PR #44232 完整报告

vllm-project/vllm

[Bugfix] Fix Gemma4 startup crash with recent transformers multimodal processor

合并时间：2026-06-02 21:42

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44232

执行摘要

- 一句话：修复 Gemma4 启动时因 transformers 升级导致的崩溃
- 推荐动作：建议批准合并，修复明确且无副作用。同时建议后续为 Gemma4MultiModalProcessor 的 _apply_hf_processor_text_only 添加单元测试，防止类似回归。

功能与动机

近期 transformers 重构了 ProcessorMixin.__call__，强制要求 multimodal placeholder 与 replacement data 1:1 匹配。Gemma4 启动时的 KV cache profiling 阶段会传入仅含 <|video|> 的 dummy prompt，但无对应 multimodal 数据，导致 HF processor 的 get_text_with_replacements 抛出 StopIteration，进而引发 ValueError 导致服务崩溃。

实现拆解

1. 覆写 _apply_hf_processor_text_only 方法（vllm/model_executor/models/gemma4_mm.py:522-539）：在 Gemma4MultiModalProcessor 类中添加 _apply_hf_processor_text_only 方法，该方法直接调用 HF processor 的 tokenizer 对传入文本进行分词，返回 token ID 列表，而不经过 HF processor 的 __call__ 方法。这样避免了 multimodal placeholder 扩展管道，token 中保留 <|video|> 等占位符的原始 token ID，供后续 _apply_prompt_updates 使用。
2. 保持其他路径不变：_call_hf_processor 和 _apply_hf_processor_mm_only 不修改，确保图像、视频、音频等正常 multimodal 处理路径不受影响。
3. 无测试文件变更：本次修改未附带新增测试，仅依赖手动启动验证（通过 vllm serve google/gemma-4-26B-A4B-it --max-model-len 16384 确认启动成功并正常服务请求）。不过本修复属于补丁性质，已有手动验证覆盖。

关键文件：

- vllm/model_executor/models/gemma4_mm.py（模块 模型执行器；类别 source；类型 data-contract；符号 _apply_hf_processor_text_only）：核心修复文件，新增 _apply_hf_processor_text_only 方法绕过 HF processor 的 multimodal 扩展管道。

关键符号：_apply_hf_processor_text_only

关键源码片段

vllm/model_executor/models/gemma4_mm.py

核心修复文件，新增 _apply_hf_processor_text_only 方法绕过 HF processor 的 multimodal 扩展管道。

```
def _apply_hf_processor_text_only(
    self,
    prompt_text: str,
    tokenization_kwargs: Mapping[str, object],
) -> list[int]:
    # 绕过 HF processor 的 __call__，直接使用 tokenizer 进行分词。
    # 原因是 HF processor 会通过 get_text_with_replacements 扩展的多模态
    # 占位符（如 <|video|>），当 prompt 中包含占位符但无对应替换数据
    # 时，会抛出 StopIteration 错误。text-only 路径只需要 token ID，
    # 因此仅用 tokenizer 就足够了。
    processor = self.info.get_hf_processor()
    text_inputs = processor.tokenizer([prompt_text], **tokenization_kwargs)
    input_ids = text_inputs["input_ids"]
    if not isinstance(input_ids, list):
        input_ids = input_ids.tolist()
    (prompt_ids,) = input_ids
    return prompt_ids
```

评论区精华

本 PR 的 review 过程简洁：唯一审核者 Isotr0py 直接批准（APPROVED），无 review 评论。Issue 评论区中，另一位贡献者 Oxygen56 提交了修复相同问题的 PR #44242，但作者 lucianommartins 指出本 PR 已实现相同修复，避免了重复工作。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 回归风险低：变更仅新增一个方法覆写，原有逻辑（_call_hf_processor、_apply_hf_processor_mm_only）未修改，且手动验证了 multimodal 路径不受影响。
  - 兼容性风险：直接调用 tokenizer 可能遗漏 HF processor 的其他预处理（如 truncation 配置），但该方法只用于 profiling 阶段的 dummy prompt，且传入了 tokenization_kwargs 保持一致性。
  - 测试覆盖不足：未包含自动化测试，未来 transformers 再次变更可能再次失效，建议后续补充单元测试。
- 影响：
  - 用户影响：Gemma4 模型用户不再因 transformers 版本升级无法启动服务，修复向后兼容 transformers 的破坏性变更。
  - 系统影响：仅影响 Gemma4 模型的多模态处理器路径，其他模型不受影响。
  - 团队影响：维护成本极低，代码简洁。

- 风险标记：缺少测试覆盖

关联脉络

- PR #44242 Fix Gemma4 startup crash with recent transformers: Oxygen56 提交了相同修复的 PR，说明多人遇到相同问题并各自提出解决方案，本 PR 先被合并。