rescale_factor: float,
        do_normalize: bool,
        image_mean: float | list[float] | None,
        image_std: float | list[float] | None,
        patch_size: int,
        temporal_patch_size: int,
        merge_size: int,
        disable_grouping: bool | None,
        return_tensors: str | TensorType | None,
        **kwargs,
    ):
        # ... (省略中间处理逻辑)
        if do_resize:
            stacked_images = self.resize(
                stacked_images,
                size=SizeDict(height=resized_height, width=resized_width),
                # 旧版: interpolation=interpolation
                # 新版: resample=resample
                resample=resample,
            )
        # ...
        return BatchFeature(...)
    return _preprocess
```

视频预处理函数也做了相同修改

同时移除 import typing.Optional (无实际影响的清理)

评论区精华

无 reviewer 评论讨论，仅由 [sglang-npu-bot](#) 自动 approve。PR 中 [gemini-code-assist\[bot\]](#) 的评论是配额警告，不影响评审。

- 暂无高价值评论线程

风险与影响

- 风险：变更极小（+4/-6），仅影响 NPU 后端的 GLM-4.6V 处理器，且只是参数名对齐上游，不涉及逻辑改动。风险极低，但需确认上游 resample 参数的行为与原先 interpolation 完全一致，以及是否兼容旧版 transformers。
- 影响：直接影响 Ascend NPU 上运行 GLM-4.6V 模型的用户，修复了预处理阶段的崩溃问题。对其他硬件后端或模型无影响。程度：小范围修复。
- 风险标记：依赖上游 API 变更

关联脉络

- 暂无明显关联 PR