

PR #48413 完整报告

vllm-project/vllm

[Bugfix][MM] Fix MiniCPM-V placeholder replacement and image processor loading on Transformers v5

合并时间: 2026-08-06 12:12

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48413>

执行摘要

- 一句话: 修复 MiniCPM-V 在 Transformers v5 下的占位符替换与图像处理器复用
- 推荐动作: 值得精读。三个设计决策有借鉴价值: 一是用 " 先 token 匹配、失败后分段文本编码 " 的双路径策略兼容 BPE 边界差异, 并把 workaround 下沉为公共层工具而非堆在模型文件里; 二是通过 processor_cls_overrides 绕过 AutoImageProcessor 按类名共享缓存的机制, 解决多 checkpoint 类名冲突; 三是在 review 中把 " 复现原始失败 " 作为测试验收标准, test_image_processor_for_dif_model 和 test_prompt_has_dif_BPE_boundaries_in_context 在 main 上失败、分支上通过, 是高质量回归测试的范例。阅读时可重点关注 _apply_prompt_updates 覆写与公共实现之间的重复代码边界, 以及上游 transformers 修复合入后的清理机会。

功能与动机

PR body 说明两个独立但相关的阻塞问题: 其一, 升级 Transformers v5 后 vllm serve 初始化即报 `RuntimeError: Expected there to be 1 prompt placeholders ... but instead found 0 prompt placeholders!`, 根因是 MiniCPM-V 占位符为普通文本而非 special token, BPE 下单独编码与在完整 prompt 上下文中编码的 token 边界不同; 其二, 多个 MiniCPM-V repo 在 preprocessor_config.json 中声明相同的 MiniCPMVImageProcessor 类名但实现各异, 同进程串行加载时后加载模型复用先前模型的处理类, 报 `TypeError: ... got an unexpected keyword argument 'max_slice_nums'`。作者同时移除 `max_transformers_version="4.57"` 上限, 因为 #44282 已 vendor MiniCPMV processor 并将 MiniCPMVBatchFeature 别名到 BatchFeature。

实现拆解

1. 占位符替换修复: 覆写 _apply_prompt_updates

在 vllm/model_executor/models/minicpmv.py 的 MiniCPMVMultiModalProcessor 中覆写 _apply_prompt_updates (+111/-3, 本 PR 源码主改动)。流程为先按 token 匹配 (_apply_token_matches), 若任一占位符未匹配到, 则回退到 _apply_text_matches_as_segmented_tokens: 把 token 序列解码为文本、按段做字符串匹配, 再逐段编码拼回 token ID, 避免整段重编码改变占位符的 BPE 边界。该覆写复用 _find_mm_placeholders 与 _seq2text 等公共工具, 最终返回新 token 序列与 PlaceholderFeaturesInfo 列表。

2. 公共层新增分段编码工具

在 `vllm/multimodal/processing/processor.py` 中新增模块级函数

`apply_text_matches_as_segmented_tokens` 与 `BaseMultiModalProcessor._apply_text_matches_as_segmented_tokens` (+33/-0)。核心思想：先 `_apply_matches` 得到切分后的文本段，再对每一段独立 `_seq2tokens` 编码后 `flatten_2d_lists` 拼接，避免整段重编码时 BPE 跨段合并改变占位符 token。此举把 MiniCPM-V 的 workaround 下沉为公共能力，其他文本占位符模型可复用。

3. 图像处理器按 repo 解析

在 `MiniCPMVProcessingInfo` 中新增 `@cached_property _image_processor_cls`：通过 `ImageProcessingMixin.get_image_processor_dict` 读取仓库配置中的 `auto_map['AutoImageProcessor']`，先 `resolve_trust_remote_code` 显式校验信任开关，再 `get_class_from_dynamic_module` 加载类。`get_hf_processor` 改为：tokenizer 仍走 `ctx.get_hf_processor`，image processor 则通过 `cached_get_image_processor(..., processor_cls_overrides=processor_cls)` 按 repo 特定类实例化，并用 `_merge_mm_kwargs` 归一化 kwargs——既避免 `AutoImageProcessor` 跨版本类名共享错配，也保证每次调用的 `max_slice_nums` 等参数生效。

