PR #45868 完整报告

vllm-project/vllm

[ModelRunnerV2] Various model/config compatibility fixes

合并时间：2026-06-17 11:23

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45868

执行摘要

- 一句话：修复 MRV2 与多种模型和配置的兼容性问题
- 推荐动作：建议相关开发者重点阅读 model_runner.py 中 input_ids 置 None 的条件逻辑以及 encoder_runner.py 中 is_embed 累加的修正，这是典型的边缘情况处理。整体 PR 可作为 MRV2 兼容性修复的参考模式。

功能与动机

这些修复来源于 PR#44443 的 CI 故障分析，旨在解决 MRV2 在 CohereASR、Whisper 等模型以及 external_launcher + PP 配置下无法正常运行的问题。

实现拆解

1. input_ids 在输入嵌入时置 None：在 vllm/v1/worker/gpu/model_runner.py 的 execute_model 方法中，当 inputs_embeds 非 None 且模型声明 requires_raw_input_tokens 为 False 时，将 input_ids 设为 None。
2. 补充 intermediate_tensors 默认值：在同一方法的 model_inputs 字典中始终设置 "intermediate_tensors": None，因为部分模型的前向函数期望该参数但未提供默认值。
3. 修复多模态 is_embed 累加：在 vllm/v1/worker/gpu/mm/encoder_runner.py 的 gather_mm_embeddings 中将一行 = 改为 |=，确保当同一条请求有多个多模态特征区域时，is_mm_embed 标记正确累积而不被覆盖。
4. 修复 Whisper 全 CUDA 图的请求数：在 vllm/v1/worker/gpu/model_states/whisper.py 的 prepare_attn 中将 _get_encoder_seq_lens 调用改为传递 num_reqs（已填充后的值），而非内部从 req_ids 推导。
5. 回退到 MRV1 对于 external_launcher + PP：在 vllm/config/vllm.py 的 _get_v2_model_runner_unsupported_features 中新增条件，当 distributed_executor_backend 为 external_launcher 且 pipeline_parallel_size > 1 时，将该组合标记为 MRV2 不支持，并自动回退到 MRV1。

关键文件：

- vllm/v1/worker/gpu/model_runner.py（模块 模型运行器；类别 source；类型 data-contract；符号 execute_model）：核心变更：在 input_embeds 非 None 时传递 input_ids=None，并始终提供 intermediate_tensors=None 默认值，影响 MRV2 所有请求的输入组装。

- vllm/config/vllm.py（模块 配置检查；类别 source；类型 core-logic；符号 _get_v2_model_runner_unsupported_features）：添加 external_launcher + PP 为 MRV2 不支持特性，自动回退到 MRV1，确保该配置下 V2 引擎不会崩溃。
- vllm/v1/worker/gpu/model_states/whisper.py（模块 Whisper 状态；类别 source；类型 data-contract；符号 prepare_attn, _get_encoder_seq_lens）：修复 Whisper 全 CUDA 图模式下使用填充后请求数，使 Whisper 模型在 MRV2 下能正常使用 full cudagraph。
- vllm/v1/worker/gpu/mm/encoder_runner.py（模块 MM 编码器；类别 source；类型 core-logic；符号 gather_mm_embeddings）：一行修正：将 is_mm_embed 的赋值改为或赋值，避免多区域重复覆盖，修复多模态嵌入标志累积错误。

关键符号：execute_model, _get_v2_model_runner_unsupported_features, prepare_attn, _get_encoder_seq_lens, gather_mm_embeddings

关键源码片段

vllm/v1/worker/gpu/model_runner.py

核心变更：在 input_embeds 非 None 时传递 input_ids=None，并始终提供 intermediate_tensors=None 默认值，影响 MRV2 所有请求的输入组装。

```
# 在 execute_model 方法中，首先获取 input_ids
input_ids = input_batch.input_ids
inputs_embeds = None
if self.supports_mm_inputs and self.is_first_pp_rank:
    # 多模态嵌入处理 ...
    inputs_embeds = self.model_state.get_mm_embeddings(
        scheduler_output.scheduled_encoder_inputs, input_batch
    )
    # 如果模型不需要原始 token IDs（如输入全是嵌入），则将 input_ids 置为 None
    if inputs_embeds is not None and not self.model.requires_raw_input_tokens:
        input_ids = None

model_inputs = {
    "input_ids": input_ids, # 可能是 None
    "positions": input_batch.positions,
    "inputs_embeds": inputs_embeds,
    "intermediate_tensors": None, # 始终提供默认值，避免某些模型缺少默认参数
    **self.model_state.prepare_inputs(input_batch, self.req_states),
}
# 对非首 PP rank，更新为 None 并准备 intermediate_tensors
if not self.is_first_pp_rank:
    model_inputs["input_ids"] = None
    model_inputs["inputs_embeds"] = None
    # 准备 intermediate_tensors...
```

vllm/v1/worker/gpu/mm/encoder_runner.py

一行修正：将 is_mm_embed 的赋值改为或赋值，避免多区域重复覆盖，修复多模态嵌入标志累积错误。

```
# 在 gather_mm_embeddings 方法中，累加 is_mm_embed 标志
# 修复前：直接赋值会覆盖之前的嵌入标记
# is_mm_embed[req_start_pos + start_idx : req_start_pos + end_idx] = (
# 修复后：使用或赋值，累加多个区域的嵌入标记
is_mm_embed[req_start_pos + start_idx : req_start_pos + end_idx] |= (
    True if is_embed is None else is_embed
)
```

评论区精华

审核人 yewentao256 最初要求提供复现示例和错误信息，看到多个修复合并在一个 PR 后表示理解并批准。另一个讨论线程关注 CUDA 图捕获时是否需要额外处理 input_ids 和 intermediate_tensors，作者 njhill 回应称 prepare_dummy_inputs 已处理相关逻辑。最终获得两位审核人批准。

- 要求描述中补充复现命令和错误信息 (question): 作者未在评论中直接回复，但 PR body 已包含足够描述，随后 yewentao256 批准。
- CUDA 图捕获时 input_ids 和 intermediate_tensors 的处理 (correctness): njhill 回复称 prepare_dummy_inputs 已处理这些情况，无需额外修改。

风险与影响

- 风险：虽然变更量少，但存在以下风险： 1) input_ids 置 None 可能影响依赖原始 ID 的下游逻辑，需确保 requires_raw_input_tokens 标志设置正确。 2) is_embed 改为 |= 若原本覆盖行为是被依赖的，则可能引入回归。 3) Whisper 的 num_reqs 改为 padded 值，若其他模式（非 full cudagraph）也需要 padded 但未使用，可能失效。 4) external_launcher + PP 回退到 MRV1 是临时方案，可能隐藏 V2 的真正限制。整体风险较低，因为变更被 CI 验证。
  - 影响：受益模型包括 CohereASR、Whisper 及其他需要 intermediate_tensors 默认参数的模型。受益配置为 external_launcher 与流水线并行组合。这些修复使 MRV2 能覆盖更广泛的模型和部署场景，推动 V2 模型运行器的生产就绪程度。
  - 风险标记：核心路径变更，缺少测试覆盖，多模态累加逻辑修复，回退到旧版本

关联脉络

- PR #44443 [Core][ModelRunnerV2] Add model runner v2: 本 PR 修复了从此 PR 的 CI 失败中发现的多个 MRV2 兼容性问题。