# PR #44205 完整报告

vllm-project/vllm

[Bugfix] fix EVS for qwen3-vl

合并时间：2026-06-04 19:06

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44205

## 执行摘要

- 一句话：修复 Qwen3-VL EVS 设备不匹配错误
- 推荐动作：这是一个针对特定模型特定功能的紧急修复，改动经过验证且明确。建议快速合并。

## 功能与动机

Issue #44204 报告 Qwen3-VL 模型启用 `--video_pruning_rate` 后服务崩溃，错误为 `RuntimeError: Expected all tensors to be on the same device, but got index is on cpu`。用户分析指出 PR #34246 引入的优化移除了对 `repl_token_ids` 设备的显式控制，导致 `F.embedding` 的 input（CPU）与 weight（GPU）不匹配。

## 实现拆解

1. 在 _create_final_video_embeddings 方法开头添加 device = video_embeddings.device，获取视频特征张量所在设备（通常为 GPU）。
2. 将 repl_token_ids = torch.tensor(video_repl.full) 改为 torch.tensor(video_repl.full, device=device)，使 token ID 张量直接创建在 GPU 上。
3. 将 _cached_tensor(self.config.video_token_id, repl_token_ids.device) 简化为 _cached_tensor(self.config.video_token_id, device=device)，避免依赖 repl_token_ids 的设备属性，同时移除临时变量。
4. 其他逻辑完全不变，仅此三处调整。测试方面，作者在 0.20.2 版本上验证修复有效。

关键文件：

- vllm/model_executor/models/qwen3_vl.py（模块 视觉模型；类别 source；类型 data-contract；符号 _create_final_video_embeddings）：唯一改动的文件，修复 Qwen3-VL 的 _create_final_video_embeddings 方法设备不匹配错误。

关键符号：_create_final_video_embeddings

## 关键源码片段

### vllm/model_executor/models/qwen3_vl.py

唯一改动的文件，修复 Qwen3-VL 的 _create_final_video_embeddings 方法设备不匹配错误。

```
    def _create_final_video_embeddings(
```

```
        self,
        video_embeddings: torch.Tensor,
        num_tokens_per_frame: list[int],
        timestamps: list[float],
        video_grid_thw: list[int],
        retention_mask: torch.Tensor,
    ) -> torch.Tensor:
        # 获取 video_embeddings 所在设备（通常为 GPU），确保后续张量与之一致
        device = video_embeddings.device

        # ... 中间生成 video_repl 的代码 ...

        # 修复：创建 repl_token_ids 时指定 device，避免默认分配到 CPU
        repl_token_ids = torch.tensor(video_repl.full, device=device)
        # 同理，embed_token_id 也显式指定 device，不再依赖 repl_token_ids.device
        embed_token_id = _cached_tensor(self.config.video_token_id, device=device)
        is_video_embed = torch.isin(repl_token_ids, embed_token_id)

        # 获取 indicator token 的文本嵌入，此时 repl_token_ids 在 GPU 上，不会报设备不匹配
        text_embeddings = self.get_language_model().embed_input_ids(repl_token_ids)
        ...
```

# 评论区精华

预提交检查要求修复代码格式，作者两次提交修复。最终 DarkLight1337 批准并合并。无实质 review 争议。

- Pre-commit 格式检查失败 (style): 作者随后两次提交修复了代码格式问题。

# 风险与影响

- 风险：风险极小。改动的本质是回退到 PR #34246 之前的行为，且仅影响 EVS（视频修剪）路径。若忘记传递 device，可能导致新的设备不匹配，但此处已修正。对非 EVS 场景无影响，因为 _create_final_video_embeddings 仅在 EVS 启用时被调用。
- 影响：直接影响：使用 Qwen3-VL 模型并且设置 --video_pruning_rate 的用户将从崩溃中恢复。间接影响：PR #34246 本来是为了优化同步性能，但该修复引入 bug；本 PR 回退可能让部分用户回到之前的同步开销。但 PyTorch 2.9+ 已解决掩码同步问题，所以性能影响可忽略。影响范围：仅 Qwen3-VL 模型 + EVS 启用场景。影响程度：中，因为该场景是关键功能。
- 风险标记：单一文件变更，极低风险

# 关联脉络

- PR #34246 [Core] Simplify multimodal masking: 本 PR 回退了 #34246 对 qwen3_vl.py 的更改，该更改引入了设备不匹配 bug。