PR #45383 完整报告

vllm-project/vllm

[BugFix] Fix prompt_embeds for multimodal models

合并时间：2026-06-14 16:44

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45383

执行摘要

- 一句话：修复多模态 prompt_embeds 被覆盖及异步调度兼容问题
- 推荐动作：值得精读，尤其是 _preprocess 中巧妙使用 torch.where 避免覆盖用户 embedding 的设计，以及配置层兼容性阻断模式。可与后续 PR #45673 对照阅读，了解如何逐步启用异步调度支持。

功能与动机

关联 Issue #44842 报告了多模态模型设置 prompt_embeds 时崩溃。PR body 指明两种根因：

1) 多模态路径无条件重新 embed（覆盖用户输入）； 2) 异步调度下 is_token_ids 标记问题，导致生成退化。

实现拆解

1. 模型运行器多模态分支保护：在 vllm/v1/worker/gpu_model_runner.py 的 _preprocess 中，当多模态条件成立且 enable_prompt_embeds 且当前 batch 有 prompt_embeds 请求时，对输入 id 做 torch.where 掩码：仅对 is_token_ids=True 的位置执行 embedding gather，并将 prompt_embeds 位置的 placeholder id 置零以防止越界；随后用 torch.where 将新生成的 embedding 写回 inputs_embeds，保留 prompt_embeds 位置原有值。
2. 配置层兼容性阻断：在 vllm/config/vllm.py 的 __post_init__ 中，如果用户显式启用了异步调度且同时满足 enable_prompt_embeds 和 is_multimodal_model，则抛出 ValueError；如果异步调度为默认（None）且该组合满足，则自动禁用并 warning。
3. 测试配套：没有新增自动化测试，但 PR 提供了可复现脚本；现有 test_models 文本 -only prompt_embeds 测试不变。

关键文件：

- vllm/v1/worker/gpu_model_runner.py（模块 模型运行器；类别 source；类型 core-logic；符号 _preprocess）：核心修复：多模态分支中保护 prompt_embeds 不被覆盖，使用掩码只对 token 位置做 embedding。
- vllm/config/vllm.py（模块 配置层；类别 source；类型 core-logic；符号 SchedulerConfig.post_init）：配置层兼容性阻断：异步调度与多模态 prompt_embeds 组合自动禁用或硬错误。

关键符号：GpuModelRunner._preprocess, SchedulerConfig.post_init

关键源码片段

vllm/v1/worker/gpu_model_runner.py

核心修复：多模态分支中保护 prompt_embeds 不被覆盖，使用掩码只对 token 位置做 embedding。

```
# vllm/v1/worker/gpu_model_runner.py（关键改动位置）

if self.enable_prompt_embeds and self.input_batch.req_prompt_embeds:
    # 部分位置携带预计算 prompt_embeds：它们已在 self.inputs_embeds 中，
    # 且标记 is_token_ids=False。此处仅对 token-id 位置做 embedding。
    # 先将 placeholder id 置零（防止越界），再通过 torch.where 写回。
    is_token_ids = self.is_token_ids.gpu[:num_scheduled_tokens]
    safe_input_ids = torch.where(
        is_token_ids,
        self.input_ids.gpu[:num_scheduled_tokens],
        0,
    )
    inputs_embeds_scheduled = self.model.embed_input_ids(
        safe_input_ids,
        multimodal_embeddings=mm_embeds,
        is_multimodal=is_mm_embed,
    )
    target = self.inputs_embeds.gpu[:num_scheduled_tokens]
    self.inputs_embeds.gpu[:num_scheduled_tokens] = torch.where(
        is_token_ids.unsqueeze(-1),
        inputs_embeds_scheduled,
        target,
    )
else:
    # 原始逻辑：无 prompt_embeds 时直接覆盖
    inputs_embeds_scheduled = self.model.embed_input_ids(
        self.input_ids.gpu[:num_scheduled_tokens],
        multimodal_embeddings=mm_embeds,
        is_multimodal=is_mm_embed,
    )
    self.inputs_embeds.gpu[:num_scheduled_tokens].copy_(
        inputs_embeds_scheduled
    )
```

评论区精华

Reviewer @qthequartermasterman 提问：异步调度 + prompt_embeds + 多模态需要哪些额外支持才能工作？作者 @mrn3088 回答：需要另一套 fix 来处理异步调度下的标记问题，已有一个未充分测试的实现，后续以 PR #45673 跟进。

- 异步调度与多模态 prompt_embeds 的兼容性 (question): 当前行为：禁用或报错；后续 PR #45673 将提供完整支持。

风险与影响

- 风险：
  1. 核心路径变更：修改了 _preprocess，可能影响其他多模态模型的 embedding 生成路径，但变更仅当 enable_prompt_embeds=True 且 batch 包含 prompt_embeds 请求时触发，风险可控。
  2. 功能默认降级：异步调度被自动禁用，可能降低吞吐量，但这是临时方案；用户若显式启用会硬错误，避免静默退化。
  3. 缺少测试覆盖：没有针对多模态 prompt_embeds 的自动化回归测试，依赖手动验证。
     - 影响：直接使用 prompt_embeds 的多模态用户从 EngineCore 崩溃或错误输出变为正确结果；文本 -only prompt_embeds 用户不受影响；使用异步调度 + 多模态 prompt_embeds 的用户会收到警告或错误提示，之后可升级到后续 PR #45673 获得完整支持。 - 风险标记：核心路径变更，功能默认降级（异步调度禁用），缺少自动化测试覆盖

关联脉络

- PR #44842 [Bug]: Setting prompt_embeds does not work for vision-language models: 关联 issue，报告了该 bug
- PR #45252 [Bugfix] Fix M-RoPE assertion for multimodal prompt_embeds: PR body 指出本 PR 是互补修复（不同根因）
- PR #45673 [Bugfix] Fix prompt_embeds for multimodal models with async scheduling: 作者提及的后续 PR，用于在异步调度下启用支持