PR #38018 完整报告

vllm-project/vllm

[Model] Use helper function to run MM processors with token inputs (where applicable)

合并时间：2026-03-26 16:44

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38018

执行摘要

- 一句话：为多模态处理器引入助手函数，避免在 token-only 输入时生成虚拟文本。
- 推荐动作：建议团队精读此 PR，重点关注 call_hf_processor_mm_only 的设计决策，以及各模型特定重写（如 pixtral.py 中的 batch 维度处理）的逻辑，以理解多模态输入处理的最佳实践和潜在风险。

功能与动机

根据 PR body，定义助手函数 call_hf_processor_mm_only 以替换 ProcessorMixin.__call__，从而处理 token 输入，因为不是所有 HF 处理器都支持空文本，避免生成 dummy text。

实现拆解

实现主要分为三部分：1) 在 vllm/transformers_utils/processor.py 中新增 call_hf_processor_mm_only 函数，处理多模态输入（图像、视频、音频）；2) 在多个模型文件（如 keye.py、qwen2_5_vl.py）中重写 _call_hf_processor 方法，以使用文本路径或处理特定逻辑（如视频、音频）；3) 在 vllm/multimodal/processing/processor.py 中修改 _apply_hf_processor_mm_only 方法，集成新助手函数，并优化空输入处理。同时，修复了 Isaac、Pixtral、Voxtral 等处理器的兼容性问题。

关键文件：

- vllm/transformers_utils/processor.py（模块 transformers_utils）：新增核心助手函数 call_hf_processor_mm_only，用于处理多模态输入，避免虚拟文本生成。
- vllm/multimodal/processing/processor.py（模块 multimodal/processing）：修改 _apply_hf_processor_mm_only 方法，集成新助手函数，优化 token-only 输入处理逻辑。
- vllm/model_executor/models/qwen2_5_vl.py（模块 model_executor/models）：重写 _call_hf_processor 方法，以使用视频特定逻辑，避免 token 路径问题。
- vllm/model_executor/models/pixtral.py（模块 model_executor/models）：重写 _call_hf_processor 方法，处理 batch 维度缺失问题，影响核心模型处理。

关键符号：call_hf_processor_mm_only, _call_hf_processor, _apply_hf_processor_mm_only

评论区精华

review 中，gemini-code-assist[bot] 在 vllm/transformers_utils/processor.py:547 指出，在 call_hf_processor_mm_only 函数中直接使用 pop('attention_mask') 可能导致 KeyError，建议检查键存在性。作者 DarkLight1337 回应无需修改，遵循 HF 实现。此风险未解决，可能影响代码健壮性。

- KeyError 风险 in call_hf_processor_mm_only (correctness): 风险未解决，作者坚持原实现，可能影响代码稳定性。

风险与影响

- 风险：技术风险包括：1) 在 call_hf_processor_mm_only 函数中潜在 KeyError，如果 attention_mask 不存在于 audio_inputs 中；2) 修改多个模型处理器（如 keye.py、pixtral.py）可能引入回归错误，特别是视频或音频特定逻辑；3) 依赖 Hugging Face 实现可能导致兼容性问题，如不同 Transformers 版本行为差异；4) 在 context.py 和 processor.py 中对 return_tensors 的改动可能影响其他处理路径。
- 影响：影响范围：对用户透明，提升多模态处理性能，避免虚拟文本生成，减少计算开销；系统层面，更高效处理 token-only 输入，增强多模态模型支持；团队需要更新对新助手函数和重写逻辑的理解，并在未来模型集成中应用此模式。
- 风险标记：潜在 KeyError，多模型变更风险，依赖外部实现

关联脉络

- PR #38119 [MultiModal] add support for numpy array embeddings: 同样涉及多模态改进，共享多模态处理逻辑，可能相互影响。
- PR #38127 Various Transformers v5 fixes: 涉及 Transformers 兼容性修复，与本 PR 的 HF 处理器依赖相关。