

# PR #31663 完整报告

sgl-project/sglang

[Bugfix] Place empty Qwen encoder-DP embeddings on the communication device

合并时间: 2026-07-21 19:54

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31663>

## 执行摘要

- 一句话: 修复 Qwen encoder DP 空 rank 时 embedding 设备不匹配
- 推荐动作: 建议精读该 PR: 它是一个经典的多设备通信 bug fix, 展示了分布式推理中张量设备一致性的重要性。同时体现了一行 device 指定错误的连锁反应。

## 功能与动机

Fix Qwen multimodal inference when encoder data parallelism is enabled and an attention-TP rank is assigned no images. In this case, the empty Qwen rope\_3d embedding can remain on CPU and then be passed to the device communication group, causing HCCL/NCCL all\_gather to fail with: RuntimeError: No backend type associated with device type cpu

## 实现拆解

仅修改一个文件 `python/sglang/srt/multimodal/mm_utils.py` 中的 `run_dp_sharded_mrope_vision_model` 函数:

1. `packed_2d_rope` 空分支: 第 707-712 行, 当 `pixel_values_local` 为空时, 将 `torch.empty` 的 `device` 参数从 `input_device` 改为 `vision_model.device`, 确保空张量创建在 vision model 所在 GPU 上。
2. 非 `packed_2d_rope` 空分支: 第 726-730 行, 同样将 `device` 从 `input_device` 改为 `vision_model.device`。

改动共 2 行, +2/-2, 不涉及新增测试或配置。其他路径 (非空 rank、packed 2D RoPE 正常模型等) 保持不变。

关键文件:

- `python/sglang/srt/multimodal/mm_utils.py` (模块 多模态; 类别 source; 类型 core-logic; 符号 `run_dp_sharded_mrope_vision_model`): 核心修复文件, 修改了 `run_dp_sharded_mrope_vision_model` 函数中空 embedding 张量的设备指定。

关键符号: `run_dp_sharded_mrope_vision_model`

## 关键源码片段

`python/sglang/srt/multimodal/mm_utils.py`

核心修复文件，修改了 `run_dp_sharded_mrope_vision_model` 函数中空 embedding 张量的设备指定。

```
# python/sglang/srt/multimodal/mm_utils.py
# 在 run_dp_sharded_mrope_vision_model 函数中，处理 encoder-DP 下无图片的 rank

if packed_2d_rope:
    if pixel_values_local is not None and pixel_values_local.shape[0] > 0:
        # 正常分支：有图片时在 pixel_values 设备上运行 vision model
        local_grid_thw = torch.tensor(
            local_grid_thw_list, device=pixel_values_local.device
        )
        # ... 调用 vision_model ...
    else:
        # 空 rank 分支：创建空张量，device 从 input_device 改为 vision_model.device
        out_dim = getattr(vision_model.config, "hidden_size", None)
        image_embeds_local = torch.empty(
            (0, embed_dim_reduction_factor, out_dim),
            device=vision_model.device, # 之前是 input_device，可能为 CPU
            dtype=input_dtype,
        )
else:
    # 非 packed_2d_rope 路径（如 rope_3d）
    if pixel_values_local is not None and pixel_values_local.shape[0] > 0:
        # ... 正常处理 ...
    else:
        out_dim = getattr(vision_model, "out_hidden_size", None)
        if out_dim is None:
            out_dim = vision_model.config.hidden_size
        image_embeds_local = torch.empty(
            (0, out_dim),
            device=vision_model.device, # 同样从 input_device 改为 vision_model.device
            dtype=input_dtype,
        )
```

## 评论区精华

review 中 `gemini-code-assist[bot]` 指出相同的 bug 也存在于 `packed_2d_rope` 路径（原代码第 686 行），并建议一并修复。提交历史显示该建议已被采纳：作者在后续提交中将 `packed_2d_rope` 空分支的 `device` 也改为了 `vision_model.device`。sglang-npu-bot 最终批准。

- `packed_2d_rope` 路径存在相同 bug (correctness): 作者接受建议，在后续提交中修复了 `packed_2d_rope` 路径。

## 风险与影响

- 风险：变更范围极小（仅 2 行），风险低。但需注意：`vision_model.device` 必须与 attention TP 组设备一致，若 vision model 因某些配置被放置到不同设备（目前不支持），可能引入新的不匹配。另外未添加自动化测试，回归依赖手动验证和 CI 覆盖。

- 影响：影响范围窄：仅影响 Qwen 系列模型在启用 encoder DP 且图片数量少于 encoder DP rank 数时的场景。对于其他模型、非 encoder DP 配置或无空 rank 的情况，无行为变化。
- 风险标记：缺少测试覆盖

## 关联脉络

- 暂无明显关联 PR