

PR #51827 完整报告

vllm-project/vllm

[3/N] Harden Transformers modelling backend multi-modal path

合并时间: 2026-08-20 18:24

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51827>

执行摘要

- 一句话: 按 transformers 版本拆分多模态处理器并启用缓存
- 推荐动作: 值得精读, 特别是关注多模态处理器与上游 transformers 的契约设计: 版本门控双处理器、把 HF 展开声明为 PromptReplacement 以复用统一缓存路径、通过 unpad 消除批量 padding 对缓存可复用性的破坏、错误信息直接指导用户上报上游或固定版本。这些设计对理解 vLLM 多模态输入流水线以及后续如何与 HuggingFace 生态协同演进很有价值。

功能与动机

PR body 说明这是对 #51408 和 #51657 的延续, 取代 #50107 和 #50899, 目标是把 Transformers modeling backend 的多模态路径加固到可缓存、可依赖的程度。关键动机包括: 重复图片应命中处理器缓存 ("A repeated image now hits the cache rather than being reprocessed and re-shipped over IPC"); 处理器按 batch 公共形状 padding 使单条数据依赖批内其他数据而不可缓存 ("an item's data depend on what it was processed beside and so uncacheable"); 以及修复 Idefics3 返回 `[1, total_patches, C, H, W]` 时 `flat_from_sizes` 在错误维度切片导致的静默错误 ("Image fields with a leading batch dimension were sliced on the wrong axis")。

实现拆解

1. 处理器按版本分流并抽取公共基类: `vllm/model_executor/models/transformers/multimodal.py` 新增 transformers 与 `packaging.version` 导入, 按 `Version(transformers.__version__)` 选择 `LegacyMultiModalProcessor` (低于 5.15.0) 或 `OffsetsMultiModalProcessor` (5.15.0 及以上); 两者公共逻辑 (调用 HF processor、尺寸计算、输出按模态归位) 上提为 `_MultiModalProcessorBase`, 占位符定位策略留给子类实现。
2. Offsets 路径接入处理器缓存: `OffsetsMultiModalProcessor._get_prompt_updates` 把 HF 报告的 `text_replacement_offsets` 逐个转成 `PromptReplacement`, 因此走基类 `apply()` 的缓存路径而非重写 `apply()`; 新增 `_get_embed_token_id` (取替换 token 众数作为承载 embedding 的 token)、`_get_slice_dim` (处理 leading batch 维)、`_unpad_images/_unpad_audios` (去掉 Emu3 等按 batch 最大形状 padding 造成的尺寸依赖), `num_image_patches` 在 offsets 路径上改从 processor 的 `_get_num_multimodal_tokens` 获取而非逐行猜测。

3. 行为变更与错误提示：从 5.15.0 起，不报告替换 offsets 的 processor 被直接拒绝，错误信息点名 processor 与其缺失的 replace_<modality>_token hook，并引导用户上报上游或固定 transformers<5.15.0；纯文本提示词走 _apply_hf_processor_text_only（传空 mm_data），只有纯模态提示词才使用 dummy text（_apply_hf_processor_mm_only）。
4. Bug 修复：修复 Idefics3 pixel_values=[1, total_patches, C, H, W] 在 flat_from_sizes 上切错轴的问题；Idefics3Processor 拒绝含 <image> 但无图片的提示词时回退 tokenizer（与 idfics3.py 现有处理一致）；Emu3 在未展开 prompt 上丢失 BOS token，标记 xfail 等待上游修复。
5. 测试配套：新增 tests/models/multimodal/processing/transformers_backend.py，提供 create_processor/create_cached_processor 直接构造两类处理器（绕过注册表的版本门控，保证两条路径都有覆盖），并新增 offsets_only 标记与 PROCESSOR_CLASSES 参数化；图像、音频测试全部改为双路径参数化，新增缓存命中断言（make_stats().hits）、拼接 prompt 与 HF 展开一致性、Emu3 padding 去除、嵌套图像字段按图拆分、Gemma3 结构 token 掩码等测试。

