PR #47071 完整报告

vllm-project/vllm

[Bugfix] Fix pooled Whisper sliding-window KV sizing

合并时间：2026-07-01 17:33

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47071

执行摘要

- 一句话：修复 pooled Whisper 滑动窗口 KV 缓存大小错误
- 推荐动作：值得精读。这是一个典型的数据契约 bug：同一个物理量在不同抽象层次使用了不同单位，导致资源预留错误。设计决策是将缩放点放在 get_kv_cache_spec 中，而不影响注意力内核，保持了关注点分离。

功能与动机

Voxtral Realtime 的 causal Whisper 编码器使用 block_pool_size 个编码器 token 组成一个 pooled KV 块，但其 SlidingWindowSpec.sliding_window 仍以编码器 token 为单位表达。滑动窗口 KV 缓存管理器随后为编码器缓存预留了大约 block_pool_size 倍过多的块。

实现拆解

1. 导入调整：在 whisper_causal.py 中添加 from vllm.utils.math_utils import cdiv 和 from vllm.v1.kv_cache_interface import SlidingWindowSpec。
2. 核心修复：在 PooledBlockPooledAttention.get_kv_cache_spec 方法中，当返回的 kv_cache_spec 是 SlidingWindowSpec 类型时，使用 cdiv(sliding_window, block_pool_size) 将其 sliding_window 字段缩放到 pooled 块单位。
3. 测试增强：在 test_voxtral_realtime.py 中新增 assert_encoder_kv_cache_spec 函数，验证编码器 KV 缓存 spec 的 sliding_window 为 cdiv(750, 4) == 188，max_admission_blocks_per_request 返回正确值（13）。同时修改 engine fixture 以禁用多进程，使得测试能够直接访问 model_executor。

关键文件：

- vllm/model_executor/models/whisper_causal.py（模块 模型执行器；类别 source；类型 data-contract）：核心修复文件：在 PooledBlockPooledAttention.get_kv_cache_spec 中缩放 sliding_window 为 pooled 块单位。
- tests/models/multimodal/generation/test_voxtral_realtime.py（模块 测试；类别 test；类型 test-coverage；符号 assert_encoder_kv_cache_spec, engine）：新增测试函数 assert_encoder_kv_cache_spec 验证修复后的 KV 缓存 spec 正确性，并修改 engine fixture 以支持访问 model_executor。

关键符号：PooledBlockPooledAttention.get_kv_cache_spec

关键源码片段

vllm/model_executor/models/whisper_causal.py

核心修复文件：在 PooledBlockPooledAttention.get_kv_cache_spec 中缩放 sliding_window 为 pooled 块单位。

```
# 文件：vllm/model_executor/models/whisper_causal.py（片段）
# 在 PooledBlockPooledAttention 类中，重写 get_kv_cache_spec 方法

def get_kv_cache_spec(self, vllm_config: VllmConfig):
    kv_cache_spec = super().get_kv_cache_spec(vllm_config)
    assert isinstance(kv_cache_spec, AttentionSpec)
    # 先按原有逻辑缩放 num_kv_heads
    kv_cache_spec = replace(
        kv_cache_spec,
        num_kv_heads=self.block_pool_size * kv_cache_spec.num_kv_heads,
    )
    # 新增：如果 spec 是 SlidingWindowSpec，将其 sliding_window 也缩放到 pooled 单位
    # 注意：这里只影响 KV 缓存管理器的块预留，不影响注意力内核的窗口（内核使用原始窗口）
    if isinstance(kv_cache_spec, SlidingWindowSpec):
        kv_cache_spec = replace(
            kv_cache_spec,
            sliding_window=cdiv(kv_cache_spec.sliding_window, self.block_pool_size),
        )
    return kv_cache_spec
```

tests/models/multimodal/generation/test_voxtral_realtime.py

新增测试函数 assert_encoder_kv_cache_spec 验证修复后的 KV 缓存 spec 正确性，并修改 engine fixture 以支持访问 model_executor。

```
# 文件：tests/models/multimodal/generation/test_voxtral_realtime.py（片段）
# 新增验证函数，用于检查修复后的 KV 缓存 spec

def assert_encoder_kv_cache_spec(engine: LLM) -> None:
    vllm_config = engine.llm_engine.vllm_config
    audio_config = vllm_config.model_config.hf_config.audio_config
    kv_cache_specs_per_rank = engine.llm_engine.model_executor.get_kv_cache_specs()

    assert len(kv_cache_specs_per_rank) == 1
    kv_cache_specs = kv_cache_specs_per_rank[0]
    # 断言音频编码器层的 spec 存在
    assert AUDIO_LAYER_NAME in kv_cache_specs, kv_cache_specs.keys()
    spec = kv_cache_specs[AUDIO_LAYER_NAME]

    # 验证已知配置值
    assert audio_config.sliding_window == 750
    assert audio_config.block_pool_size == 4
    assert isinstance(spec, SlidingWindowSpec)
    assert spec.block_size == 16
```

```
assert spec.num_kv_heads == 128
# 核心断言：缩放后的 sliding_window 应为 cdiv(750, 4) = 188
assert spec.sliding_window == cdiv(750, 4) == 188
# 验证最大准入块数计算正确
assert (
    spec.max_admission_blocks_per_request(
        max_num_batched_tokens=1,
        max_model_len=vllm_config.model_config.max_model_len,
    )
    == 13
)
```

评论区精华

无 review 评论。PR 由 NickLucche 批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅限于 PooledBlockPooledAttention.get_kv_cache_spec 方法中的滑动窗口缩放逻辑，并且有新增的测试覆盖。主要风险是如果 block_pool_size 不能整除 sliding_window，使用 cdiv 可能导致窗口略微扩大，但这比之前预留过多块更合理。
- 影响：影响范围限于使用 PooledBlockPooledAttention 和 SlidingWindowSpec 的模型（目前为 Voxtral Realtime）。修正后，滑动窗口 KV 缓存将准确预留块，减少 GPU 内存浪费，提高内存利用率。
- 风险标记：核心路径变更

关联脉络

- 暂无明显关联 PR