PR #44568 完整报告

vllm-project/vllm

[Model Runner V2] Fix v2 `AttributeError: 'CohereASRDecoder' object has no attribute 'embed_input_ids'`

合并时间：2026-06-11 02:06

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44568

执行摘要

- 一句话：修复 V2 Model Runner 下 CohereASRDecoder 缺少 embed_input_ids 属性
- 推荐动作：该 PR 作为 bugfix 直接且安全，建议合并。对于架构名称的硬编码，未来可考虑更抽象的接口（如 is_encoder_decoder 或 supports_embed_input_ids 属性），但当前最小化修复策略合理。

功能与动机

修复在 V2 Model Runner 下使用 pytest tests/models/test_initialization.py::test_can_initialize_large_subset[CohereAsrForConditionalGeneration] 测试时引发的 AttributeError: 'CohereASRDecoder' object has no attribute 'embed_input_ids'，该错误会导致模型初始化失败。

实现拆解

1. 在 vllm/v1/worker/gpu/model_states/__init__.py 中修改 init_model_state 函数的条件判断。
2. 将原有的仅检查 WhisperForConditionalGeneration 架构的条件扩展为同时检查 CohereAsrForConditionalGeneration 架构。
3. 当检测到 CohereAsrForConditionalGeneration 架构时，复用 WhisperModelState 来处理模型状态初始化，从而获得 embed_input_ids 实现。
4. 该改动确保了 CohereASRDecoder 在 V2 Model Runner 的 _dummy_run 和 execute_model 流程中能正确调用 embed_input_ids，避免了 AttributeError。

关键文件：

- vllm/v1/worker/gpu/model_states/__init__.py（模块 模型状态；类别 source；类型 data-contract）：唯一的变更文件，修改了 init_model_state 函数的分发逻辑，将 CohereAsrForConditionalGeneration 纳入 WhisperModelState 路径。

关键符号：未识别

关键源码片段

vllm/v1/worker/gpu/model_states/__init__.py

唯一的变更文件，修改了 init_model_state 函数的分发逻辑，将 CohereAsrForConditionalGeneration 纳入 WhisperModelState 路径。

```
# vllm/v1/worker/gpu/model_states/__init__.py

import torch
import torch.nn as nn

from vllm.config import VllmConfig
from vllm.v1.worker.gpu.mm.encoder_cache import EncoderCache


def init_model_state(
    vllm_config: VllmConfig,
    model: nn.Module,
    encoder_cache: EncoderCache | None,
    device: torch.device,
):
    # 将 Whisper 和 CohereASR 模型统一分发到 WhisperModelState，
    # 因为 CohereASRDecoder 缺少 embed_input_ids 方法，
    # 而 WhisperModelState 提供了该方法的实现。
    if (
        "WhisperForConditionalGeneration" in vllm_config.model_config.architectures
        or "CohereAsrForConditionalGeneration" in vllm_config.model_config.architectures
    ):
        from vllm.v1.worker.gpu.model_states.whisper import WhisperModelState
        return WhisperModelState(vllm_config, model, encoder_cache, device)

    if vllm_config.model_config.is_hybrid:
        from vllm.v1.worker.gpu.model_states.mamba_hybrid import MambaHybridModelState
        return MambaHybridModelState(vllm_config, model, encoder_cache, device)

    from vllm.v1.worker.gpu.model_states.default import DefaultModelState
    return DefaultModelState(vllm_config, model, encoder_cache, device)
```

评论区精华

在 review 中，njhill 询问是否还有其他模型存在类似问题，并探讨是否有更通用的判断方式（如根据 config 属性）。贡献者 yewentao256 回应：已检查 funasr.py、fireredasr2.py、fireredlid.py 等模型，它们都实现了 embed_input_ids，不存在此问题。如果使用 supports_transcription_only 等属性判断可能影响其他模型，因此选择当前的最小且精准的修复方案。

- 是否存在其他缺失 embed_input_ids 的模型，以及是否有更通用的判断方式 (design): 当前修复为最小且精准的方案，暂无其他模型需要处理。

风险与影响

- 风险：该 PR 风险极低。仅将 CohereAsrForConditionalGeneration 加入条件分支以复用现存的 WhisperModelState，未改变任何核心逻辑。但需关注：若后续 WhisperModelState 的实现发生变化，CohereASRDecoder 的行为可能意外受影响；另外当前方案依赖架构名称硬编码，未来若引入类似模型需要手动扩充列表。
- 影响：直接影响 CohereAsrForConditionalGeneration 模型在 V2 Model Runner 下的可用性。其他模型无影响。由于 V2 Model Runner 是较新的实验性功能，实际影响面局限于启用该标志的用户。
- 风险标记：依赖架构名称硬编码，修复范围精准

关联脉络

- PR #44443 [Model] Cohere ASR Model Support: 本 PR 修复了 #44443 引入的 Cohere ASR 模型在 V2 Model Runner 下的初始化崩溃问题，是直接的后继 bugfix。