关键文件：

- vllm/model_executor/models/transformers/multimodal.py（模块 多模态处理；类别 source；类型 data-contract；符号 _get_embed_token_id, _get_audio_token_id, MultiModalProcessor, _get_prompt_updates）：核心源码文件：新增 _MultiModalProcessorBase 公共基类、LegacyMultiModalProcessor 与 OffsetsMultiModalProcessor 版本分流、_get_embed_token_id、_get_slice_dim、缓存路径与 unpad 逻辑，并包含 Idefics3 切片修复与无 offsets 拒绝的行为变更。
- tests/models/multimodal/processing/test_transformers_image.py（模块 图像测试；类别 test；类型 test-coverage；符号 test_multimodal_processor, _process_two_images, test_image_multiple_inputs, test_image_adjacent_inputs）：图像侧测试全套参数化为 legacy/offsets 双路径，新增缓存命中、Emu3 padding 去除、Gemma3 结构 token 掩码等关键用例。
- tests/models/multimodal/processing/test_transformers_audio.py（模块 音频测试；类别 test；类型 test-coverage；符号 test_audio_multimodal_processor, _process_granite_speech, test_audio_multiple_inputs, process）：音频侧测试双路径参数化，多音频输入按占位符逐一拆分，Voxtral 的 xfail 原因更新为新行为。
- tests/models/multimodal/processing/transformers_backend.py（模块 测试工具；类别 test；类型 test-coverage；符号 create_processor, create_cached_processor）：新增测试工具模块：直接构造两类处理器与真实缓存，保证版本门控下两条路径都有 CI 覆盖。

关键符号：_get_embed_token_id, _get_prompt_updates, _get_hf_mm_data, _get_slice_dim, _call_hf_processor, create_processor, create_cached_processor, test_repeated_image_hits_the_processor_cache, test_spliced_prompt_matches_hf_expansion, test_batch_padding_removed_from_image_items

关键源码片段

[vllm/model_executor/models/transformers/multimodal.py](#)

核心源码文件：新增 `_MultiModalProcessorBase` 公共基类、`LegacyMultiModalProcessor` 与 `OffsetsMultiModalProcessor` 版本分流、`_get_embed_token_id`、`_get_slice_dim`、缓存路径与 `unpad` 逻辑，并包含 `Idefics3` 切片修复与无 `offsets` 拒绝的行为变更。

```
# vllm/model_executor/models/transformers/multimodal.py
```

```
# 按 transformers 版本选择多模态处理器：低于 5.15.0 用 Legacy，5.15.0 及以上用 Offsets。
```

```
_MODALITY_SIZE_KEYS = {"audio": "num_audio_tokens", "image": "num_image_patches"}
```

```
def _get_embed_token_id(replacement_ids: torch.Tensor) -> int:
```

```
    """HF 展开里重复的那个 token 就是承载 embedding 的 token，取众数得到它。
```

```
    PromptReplacement 展开时会重复同一 token，mode() 拿到出现次数最多的值，  
    即可作为该模态的 embedding token。
```

```
    """
```

```
    return int(replacement_ids.mode().values)
```

```
class _MultiModalProcessorBase(BaseMultiModalProcessor[MultiModalProcessingInfo]):
```

```
    """两种后端处理器共享的公共流程：调用 HF processor、计算图像尺寸、  
    输出按模态归位；占位符定位策略由子类决定。
```

```
    Legacy 子类：基于 mm_token_type_ids 在扩展后的 prompt 上定位占位符，  
    无法命中缓存；
```

```
    Offsets 子类：把每个展开声明为 PromptReplacement，走基类的缓存路径，  
    重复的图片命中处理器缓存而不是重新处理再走 IPC。
```

```
    """
```

```
def _get_prompt_updates(self, ...):
```

```
    # 子类实现：返回 PromptUpdate 列表，描述 prompt 中每个模态占位符  
    # 如何被展开替换（offsets 路径）或保持原样（legacy 路径）。
```

```
    ...
```

```
def _get_hf_mm_data(self, ...):
```

```
    # 组装喂给 HF processor 的 mm_data，并记录各模态的尺寸信息，  
    # 供后续 flat_from_sizes 按 item 切回独立数据。
```

```
    ...
```

```
def _get_slice_dim(self, field: str) -> int:
```

```
    # 部分 processor（如 Idefics3）返回 [1, total_patches, C, H, W],  
    # 带 leading batch 维，切片必须从 dim 1 开始，否则第一张图会拿到  
    # 每一行的内容而第二张图拿到空张量。
```

```
    ...
```

```
class OffsetsMultiModalProcessor(_MultiModalProcessorBase):
```

```
    """根据 HF processor 报告的 text_replacement_offsets 定位占位符，  
    每个展开声明为一个 PromptReplacement。
```

把展开表达成 `update` 之后，就能从不含模态的原始 `prompt` 出发重建，因此走基类的通用处理路径，也就能享受处理器缓存：重复图片命中缓存，不再重新处理、不再跨 IPC 传输。

"""

```
def _get_prompt_updates(self, ...):
    # 读取 hf_inputs 里的 text_replacement_offsets，逐个构造 PromptReplacement;
    # 按模态取 num_image_patches / num_audio_tokens 决定占位符长度，
    # 并剔除仅供结构化 prompt 使用的非 embedding 辅助 token。
    ...
```

