PR #46753 完整报告

vllm-project/vllm

[ModelRunner V2] Fix cross-attention block table sizing

合并时间：2026-06-27 07:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46753

执行摘要

- 一句话：修复 MRV2 跨注意力块表大小不足问题
- 推荐动作：该 PR 变更简单但至关重要，建议所有使用 MRV2 且涉及编码器 - 解码器模型的用户合入。值得关注的是 scheduler_config.max_num_encoder_input_tokens 作为新数据源的使用，体现了 V2 架构对调度器配置的依赖。

功能与动机

PR body 指出："Without this there can be OOB errors with e.g. Nemotron-Parse"。此外，关联 CI 失败（https://buildkite.com/vllm/ci/builds/74149#019efb55-7f93-4e20-8d30-0d7adc592b13）进一步证实缺少此修复会导致运行时错误。

实现拆解

1. 在 vllm/v1/worker/gpu/model_runner.py 的 initialize_kv_cache 方法中，当模型为编码器 - 解码器（self.is_encoder_decoder）时，计算 block_table_max_model_len 时除了取 self.max_model_len 和 hf_config.max_source_positions 的最大值，额外加入 self.scheduler_config.max_num_encoder_input_tokens 作为上限。
2. 这一修改确保跨注意力块表能够索引所有编码器 token，无论 hf_config.max_source_positions 是否准确或缺失，都能利用调度器配置中的实际编码器输入上限。
3. 变更仅涉及一行加法与两行注释润色，无其他文件改动。

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 initialize_kv_cache）：MRV2 模型运行器，核心变更位置：在 initialize_kv_cache 中补充了跨注意力块表大小计算的数据源。

关键符号：initialize_kv_cache

关键源码片段

vllm/v1/worker/gpu/model_runner.py

MRV2 模型运行器，核心变更位置：在 initialize_kv_cache 中补充了跨注意力块表大小计算的数据源。

```
# vllm/v1/worker/gpu/model_runner.py
# 在 initialize_kv_cache 中，当模型是编码器 - 解码器时，
# 计算块表所需的最大模型长度，确保跨注意力块表能容纳所有编码器 token。
# 关键变更：新增 scheduler_config.max_num_encoder_input_tokens 作为上限参考。

block_table_max_model_len = self.max_model_len
if self.is_encoder_decoder:
    # Cross-attention block tables need to index encoder tokens, which
    # can exceed the decoder's max_model_len.
    block_table_max_model_len = max(
        block_table_max_model_len,
        self.scheduler_config.max_num_encoder_input_tokens, #
        新增：利用调度器中的编码器输入上限
        getattr(self.model_config.hf_config, "max_source_positions", 0),
    )

block_sizes = []
max_num_blocks_per_group = []
for kv_cache_group in kv_cache_config.kv_cache_groups:
    spec = kv_cache_group.kv_cache_spec
    block_sizes.append(spec.block_size)
    max_num_blocks = cdiv(
        block_table_max_model_len, spec.block_size * self.dcp_size
    )
    # ... 后续对齐和 Mamba 处理 ...
    max_num_blocks_per_group.append(max_num_blocks)
```

评论区精华

Review 中 yewentao256 提问 "What is the CI failure this PR is going to fix?"，作者 njhill 给出了具体的 Buildkite 链接。之后 WoosukKwon 和 yewentao256 均审批通过。无其他讨论。

- CI 失败关联 (question): 确认修复了一个已知的 CI 失败，涉及 Nemotron-Parse 模型的越界错误。

风险与影响

- 风险：风险极低：变更仅新增一个 max() 参数，不会破坏现有非编码器 - 解码器模型的逻辑（在 is_encoder_decoder 分支内）。但由于缺乏对应测试文件变更，回归风险需依靠已有 CI 覆盖。建议后续补充针对编码器 - 解码器场景的 MRV2 集成测试。
- 影响：直接影响在使用 ModelRunner V2 的编码器 - 解码器模型（如 Nemotron-Parse、Whisper）上，修复了潜在的越界错误。对解码器 -only 模型无影响。影响范围小，但修复了关键正确性问题。
- 风险标记：缺少测试覆盖

关联脉络

- PR #46437 [Frontend][Gpt-oss] Use process_eos() to flush Harmony Parser outputs.: Nemotron-Parse 可能使用了 Harmony 解析器，该 PR 也与解析器输出刷新相关，但直接关联性较弱。
- PR #46771 [ModelRunner V2] Update scheduler tests to cover MRV2 paths: 和本 PR 同属 MRV2 改进线，前者补充调度器测试，本 PR 修复 MRV2 的 bug，两者共同提升 MRV2 的稳定性。