PR #49403 完整报告

vllm-project/vllm

[BugFix] Increase the max supported duration for MOSS-TD

合并时间：2026-07-25 22:48

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49403

执行摘要

- 一句话：修复 MOSS-TD 长音频被拒绝问题
- 推荐动作：建议阅读此 PR 以了解基于时长估算 token 的通用模式；对 MOSS-TD 用户来说这是关键修复。但需注意缺少测试，建议后续补充相应测试。

功能与动机

之前代码使用 feature_extractor.chunk_length 估算最大音频时长，但 Whisper 的 chunk_length 是音频切片的持续时间，不是完整音频的最大长度，导致 max_tokens_per_mm_item 只有约 375 个 token，进而 encoder_cache_size 回退到 max_num_batched_tokens（如 8192），超过 8192 个 embedding token 的长音频请求被拒绝。

实现拆解

1. 定义最大音频时长常量：在文件顶部新增 MAX_AUDIO_DURATION_S = 90 * 60，表示 MOSS-TD 支持的最大音频时长为 90 分钟。
2. 简化 _get_max_audio_samples：移除原来依赖 feature_extractor 属性的逻辑（检查 chunk_length 并回退到 n_samples），直接返回 MAX_AUDIO_DURATION_S * feature_extractor.sampling_rate，即 90 分钟对应的采样点数。
3. 间接影响 encoder cache 容量：通过 get_max_tokens_per_mm_item 等上游函数调用 _get_max_audio_samples，使 max_tokens_per_mm_item 提高到约 67500 个 token，从而 encoder_cache_size = max(max_num_batched_tokens, max_tokens_per_mm_item) 能容纳更长音频。
4. 无测试或配置文件变更：仅修改源码逻辑，未引入单元测试或 e2e 测试。

关键文件：

- vllm/model_executor/models/moss_transcribe_diarize.py（模块 模型层；类别 source；类型 data-contract）：核心改动文件：新增 MAX_AUDIO_DURATION_S 常量并简化 _get_max_audio_samples 函数，直接修复了长音频被拒绝的 bug。

关键符号：_get_max_audio_samples

关键源码片段

vllm/model_executor/models/moss_transcribe_diarize.py

核心改动文件：新增 MAX_AUDIO_DURATION_S 常量并简化 _get_max_audio_samples 函数，直接修复了长音频被拒绝的 bug。

```
# vllm/model_executor/models/moss_transcribe_diarize.py

# 新增：MOSS-TD 支持的最大音频时长 90 分钟
MAX_AUDIO_DURATION_S = 90 * 60

# ...

def _get_max_audio_samples(feature_extractor: Any) -> int:
    """
    返回 MOSS-TD 支持的最大音频采样点数。
    原逻辑使用 feature_extractor.chunk_length (Whisper 的 30 秒 chunk)
    作为最大时长，导致长音频被拒绝；现在改为硬编码 90 分钟。
    """
    return int(MAX_AUDIO_DURATION_S * feature_extractor.sampling_rate)
```

评论区精华

仅有自动化 bot claude[bot] 评论提示该 PR 来自 fork 且未触发自动 review，无人工 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 最大时长硬编码：MAX_AUDIO_DURATION_S 被硬编码为 90 分钟，若后续模型支持更长音频需要手动修改常量，缺乏配置化支持。
  2. 内存使用增加：允许更大 max_tokens_per_mm_item 会导致 encoder_cache_size 膨胀，可能增加显存占用，但仅在用户输入超过原限制的长音频时实际分配。
  3. 无测试覆盖：该变更缺少对应的单元测试或集成测试，对回归风险缺少验证。 - 影响：影响范围：仅 MOSS-TD 单个模型文件，一行常量新增 + 一行函数简化，变更极小。用户影响：修复了长音频（超过 30 秒但小于 90 分钟）被拒绝的错误，对使用该模型做长音频转写的用户是重要改进。系统影响：无，因为变更仅在模型特定的多模态处理流程中生效，不影响其他模型或系统全局行为。 - 风险标记：缺少测试覆盖，硬编码常量

关联脉络

- 暂无明显关联 PR