4. 测试与注册表配套

新增 `tests/models/multimodal/processing/test_minicpmv.py` (+74 行) 三个回归测试：

`test_image_processor_for_dif_model` 复现 2.5 后接 4.0 的类错用；

`test_prompt_has_dif_BPE_boundaries_in_context` 复现 BPE 边界差异；

`test_get_hf_processor_for_same_model_different_kwargs` 验证不同 kwargs 不会复用陈旧实例。`tests/models/registry.py` 删除 MiniCPM-V 的 `max_transformers_version="4.57"` 与对应 reason (-7 行)，使处理测试在 Transformers v5 下运行；2.6 因 gated repo 被跳过。

关键文件：

- `vllm/model_executor/models/minicpmv.py` (模块 模型层；类别 source；类型 core-logic；符号 `_image_processor_cls`, `get_hf_processor`, `_apply_prompt_updates`)：PR 源码主改动：新增 `_image_processor_cls` 按 repo 解析图像处理器类并改造 `get_hf_processor`，新增 `_apply_prompt_updates` 覆写实现 BPE 分段回退，直接修复两个核心问题。
- `vllm/multimodal/processing/processor.py` (模块 多模态处理；类别 source；类型 core-logic；符号 `apply_text_matches_as_segmented_tokens`, `_apply_text_matches_as_segmented_tokens`)：公共多模态处理器层新增分段编码工具，把 MiniCPM-V 的 BPE workaround 下沉为公共能力，是本 PR 影响面最大的复用点。
- `tests/models/multimodal/processing/test_minicpmv.py` (模块 回归测试；类别 test；类型 test-coverage；符号 `test_get_hf_processor_for_same_model_different_kwargs`, `test_image_processor_for_dif_model`, `test_prompt_has_dif_BPE_boundaries_in_context`)：新增 3 个回归测试，分别在 main 上失败、分支上通过，直接复现 2.5/V-4 处理器类错用与 BPE 边界差异两个原始 bug。
- `tests/models/registry.py` (模块 测试配置；类别 test；类型 test-coverage)：移除 MiniCPM-V 的 `max_transformers_version="4.57"` 上限，使 MiniCPM-V 处理测试可在

Transformers v5 下运行，是 v5 兼容目标落地的配套开关。

关键符号：_image_processor_cls, get_hf_processor, _apply_prompt_updates, apply_text_matches_as_segmented_tokens, _apply_text_matches_as_segmented_tokens, test_get_hf_processor_for_same_model_different_kwargs, test_image_processor_for_dif_model, test_prompt_has_dif_BPE_boundaries_in_context

评论区精华

审阅人 tc-mb 与作者 YunzhuLu 围绕缓存设计、测试质量与安全边界进行了多轮交锋：

- 缓存忽略 per-call kwargs: tc-mb 实测 get_hf_processor(max_slice_nums=1) 后再调用 max_slice_nums=2 返回同一实例，要求按 normalized kwargs 键控或直接移除缓存。作者先改为按 _merge_mm_kwargs 结果键控，随后在下一轮 review 中彻底移除该缓存。
- 无界缓存与既有 LRU 重复: tc-mb 指出长运行服务中 per-request kwargs 会让缓存无限增长，而 cached_get_image_processor 已有有界 LRU、cached_property 已记忆类解析，额外缓存和对象同一性测试均不必要。作者最终只保留 @cached_property 记忆类解析。
- 测试应复现原始失败: tc-mb 要求测试覆盖 2.5 后接 4.0 的类错用和 BPE 边界差异 prompt，作者用 test_image_processor_for_dif_model 与 test_prompt_has_dif_BPE_boundaries_in_context 替换了基于对象同一性的断言，两者在 main 上失败、分支上通过。
- 公共层复用与安全: tc-mb 指出 workaround 复制了公共处理器私有逻辑、类引用应来自 `auto_map.AutoImageProcessor` 而非硬编码，且 `get_class_from_dynamic_module` 会执行仓库代码，必须显式强制 `trust_remote_code`。作者将逻辑下沉至 `processor.py::apply_text_matches_as_segmented_tokens`，类引用改为从配置读取并强制信任校验，还顺带发现 transformers 上游 `get_image_processor_config` 丢失 `auto_map` 的 bug（已提 upstream PR#47628）。
- 新增 _minicpmv_hf_processor_cache 忽略 per-call mm_processor_kwargs (correctness): 作者先改为按 _merge_mm_kwargs 归一化后的 kwargs 键控，后续 review 中彻底移除该缓存，仅以 cached_property 记忆类解析。
- 无界缓存与既有 LRU 缓存重复 (design): 删除 _minicpmv_hf_processor_cache，保留 cached_get_image_processor 的 LRU 与 cached_property 类解析记忆，避免长运行服务缓存膨胀。
- 测试应复现原始失败场景 (testing): 新增 test_image_processor_for_dif_model 与 test_prompt_has_dif_BPE_boundaries_in_context，作者确认两者在 main 上失败、分支上通过。
- 占位符 workaround 应下沉到公共处理器层 (design): 移除模型内私有实现，逻辑提炼为 processor.py 的 apply_text_matches_as_segmented_tokens，MiniCPM-V 只做覆写调用。
- image processor 类来源与 trust_remote_code 强制 (security): 类引用改为从 ImageProcessingMixin.get_image_processor_dict 读取 auto_map，resolve_trust_remote_code 显式校验；过程中发现上游 get_image_processor_config 丢 auto_map 的 bug，已提 transformers#47628。

