

PR #46759 完整报告

vllm-project/vllm

[Bugfix][MRV2] Forward seq_lens_cpu_upper_bound for mamba hybrid models

合并时间: 2026-06-26 10:03

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46759>

执行摘要

- 一句话: 修复 mamba 混合模型启动崩溃
- 推荐动作: 低风险紧急 bugfix, 建议快速合并。可配合后续 PR 添加 mamba 混合模型的 V2 运行器测试。

功能与动机

MambaHybridModelState.prepare_attn 计算了 seq_lens_cpu_upper_bound 但未转发给 build_attn_metadata, 导致 mamba 元数据构建器收到 None, 并在 warmup 时触发 assert seq_lens_cpu is not None 失败。这阻止了 V2 模型运行器 (VLLM_USE_V2_MODEL_RUNNER=1) 启动任何 mamba/conv-hybrid 模型。

实现拆解

1. 在 vllm/v1/worker/gpu/model_states/mamba_hybrid.py 的 prepare_attn 方法中, 向 build_attn_metadata 调用添加 seq_lens_cpu_upper_bound=seq_lens_cpu_upper_bound 参数, 该变量已在作用域内计算但未传递。
2. 该参数是标准注意力元数据构建所需的字段, 缺失导致下游断言失败。
3. 仅 1 行新增, 无其他变更, 无需测试或配置改动。

关键文件:

- vllm/v1/worker/gpu/model_states/mamba_hybrid.py (模块 模型运行器; 类别 source; 类型 data-contract; 符号 prepare_attn): 核心修复文件, 增加了缺失的参数传递。

关键符号: prepare_attn

关键源码片段

`vllm/v1/worker/gpu/model_states/mamba_hybrid.py`

核心修复文件, 增加了缺失的参数传递。

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

def prepare_attn(self, ...):
    # ... 前面的代码已计算出 seq_lens_cpu_upper_bound ...
    return build_attn_metadata(
```

```
attn_groups=attn_groups,
num_reqs=num_reqs,
num_tokens=num_tokens,
query_start_loc_gpu=input_batch.query_start_loc,
query_start_loc_cpu=query_start_loc_cpu,
max_query_len=max_query_len,
seq_lens=input_batch.seq_lens,
max_seq_len=max_seq_len,
block_tables=block_tables,
slot_mappings=slot_mappings,
kv_cache_config=kv_cache_config,
# 此行是修复：将已计算的 seq_lens_cpu_upper_bound 显式传递,
# 之前遗漏导致下游 builder 收到 None 并触发 assert 失败。
seq_lens_cpu_upper_bound=seq_lens_cpu_upper_bound,
dcp_local_seq_lens=input_batch.dcp_local_seq_lens,
model_specific_attn_metadata=mamba_attn_metadata,
for_cudagraph_capture=for_capture,
)
```

评论区精华

无 review 评论。审核者 njhill 和 yewentao256 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅添加一行参数传递，变量已在作用域内，不会引发新问题。但缺乏对应测试（如 mamba 混合模型集成测试），回归风险依赖现有 CI 覆盖。
- 影响：影响范围：仅影响使用 V2 模型运行器 (VLLM_USE_V2_MODEL_RUNNER=1) 的 mamba/conv-hybrid 模型用户。修复后这些模型可正常启动并推理，例如 LiquidAI/LFM2.5-230M。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR