PR #17122 完整报告

sgl-project/sglang

[bugfix]GLM-4V model

合并时间：2026-04-01 10:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/17122

执行摘要

此 PR 修复了 GLM-4V 模型在 VisionAttention 中因 num_dummy_heads 参数缺失导致的 dist_utils.divide 异常，确保模型在 NPU 和 GPU 上稳定运行；但 review 中暴露了跨硬件兼容性风险和测试准确性疑虑，需团队跟进验证。

功能与动机

修复 GLM-4V 模型 bug，解决当 num_dummy_heads=0 且与 num_heads 之和不是 tp_size 整数倍时，dist_utils.divide 抛出异常的问题。动机源于模型在 NPU 硬件上的运行崩溃，PR body 中明确描述："In VisionAttention, it is necessary to use num_dumy_heads to calculate. If num_dumy_heads=0 and the result of num_dumy_heads+num_heads is not an integer multiple of self.tp_size, dist_utils.divide will throw an exception"。

实现拆解

实现分为三个模块：

- 核心模型层(python/sglang/srt/models/glm4v.py)：
  - 在 VisionAttention 的 __init__ 方法中添加 num_dummy_heads=vision_config.num_dummy_heads 参数传递。
  - 调整 vision_utils.update_vit_attn_dummy_heads_config(self.config) 调用顺序至 Glm4vVisionModel 初始化前，确保配置正确。
- 处理器层(python/sglang/srt/multimodal/processors/base_processor.py)：
  - 修改 process_mm_data 函数，通过检查 processor.__class__.__name__ 排除 'Glm4vProcessor'，避免在 NPU 上应用不兼容的补丁。
- 测试层(test/registered/ascend/vlm_models/test_ascend_glm_4_5v.py)：
  - 新增 TestGLM4Models 类，为 GLM-4.5V 模型添加 NPU 测试用例，但设置 mmm_accuracy = 0.2。

评论区精华

Review 讨论中最有价值的交锋集中在兼容性和测试准确性：

- 跨硬件兼容性：reviewer xiaobaicxy 提问："Does it affect the normal operation of the model on other hardware platforms? Please add the GPU test results in the PR

description." 这引发了关于修改是否通用化的疑虑，但未在后续讨论中明确解决。
- 测试准确性: reviewer iforgetmyname 直接指出："this accuracy is incorrect https://docs.z.ai/guides/vlm/glm-4.5v"，质疑测试用例的基准值，此问题未被回复或修正。

风险与影响

- 技术风险:
  1. 跨硬件兼容性：修改可能意外影响 GPU 或其他平台的模型行为，因缺乏 GPU 测试结果验证。
  2. 测试覆盖不足：新增测试的准确性被质疑，可能无法有效捕捉回归错误。
  3. 配置时序：调整 vision_utils.update_vit_attn_dummy_heads_config 调用顺序可能引入副作用，影响其他视觉模型初始化流程。
- 影响范围:
  - 对用户：修复了 GLM-4V 模型在 NPU 上的潜在崩溃，提升多模态应用稳定性。
  - 对系统：支持更广泛的硬件部署，但需确保修改不破坏现有 GPU 工作流。
  - 对团队：新增测试为未来回归提供基础，但准确性问题需文档或代码更新以对齐标准。

关联脉络

从近期历史 PR 分析中，未发现直接修改相同文件或针对 GLM-4V 模型的 PR，表明此变更可能是一个独立的 bugfix。然而，标签 npu 和 multimodal 与其他 PR（如 PR 21763 关于 multimodal CI 改进）共享，提示团队在持续优化多模态和 NPU 支持；本 PR 的测试准确性讨论可能关联到更广泛的模型评估标准演进。