PR #22293 完整报告

sgl-project/sglang

[fix] [whisper] ensure inputs are moved to the correct device before processing.

合并时间：2026-04-08 23:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22293

执行摘要

本次 PR 修复了 Whisper 模型因未覆盖 #22038 引入的延迟设备转移机制而导致的设备不匹配错误。通过在 forward 方法中显式将输入张量移动到模型权重所在设备，解决了 CUDA 图捕获时出现的 "Expected all tensors to be on the same device" 运行时异常。这是一个针对特定模型的关键 bugfix，影响范围有限但解决了稳定性问题。

功能与动机

问题根源：Whisper 模型未覆盖 #22038 引入的延迟设备转移机制，导致输入特征 input_features 和位置 IDposition_ids 可能停留在 CPU，而模型权重已在 GPU 上。

具体表现：执行手动测试 test/manual/test_whisper_cuda_graph.py 时出现运行时错误：

```
RuntimeError: Expected all tensors to be on the same device, but got weight is on cuda:0, different from other tensors on cpu
```

修复目标：确保在卷积操作 self.conv1(input_features) 执行前，所有张量都在同一设备上。

实现拆解

仅修改 python/sglang/srt/models/whisper.py 文件的 forward 方法，在原有逻辑前插入设备同步代码：

```
device = self.conv1.weight.device
input_features = input_features.to(device=device)
position_ids = position_ids.to(device=device)
```

关键设计点：

1. 使用 self.conv1.weight.device 作为目标设备参考点
2. 同时移动 input_features 和 position_ids 两个张量
3. 在第一个卷积操作前完成设备转移

评论区精华

review 中只有一条实质性技术讨论：

> mickqian: "nit: device=next(self()).device is more robust"

讨论要点：

- mickqian 建议使用迭代器方式获取设备，认为比直接引用特定层权重更稳健
- 作者最终未采纳此建议，保持原方案
- 可能原因：conv1 作为模型第一个卷积层，其设备位置具有代表性，且代码更直观

风险与影响

技术风险：

1. 性能开销：增加两次 to() 调用，但在 CUDA 图场景下这是必要成本
2. 覆盖不全：如果模型其他部分也有类似设备依赖，可能遗漏
3. 测试缺失：缺少针对此修复的单元测试

影响评估：

- 用户影响：仅影响使用 Whisper 模型且启用 CUDA 图的用户，解决崩溃问题
- 系统影响：无架构或接口变更，不影响其他模型
- 团队影响：提醒需要检查其他模型是否也存在 #22038 覆盖不全问题

关联脉络

与历史 PR 的关联：

1. #22038：引入了延迟设备转移机制，是本 bug 的根本原因
2. #21817：类似的多模态设备同步修复，解决 warmup 图像初始化并发问题
3. #22266：NPU 上 Qwen3.5 视频处理器的设备相关修复

演进趋势：

- 随着 #22038 机制的推广，需要确保所有模型都正确覆盖设备转移逻辑
- 多模态模型（Whisper、扩散模型等）在设备同步方面有特殊需求
- CI 标签 run-ci 和 multimodal 频繁出现，显示团队对多模态测试的重视