ChatGPT

错误原因分析与背景

从报错信息来看，错误发生在 vLLM 底层的多模态输入处理阶段，具体是 **MRotaryEmbedding** 模块中试图访问图像网格（`image_grid_thw`）时出现索引越界[1]。也就是说，当模型处理输入图像时，生成的位置编码超出了预期范围。类似的堆栈信息已在 vLLM 的 GitHub issue 中被多次提及。例如，一位用户在使用 Qwen2.5-VL 模型时也遇到了相同的 IndexError[1]。该错误并非源于 MinerU 模型自身，而是 vLLM 对图像输入处理的已知问题。

根据模型报告，**MinerU2.5-2509-1.2B** 基于 Qwen2.5-VL 架构（一个视觉-语言模型），其多模态处理依赖 vLLM 提供的多媒体输入支持。vLLM 0.10.x 版本在处理此类 Qwen2.5-VL 图像输入时存在缺陷，容易触发这一错误[1]。简单而言，这不是普通 Python 列表越界问题，而是 vLLM 处理图像 token 时的内部错误。

已知相关问题与讨论

- **vLLM GitHub Issue**：vLLM 官方仓库已有关于 Qwen2.5-VL 附加图像时报错的 issue（如 #23538、#23797），堆栈信息与您的情况几乎一致[1]。这些 issue 均指向 `MRotaryEmbedding._vl_get_input_positions_tensor` 方法在计算图像位置编码时出错。
- **vLLM 社区讨论**：在 vLLM 社区论坛（discuss.vllm.ai）中，也有用户反馈“Qwen2.5-VL 模型接入图像时出错，返回 ‘IndexError: list index out of range’ ”的问题。论坛指出，此为 vLLM 0.10.x 版本处理某些图像输入时的已知 bug[2]。官方答复提到，这并非请求格式问题，而是内核尚未修复的漏洞。
- **硬件兼容性**：您的 GPU 是 Tesla T4（CUDA CC 7.5），不支持 vLLM 的 V1 引擎（只有 CC ≥8.0 支持）。因此无法借助 V1 引擎绕过此问题[2]。

至今，MinerU 官方仓库中并未针对该错误给出解决方案，因为问题出在 vLLM 层面，而非模型本身。我们需要根据社区信息和现有讨论采取应对措施。

解决方案建议

1. **降级 vLLM 版本**：由于该问题在 vLLM 0.10.1 及以上版本中存在，**可尝试使用 vLLM 0.10.0** 版本运行 MinerU2.5。社区反馈表明，在 vLLM 0.10.0 下 Qwen2.5-VL 的图像输入问题消失[2]。您可以更换镜像标签（如使用 `vllm/vllm-openai:v0.10.0`，或在容器内 `pip install vllm==0.10.0`）来实现降级。务必清除缓存并确保只加载 v0.10.0 相关代码，以避免兼容性问题。
2. **尝试不同图片或尺寸**：一些用户发现错误可能与输入图片的分辨率或格式有关[2]。尽管不稳定，但可尝试使用不同来源或更大尺寸的图像（例如将 330×131 的截图放大或裁剪其他图像）来测试。如果问题与特定图像相关，这可能会暂时规避错误。但这种方法非根本之计，仅供实验验证。
3. **尝试 CPU 推理**：如果硬件允许，可考虑在 CPU 上运行 vLLM（设置 `device="cpu"` 或类似参数），使用纯 CPU 推理模式。CPU 模式下可使用 V1 引擎（因为无需显卡支持FlashAttention），理论上也可能绕过此bug。但要注意，CPU 运算会明显更慢，只适合排查或少量推理。
4. **关注 vLLM 更新**：该问题已被 vLLM 团队收录为 Bug。建议关注 vLLM 后续版本（如 v0.10.2 及以上）发布说明，看是否解决了此图像处理问题。据发布日志，v0.11.0 已完全移除 V0 引擎，仅支持 V1，新版本对旧 GPU 已不兼容，故不建议升级到 0.11.x。在满足硬件要求时，再考虑使用新版并移除 `VLLM_USE_V1=0` 限制。
5. **调整配置**：目前已设置 `VLLM_USE_V1=0` 强制使用 V0 引擎（因为 T4 不支持 V1）。如果有机会使用 CC≥8.0 的 GPU，可去掉该环境变量尝试 V1。但在当前硬件上，这通常不可行，可能导致其他错误[2][3]。

总之，结合现有讨论，**最可靠的方案是降级到 vLLM 0.10.0** [2]。该版本在处理多模态输入时未触发此错误。同时，可尝试更改输入图像（尺寸、格式）作辅助测试。如问题仍未解决，则需要等待 vLLM 官方的正式修补或升级硬件环境。

参考资料： vLLM GitHub issues [1]、vLLM 社区论坛 [2]。这些资源提供了错误定位和临时解决方案的信息。根据其中建议调整 vLLM 版本和配置通常能消除此错误。

---

[1] [Usage]: InputProcessingError: Failed to prepare inputs for sequence group with request id: 0, · Issue #23797 · vllm-project/vllm · GitHub
https://github.com/vllm-project/vllm/issues/23797

[2] IndexError: list index out of range (Qwen/Qwen2.5-VL-3B-Instruct) - General - vLLM Forums
https://discuss.vllm.ai/t/indexerror-list-index-out-of-range-qwen-qwen2-5-vl-3b-instruct/1626

[3] [Bug]: [V1] Testla T4 cannot work for V1 · Issue #15853 · vllm-project/vllm · GitHub
https://github.com/vllm-project/vllm/issues/15853