评论区精华

- zucchini-nlp 在 `multimodal.py:763` 询问：offsets 路径是否只有 `cached-processing` 适用？legacy 是否几个月后会被移除？缓存是否是 vLLM 默认模式？hmellor 回答：缓存由 `mm_processor_cache_gb > 0` 开启，默认值是 4，所以新路径默认即走缓存。
- zucchini-nlp 在 `multimodal.py:368` 询问：vLLM 对 text-only 或 mm-only 数据是否用 dummy 方式补缺失模态？hmellor 澄清：`_apply_hf_processor_mm_only` 使用 dummy text，但 `_apply_hf_processor_text_only` 只是把空 `mm_data` 传给 HF processor。
- zucchini-nlp 在 `multimodal.py:437` 以 "🙄" 表达对 legacy/offsets 双路径长期并存的无奈，hmellor 回应 "Hopefully this won't need to coexist for too long"。
- zucchini-nlp 在 `multimodal.py:1156` 提醒：`get_imagevideo_features` 的输出已在 transformers 5.15 起标准化为 `list[torch.Tensor]`（见 [huggingface/transformers#46405](https://huggingface.co/transformers#46405)），需要保持向后兼容；hmellor 确认该 PR 已随 5.15 发布，因此需要保留当前的 normalization 逻辑直到 5.15 成为最低支持版本。
- offsets 路径是否默认启用缓存 (design): hmellor 回答缓存由 `mm_processor_cache_gb > 0` 开启，默认值为 4，因此 offsets 路径默认享受缓存。
- text-only 与 mm-only 是否用 dummy 数据补齐 (design): hmellor 澄清：mm-only 走 `_apply_hf_processor_mm_only` 使用 dummy text，而 text-only 走 `_apply_hf_processor_text_only` 只传空 `mm_data`。
- `get_imagevideo_features` 输出标准化与向后兼容 (correctness): hmellor 确认该上游 PR 已随 5.15 发布，需保留 normalization 直到 5.15 成为最低支持版本。
- legacy 与 offsets 双路径并存周期 (design): hmellor 回应希望两者不需要共存太久。

风险与影响

- 风险：行为变更影响面：5.15.0 及以上不再对无 offsets 的 processor 做兜底搜索，PR 实测 87/145 个 `ProcessorMixin` 子类未实现替换 hook（尽管多数不是生成式多模态模型或已有 vLLM 原生实现），用户升级 transformers 后可能突然遇到拒绝服务的报错。缓存正确性依赖一个强前提：缓存未命中时单样本调用 `image_processor([img_1])` 必须等于批量调用 `image_processor([img_0, img_1])[1]`；`_unpad_images/_unpad_audios` 覆盖了 padding 型 processor，但未覆盖的 processor 若有 batch 依赖输出，缓存可能返回与上下文相关的错误数据。Idefics3 切片修复翻转了此前被 re-batch 掩盖的隐性错误，单图与多图边界的回归需要重点验证。Emu3 BOS token 依赖上游

huggingface/transformers#47924 发布，当前以 xfail 掩盖。此外 legacy 与 offsets 双实现并存，长期存在行为分叉与维护成本。

- 影响：用户侧：在 transformers>=5.15.0 下，重复多模态输入默认命中处理器缓存（测试中 2 个请求 1 次命中），显著降低重复预处理与跨 IPC 传输开销；但部分模型会因无替换 offsets 直接报错，需要固定版本或等待上游。系统侧：处理器缓存默认开启（mm_processor_cache_gb=4），占用额外内存，同时减少 CPU 侧重复视觉 / 音频特征计算。团队侧：测试基础设施新增双路径参数化，vLLM 与 HuggingFace transformers 在处理器接口上双向对齐（offsets 契约、embedding token 推断），后续需在最低支持版本到达 5.15 后清理 legacy 路径。
- 风险标记：行为变更：5.15.0+ 拒绝无 offsets 处理器，缓存正确性依赖单样本与批量处理等价，上游依赖：Emu3 BOS 待 transformers 修复，双处理器路径长期并存维护成本，Idefics3 切片修复回归风险

关联脉络

- PR #51408 Harden Transformers modelling backend multi-modal path（系列前序 [1/N]，见本 PR body）：本 PR 明确说明 Continues from #51408，属于同一加固系列的第一步。
- PR #51657 Harden Transformers modelling backend multi-modal path（系列前序 [2/N]，见本 PR body）：本 PR 明确说明 Continues from #51657，属于同一加固系列的第二步。
- PR #50107 被本 PR 取代的早期尝试（见本 PR body）：PR body 声明本 PR Supersedes #50107，是此前方案被取代。
- PR #50899 被本 PR 取代的早期尝试（见本 PR body）：PR body 声明本 PR Supersedes #50899，是此前方案被取代。
- PR #47924 huggingface/transformers: Leftover processors（修复 Emu3 BOS token）：上游 transformers PR，修复 Emu3 在未展开 prompt 上的 BOS 丢失，本 PR 以 xfail 等待其发布。