

PR #42332 完整报告

vllm-project/vllm

[Bugfix] Fixes MiniCPM-O resampler device placement to avoid tensor device mismatch

合并时间: 2026-06-17 16:35

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/42332>

执行摘要

- 一句话: 修复 MiniCPM-O resampler 设备放置错误
- 推荐动作: 值得精读的 small fix 示例。展示了继承层次中父类 - 子类责任链断裂的典型排错思路: 子类覆盖父类方法并重写了权重加载逻辑, 但遗漏了父类中的设备后处理步骤。对于使用多模态模型的开发者, 建议关注模型初始化时的设备一致性问题。

功能与动机

修复 Issue #42322: MiniCPM-O 2.6 启动时崩溃, 报错 'Expected all tensors to be on the same device'。根因是 Resampler2_5._set_2d_pos_cache 在 __init__ 时将 pos_embed buffer 创建在 CPU 上, 而 MiniCPMOBaseModel.load_weights 覆盖了父类方法却未调用 _ensure_resampler_device, 导致 buffer 未被移至 GPU。

实现拆解

1. 定位根因: 在 vllm/model_executor/models/minicpmo.py 的 MiniCPMOBaseModel.load_weights 方法中, 父类 MiniCPMVBaseModel 在 load_weights 末尾会调用 _ensure_resampler_device() 以确保 resampler 的 buffer 处于正确设备, 但子类 MiniCPMOBaseModel 覆盖该方法时遗漏了此调用 (仅调用 AutoWeightsLoader 加载权重后直接返回)。
2. 修复方式: 在 MiniCPMOBaseModel.load_weights 中获取 loader.load_weights(weights) 的返回值后, 立即调用 self._ensure_resampler_device(), 再返回。
3. 涉及模型: MiniCPMO2_6 和 MiniCPMO4_5 均继承自 MiniCPMOBaseModel, 所以一行修改同时修复两个模型。
4. 无测试 / 配置 / 部署修改: 修复仅涉及一行源码变更, 未增加测试或配置, 影响范围极窄。

关键文件:

- vllm/model_executor/models/minicpmo.py (模块 模型执行器; 类别 source; 类型 core-logic; 符号 MiniCPMOBaseModel.load_weights): 修复的核心文件, 在 MiniCPMOBaseModel.load_weights 中增加 _ensure_resampler_device() 调用, 解决 resampler buffer 设备不匹配问题。

关键符号: MiniCPMOBaseModel.load_weights

关键源码片段

vllm/model_executor/models/minicpmo.py

修复的核心文件，在 `MiniCPMOBaseModel.load_weights` 中增加 `_ensure_resampler_device()` 调用，解决 resampler buffer 设备不匹配问题。

```
# vllm/model_executor/models/minicpmo.py
# MiniCPMOBaseModel 重写 load_weights 以跳过 "tts" 前缀的权重
# 但忽略了父类 MiniCPMVBaseModel 中的 _ensure_resampler_device() 调用
# 导致 resampler 的 pos_embed buffer 留在 CPU 上
# 修改后，在加载权重后显式调用该方法，使 buffer 移至 GPU

def load_weights(self, weights: Iterable[tuple[str, torch.Tensor]]) -> set[str]:
    loader = AutoWeightsLoader(self, skip_prefixes=["tts"])
    loaded = loader.load_weights(weights)
    self._ensure_resampler_device() # 新增：确保 resampler buffer 在正确设备上
    return loaded
```

评论区精华

review 中无实质性讨论。tc-mb 在 issue 评论中确认根因由 PR#39862 引入，该 PR 将 resampler 设备放置集中到 `_ensure_resampler_device()`，在 `load_weights` 末尾调用，但只修改了 `minicpmv.py` 中的子类，遗漏了 `minicpmo.py` 中的 `MiniCPMOBaseModel`。修复方式简单一致，获得审批。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅新增一行调用 `_ensure_resampler_device()`，该方法已在父类中定义并稳定使用，行为可预期。可能的风险：若 `_ensure_resampler_device()` 实现本身有 bug（如错误地移动了不该移动的张量），可能引发新问题，但该函数已用于 `minicpmv.py` 且无相关 bug 报告。无回归测试覆盖，但该修复已由 PR 作者和 reviewer 在 MiniCPM-o-2.6 上本地验证。
- 影响：影响范围：仅修复 MiniCPM-O 2.6 和 4.5 两个模型在启动时的崩溃问题。对系统其他模块无影响。用户可正常启动并使用这两个模型。影响程度：高（对于使用 MiniCPM-O 的用户）。
- 风险标记：缺少测试覆盖

关联脉络

- PR #39862 [Bugfix] Fix online FP8 for MiniCPM models: 该 PR 引入了 `_ensure_resampler_device()` 集中化 resampler 设备放置，但只修改了 `minicpmv.py`，遗漏了 `minicpmo.py`，导致当前 bug。