PR #47428 完整报告

vllm-project/vllm

[ModelRunner V2] Fix Mamba2 crash on non-spec-decode

合并时间：2026-07-02 22:05

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47428

执行摘要

- 一句话：修复 V2 ModelRunner 中 Mamba2 非投机解码崩溃
- 推荐动作：值得合并的快速 Bug 修复，逻辑清晰且风险低。对于关注 V2 ModelRunner 稳定性或 Mamba2 模型的团队，建议了解此修复。

功能与动机

当 Mamba2 混合模型（如 GraniteMoeHybrid）在 V2 ModelRunner 上运行时，由于非投机解码路径下未正确限制 num_accepted_tokens 的填充，导致 selective_state_update 接收形状不匹配的张量并崩溃。PR 描述指出该 Bug 由 Claude 辅助发现。

实现拆解

在 vllm/v1/worker/gpu/model_states/mamba_hybrid.py 文件的 prepare_attn 方法中，将条件 if not for_capture: 修改为 if not for_capture and self.vllm_config.num_speculative_tokens > 0:。这确保了 num_accepted_tokens 和 num_decode_draft_tokens_cpu 仅在投机解码启用时被填充，从而避免非投机解码路径下的形状错误和断言失败。

关键文件：

- vllm/v1/worker/gpu/model_states/mamba_hybrid.py（模块 模型状态；类别 source；类型 data-contract）：核心变更文件，修复了 prepare_attn 方法中的条件判断，避免非投机解码路径下错误填充 tensor。

关键符号：未识别

关键源码片段

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

核心变更文件，修复了 prepare_attn 方法中的条件判断，避免非投机解码路径下错误填充 tensor。

```
# file: vllm/v1/worker/gpu/model_states/mamba_hybrid.py
# 修复前：if not for_capture:
# 修复后：if not for_capture and self.vllm_config.num_speculative_tokens > 0:
# 确保非投机解码路径下不填充投机相关 tensor，避免形状错误
if not for_capture and self.vllm_config.num_speculative_tokens > 0:
```

```
num_accepted_tokens = self.num_accepted_tokens_gpu.new_ones(num_reqs)
num_accepted_tokens[:input_batch.num_reqs] = self.num_accepted_tokens_gpu[
    input_batch.idx_mapping
]

# GDN 使用 >= 0 选择投机解码行，非解码行需要 -1 哨兵值
num_decode_draft_tokens_np = np.full(num_reqs, -1, dtype=np.int32)
if input_batch.num_draft_tokens_per_req is not None:
    has_draft_tokens = input_batch.num_draft_tokens_per_req > 0
    spec_decode_mask = has_draft_tokens & ~input_batch.is_prefilling_np
    num_decode_draft_tokens_np[:input_batch.num_reqs] = np.where(
        spec_decode_mask, input_batch.num_draft_tokens_per_req, -1
    )
num_decode_draft_tokens_cpu = torch.from_numpy(num_decode_draft_tokens_np)
```

评论区精华

无 review 讨论，由 claude[bot] 自动评论且未触发实际审查，yewentao256 和 mgoin 直接 LGTM 批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅在一行条件中添加了额外的投机解码检查，且与 V1 运行时的行为一致。不会影响投机解码路径，非投机解码路径中 Mamba2 相关的断言不再被触发。
- 影响：影响范围有限：仅修复 Mamba2 混合模型在 V2 ModelRunner 上的崩溃。对非 Mamba2 模型无影响，对投机解码场景无影响。团队无需额外操作。
- 风险标记：暂无

关联脉络

- PR #44443 [ModelRunner V2] Enable by default for all dense models: 同一 V2 ModelRunner 功能线的 PR，该 PR 的 Bug 修复正是为了支持 V2 模型运行的稳定性。