风险与影响

- 风险：
 - 公共层新增函数影响面：apply_text_matches_as_segmented_tokens 是对 processor.py 的纯增量 (+33)，目前仅 MiniCPM-V 使用；但 MiniCPMVMultiModalProcessor._apply_prompt_updates 仍复制了公共实现中 matched_updates 组装与 _find_mm_placeholders 的代码，未来公共 _apply_prompt_updates 演进时需同步维护。
 - BPE 回退路径行为变化：失败回退从 " 整段文本匹配后整体重编码 " 变为 " 分段编码 "，对于同样有文本占位符的其他模型（若未来复用该工具），token 序列可能与旧行为不同，需回归验证。
 - 依赖 transformers 内部 / 半私有 API：ImageProcessingMixin.get_image_processor_dict、resolve_trust_remote_code 与 vLLM 侧 _merge_mm_kwargs、cached_get_image_processor 的 kwargs 归一化行为直接耦合；上游已确认 get_image_processor_config 存在丢 auto_map 的 bug（README 中 transformers#47628 修复中），若上游接口再变动可能回归。
 - 安全与信任边界：get_class_from_dynamic_module 会执行仓库代码，虽然已用 resolve_trust_remote_code 强制校验，但 MiniCPM-V 本身要求 trust_remote_code=True；企业环境若误配信任开关，动态加载即执行第三方代码。
 - 测试覆盖缺口：MiniCPM-V-2_6 是 gated repo 被跳过；BPE 边界测试依赖特定 tokenizer 行为，若未来 tokenizer 或 chat template 变更可能失去复现力。
- 影响：
 - 用户侧：MiniCPM-V 2.5 / 2.6 / 4.0 / 4.5 / 4.6 在 Transformers v5 下可正常 vllm serve 并处理图文 / 视频请求；同进程多 checkpoint（串行 pytest、多模型服务）不再错用图像处理器。
 - 系统侧：processor.py 公共层新增分段编码工具，为后续其他非 special-token 占位符模型的 prompt 替换提供通用解法；tests/models/registry.py 解除版本上限后，MiniCPM-V 处理测试会随 CI 在 v5 下持续运行，提前暴露兼容性回归。
 - 团队侧：需要持续跟踪上游 transformers#47628 的合入状态，并维护 vendored MiniCPMVProcessor 与上游 autoclass 解析行为的对齐。
 - 风险标记：远程代码动态加载，依赖 transformers 内部 API，BPE 回退路径改变 tokenization，gated 模型测试覆盖缺口，公共处理器层新增分支

关联脉络

- PR #50411 [Model] Fused mm preprocess normalisation on the Device: 同属多模态预处理链路演进：该 PR 改造了 vllm/transformers_utils/processor.py 与多模态处理测试，把图像归一化搬上 GPU；本 PR 继续在 processor.py 公共层新增分段编码工具并引入 _merge_mm_kwargs / cached_get_image_processor，二者共同反映 vLLM 对 HF processor 兼容层的持续改造方向，属同类基础设施演进。