PR #39344 完整报告

vllm-project/vllm

fix(kimi_k25): resolve media_placeholder_token_id from tokenizer

合并时间：2026-04-12 12:10

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39344

执行摘要

修复 Kimi-K2.5 多模态推理因媒体占位符 token ID 不匹配导致的完全失效问题。通过从 tokenizer 动态解析 token ID 并打补丁配置，确保配置与运行时一致，同时添加防护检查避免静默失败。影响范围限于 Kimi-K2.5 模型的多模态功能，是解决 transformers v5 升级中配置不一致问题的关键修复。

功能与动机

问题：Kimi-K2.5 多模态推理（图像 / 视频输入）完全崩溃，抛出 AssertionError: Failed to apply prompt replacement for mm_items['vision_chunk'][0]。

根本原因：模型配置中的 media_placeholder_token_id 硬编码为 163605，但运行时 tokenizer 将 <|media_pad|> 映射到 163602。这是因为 Kimi-K2.5 缺少 tokenizer.json，transformers v5 在自动转换慢速 tokenizer 时压缩了特殊 token ID 的间隙，导致偏移。

修复目标：确保多模态处理管道使用正确的 token ID，无论上游配置值是否正确。

实现拆解

修改集中在 vllm/model_executor/models/kimi_k25.py 文件：

1. KimiK25ProcessingInfo.init：
   - 新增逻辑解析 token ID：python config_token_id = hf_config.media_placeholder_token_id resolved_token_id = tokenizer.convert_tokens_to_ids("<|media_pad|>") is_valid_resolved = isinstance(resolved_token_id, int) and (tokenizer.unk_token_id is None or resolved_token_id != tokenizer.unk_token_id) if is_valid_resolved and resolved_token_id != config_token_id: logger.warning_once(...) # 记录警告 media_token_id = resolved_token_id hf_config.media_placeholder_token_id = resolved_token_id # 打补丁 else: media_token_id = config_token_id self.media_token_id = media_token_id
   - 关键点：验证解析值不是 unk_token_id，避免静默失败。
2. KimiK25MultiModalProcessor._get_prompt_updates：
   - 将 media_token_id = hf_config.media_placeholder_token_id 替换为 media_token_id = self.info.media_token_id。

- 确保使用已解析的 token ID，避免重新读取可能不正确的配置。

评论区精华

review 中只有一个核心讨论，但具有重要技术洞察：

> gemini-code-assist[bot]："There's a potential issue here if <|media_pad|> is not found in the tokenizer's vocabulary. In that case, convert_tokens_to_ids will return the unk_token_id. The current logic will then incorrectly use this unk_token_id as the media_token_id..."
>
> r266-tech回应："Good catch — added the unk_token_id guard... If <|media_pad|> isn't in the vocab, we now fall back to the config value instead of silently using the unknown token."

决策结论：添加防护检查，确保解析的 token ID 有效（是整数且不等于 unk_token_id），否则回退到配置值。这提升了鲁棒性，避免了潜在的错误传播。

风险与影响

风险：

- 配置污染：修改 hf_config.media_placeholder_token_id 可能影响其他依赖此配置的代码，但打补丁是局部的，旨在确保一致性。
- 逻辑边界：防护检查覆盖了常见情况，但如果 tokenizer.convert_tokens_to_ids 返回非整数（如列表），逻辑可能不处理，但概率较低。
- 性能影响：初始化时增加一次 tokenizer 查询，可忽略不计。

影响：

- 用户：修复了 Kimi-K2.5 多模态推理的崩溃，恢复图像 / 视频输入功能。
- 系统：确保模型在多模态场景下的正确性，避免断言失败。
- 团队：为 transformers v5 升级提供了支持，解决了配置与 tokenizer 不一致的通用问题模式。

关联脉络

- 关联 Issue #39261：直接修复此 bug，详细描述了崩溃现象和根本原因。
- 历史 PR：与近期多模态相关 bugfix（如 PR #39526、#38907）形成脉络，共同提升多模态推理的稳定性。
- 演进趋势：随着 transformers v5 升级，模型配置与 tokenizer 不一致的问题可能更常见，本 PR 提供了一种优雅的解决方案：动态解析并打补丁，同时记录警告。这为类似问题（如其他缺少 tokenizer.json 的模型）提供了参考模式。