

PR #43194 完整报告

vllm-project/vllm

[Bugfix] fix device mismatch in MiniCPM-o-4_5 resampler

合并时间: 2026-05-26 13:12

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43194>

执行摘要

- 一句话: 修复 MiniCPM-o-4_5 resampler 设备不匹配
- 推荐动作: 可直接合入。建议后续跟进 temporal_pos_embed 的设备同步修复, 或确认当前代码路径不会触发。

功能与动机

修复 openbmb/MiniCPM-o-4_5 模型推理时的设备不匹配错误: RuntimeError: Expected all tensors to be on the same device, but found at least two devices, xpu:0 and cpu!。PR body 中附带了完整的错误 stack trace。

实现拆解

1. 定位问题: 在 vllm/model_executor/models/minicpmv.py 的 MiniCPMVResampler.forward 方法中, 第 380-382 行构建 2D 位置编码时, pos_embed 切片后仅调用 .to(dtype) 转换为指定数据类型, 未调用 .to(device) 将张量移动到当前设备 (device 参数)。当模型在 GPU 或 XPU 上运行时, 该张量留在了 CPU 上, 导致后续 x + pos_embed_2d 操作时报设备不匹配错误。
2. 修复方案: 在 pos_embed 切片后链式调用 .to(device=device, dtype=dtype), 同时指定设备和数据类型。改动仅涉及一行, 增加 3 个字符, 将 .to(dtype) 替换为 .to(device=device, dtype=dtype)。
3. 影响范围: 仅修改了单文件单行代码, 不影响任何其他函数或数据流。修复使 resampler 在 TP>1 时也能正确将位置编码放置到工作器所在设备。

关键文件:

- vllm/model_executor/models/minicpmv.py (模块 模型执行; 类别 source; 类型 data-contract): 修复位置, 在 pos_embed 的 .to() 调用中添加 device 参数, 确保 MPS 下位置编码与输入在同一设备。

关键符号: MiniCPMVResampler.forward

关键源码片段

[vllm/model_executor/models/minicpmv.py](#)

修复位置，在 pos_embed 的 .to() 调用中添加 device 参数，确保 MPS 下位置编码与输入在同一设备。

```
# vllm/model_executor/models/minicpmv.py 第 380-384 行
# 修复前: pos_embed 切片后仅指定 dtype, 未指定 device
# 导致在 GPU/XPU 上运行时张量留在 CPU, 引发 RuntimeError
pos_embed_2d.append(
    self.pos_embed[:tgt_h, :tgt_w, :]
    .reshape((tgt_h * tgt_w, -1))
    .to(device=device, dtype=dtype) # 修复: 同时指定 device 和 dtype
) # patches * D
```

评论区精华

gemini-code-assist[bot] 在 review 中指出了两个问题：一是建议合并链式 .to() 调用以简化代码；二是发现 temporal_pos_embed 在第 377 行也存在类似的设备不匹配问题（调用 .to(dtype) 但未指定 device），建议一并修复。该评论未得到作者或合并者的回复，且最终 PR 仅修复了 pos_embed_2d 一处，temporal_pos_embed 的设备问题未被处理。不过 PR 仍获得了 jikunshang 的批准。

- 设备不匹配修复及链式调用建议 (correctness): 作者仅修复了 pos_embed_2d 的 device 问题，未采纳链式调用简化，也未修复 temporal_pos_embed。PR 仍获批准。

风险与影响

- 风险：
 1. 回归风险：极低。改动为单行增加参数，不影响其他分支逻辑。
 2. 残留风险：review 指出的 temporal_pos_embed 设备不匹配问题未修复，在视频输入场景中可能仍然触发类似错误，但可能是作者判断该路径在当前版本中不会执行到。
 3. 兼容性：无影响。
- 影响：
 1. 用户影响：修复了 openbmb/MiniCPM-o-4_5 在 multi-GPU 或 XPU 上推理时的崩溃问题，使该模型可在 TP>1 的环境下正常运行。
 2. 系统影响：无。
 3. 团队影响：无。 - 风险标记：残留未修复设备不匹配，缺少测试覆盖

关联脉络

- 暂无明显关联 PR