PR #21191 完整报告

sgl-project/sglang

fix: Use base GPU ID CUDA device for multimodal processor

合并时间：2026-05-21 13:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21191

执行摘要

- 一句话：修复多模态处理器默认使用 GPU 0 的问题
- 推荐动作：最小侵入修复，逻辑清晰，推荐合并。值得关注的是多 GPU 资源隔离的设计思路：预处理应在当前进程绑定的 GPU 上进行，而非全局默认设备。

功能与动机

PR body 说明在多实例部署场景下，BaseImageProcessorFast 始终使用默认 CUDA 设备（cuda:0），导致 GPU 0 显存分配过多而其他 GPU 空闲。issue#21191 中的截图显示了修改前后显存占用分布的变化。

实现拆解

在 python/sglang/srt/multimodal/processors/base_processor.py 的 process_mm_data 方法中，将原来硬编码的 "cuda" 替换为 f"cuda:{base_gpu_id}"，其中 base_gpu_id 通过 get_global_server_args().base_gpu_id 获取。该修改仅影响非 CPU、非 XPU、非 NPU 且具备 BaseImageProcessor 的处理器路径。

关键文件：

- python/sglang/srt/multimodal/processors/base_processor.py（模块 多模态处理器；类别 source；类型 core-logic；符号 process_mm_data）：核心修改文件：将处理器设备从硬编码 'cuda' 改为使用 base_gpu_id，平衡多实例 GPU 显存分配。

关键符号：process_mm_data

关键源码片段

python/sglang/srt/multimodal/processors/base_processor.py

核心修改文件：将处理器设备从硬编码 'cuda' 改为使用 base_gpu_id，平衡多实例 GPU 显存分配。

```
# 文件：python/sglang/srt/multimodal/processors/base_processor.py
# 方法：process_mm_data（片段）

# 此前设备选择逻辑：
# elif not _is_npu:
# kwargs["device"] = "cuda" # 始终使用 GPU 0
```

```
#
# 修复后：
if hasattr(processor, "image_processor") and isinstance(
    processor.image_processor, BaseImageProcessor
) and not self.server_args.disable_fast_image_processor:
    if _is_cpu or get_global_server_args().rl_on_policy_target is not None:
        kwargs["device"] = "cpu"
    elif _is_xpu:
        kwargs["device"] = "xpu"
    elif not _is_npu:
        # 通过 get_global_server_args().base_gpu_id 获取当前进程绑定的 GPU ID
        # 确保多实例部署下各自处理器使用正确的 GPU，避免 GPU 0 显存瓶颈
        base_gpu_id = get_global_server_args().base_gpu_id
        kwargs["device"] = f"cuda:{base_gpu_id}"
    elif processor.__class__.__name__ not in {"Glm4vProcessor",}:
        # NPU 处理器特殊处理
        ...
```

评论区精华

Review 中 reviewer mickqian 建议更新文档以反映此新参数行为，但 author moehanabi 指出并无新增参数（base_gpu_id 已是已有参数），最终没有进一步文档改动。

- 文档更新建议与回应 (documentation): 无需额外文档更新，因为 base_gpu_id 已是已有参数且已有文档。

风险与影响

- 风险：风险极低：仅改动了一行硬编码字符串，且该设备选择分支已存在于现有逻辑中，仅在非 CPU/XPU/NPU 环境下生效。可能的风险是某些依赖固定 cuda:0 行为的边缘场景，但 base_gpu_id 本身就是用户显式配置的，此变更与其期望一致。
- 影响：影响范围为多 GPU 多实例部署 VLM 的场景，尤其是使用 --base-gpu-id 参数的 SGLang 实例。修正后每个实例的处理器会正确使用其绑定的 GPU，避免 GPU 0 显存压力过大。单 GPU 或单实例用户无感知。
- 风险标记：无测试覆盖

关联脉络

- 暂无明显关联 PR