PR #26146 完整报告

sgl-project/sglang

[Ascend NPU] Enable GLM-4.6V series models inference

合并时间：2026-05-28 17:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26146

执行摘要

- 一句话：支持 GLM-4.6V 模型在 NPU 上推理
- 推荐动作：该 PR 实现了对特定模型在 NPU 上的支持，设计与既有 Qwen VL NPU 补丁模式一致，具有较好的参考价值。对于需要在 NPU 上适配其他视觉语言模型的开发者，其补丁机制的架构思路值得学习。但若只是使用 GLM-4.6V 模型，可直接部署功能。

功能与动机

Ascend NPU 仅支持最多 8 维张量，而 GLM-4.6V 的图像预处理过程会创建 10 维张量，导致在 NPU 上运行失败。此外，GLM-4.6V (MoE) 模型设置 use_qk_norm=False，但 NPU decode 路径无条件访问 q_norm/k_norm 属性，需要条件保护。

实现拆解

变更分三步实现：

1. 新增 NPU 预处理补丁模块（python/sglang/srt/hardware_backend/npu/modules/glm46v_processor.py）：参考 Qwen VL 的 NPU 补丁模式，重写图像预处理中的 patch 提取逻辑，利用 transform_patches_to_flatten 将 10 维张量降为 8 维以内。同时包含视频预处理 wrapper，适用于同样需要降维的场景。
2. 在 base_processor.py 中注册补丁调用点（python/sglang/srt/multimodal/processors/base_processor.py）：在 process_mm_data 方法中识别处理器类名为 Glm46VProcessor 的特例，先应用 GLM-4.6V 专用补丁，再设置设备为 npu。同时将 Glm46VProcessor 加入白名单，避免执行默认的 Qwen VL 补丁路径。
3. 修复 q_norm 属性错误：在 NPU decode 路径中添加条件守卫 if self.use_qk_norm else None，确保 q_norm、k_norm 等属性只在启用时访问（此修改在文件 diff 中未直接体现，但 PR body 声称已包含）。
4. 测试与文档：本 PR 未包含测试或文档变更，但代码格式已通过 pre-commit 检查。

关键文件：

- python/sglang/srt/hardware_backend/npu/modules/glm46v_processor.py（模块 NPU 补丁；类别 source；类型 core-logic；符号 npu_wrapper_glm46v_preprocess, _preprocess, npu_wrapper_glm46v_video_preprocess, npu_apply_glm46v_image_preprocess_patch）：核心新增文件，实现了 NPU 兼容的图像和视频预处理逻辑。

- python/sglang/srt/multimodal/processors/base_processor.py（模块 处理器基类；类别 source；类型 dependency-wiring）：注册了 GLM-4.6V 处理器补丁的调用点，新增条件分支以跳过默认 NPU 补丁并应用专用补丁。

关键符号：npu_wrapper_glm46v_preprocess, npu_wrapper_glm46v_video_preprocess, npu_apply_glm46v_image_preprocess_patch

评论区精华

该 PR 由 sglang-npu-bot 机器人直接批准，未产生实质性的 review 讨论。所有评论仅涉及 CI 重跑指令（/rerun-failed-ci），无内容性讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。base_processor.py 的修改限制在特定处理器类名 Glm46VProcessor 分支内，不影响其他模型流程。新增的 processor 文件独立于主流程，仅在补丁注册时按需引用。主要风险在于： 1) 依赖 transformers 内部 API：补丁直接替换 Glm46VImageProcessorFast._preprocess 方法，若 transformers 版本升级导致接口或内部逻辑变化，补丁可能失效。 2) 缺少测试覆盖：未提供单元测试验证补丁在多尺寸、视频等场景的正确性。 3) 视频处理兼容性：视频预处理函数未充分验证，可能存在边界问题。
  - 影响：用户影响：使用 GLM-4.6V 系列模型的用户可在 Ascend NPU 硬件上进行推理，无需手动处理张量维度问题。系统影响：新增 287 行处理器文件，base_processor.py 增加 8 行条件分支，整体改动集中且限制在 NPU 后端内，不涉及核心调度或注意力逻辑。
  - 风险标记：缺少测试覆盖 , 依赖 transformers 内部接口

关联脉络

- 暂无明显关联 PR