PR #26320 完整报告

sgl-project/sglang

fix(gemma4): register image/video/audio token_regex for HF-expanded prompts

合并时间：2026-06-10 07:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26320

执行摘要

- 一句话：修复 Gemma4 多模态处理器 token_regex 缺失导致的 prompt 解析错误
- 推荐动作：该 PR 修复逻辑清晰、代码简洁，值得阅读以理解多模态 token 扩展的处理模式。其设计决策——通过正则表达式的备选分支同时支持新旧格式——是保持向后兼容的通用方法。不过如果时间有限，可直接参考 gemma3.py 中已生效的相同模式，本 PR 仅是对齐遗漏。

功能与动机

Gemma4SGLangProcessor 未设置 token_regex，导致 parse_regex 回退到 re.escape(token_str)，只能匹配单个 bare placeholder。当客户端使用 transformers 的 Gemma4 processor 生成的扩展 prompt（如 <|image> + N×<|image|> + <image|>）时，每个 patch token 被独立匹配，造成 token 数量远多于实际提供的多模态数据数量，从而抛出 RuntimeError: Mismatch: More 'IMAGE' tokens found than corresponding data provided.。gemma3.py 和 qwen_vl.py 已正确设置 image_token_regex，Gemma4 在 #21952 添加时遗漏了这一配置。

实现拆解

1. 在 __init__ 方法中为 MultimodalSpecialTokens 构造调用补充三个命名参数：image_token、video_token、audio_token 及其对应的正则表达式。
2. 每个正则表达式采用两个备选分支：第一个分支匹配完整的 HF 扩展块（BOI + 至少一个 patch token + EOI），作为一条 marker 整体识别；第二个分支匹配单个 bare placeholder，确保对旧格式客户端保持向后兼容。
3. 视频模态复用图像的 BOI/EOI 标记（<|image> / <image|>），中间 patch token 为 <|video|>，正则表达式据此构造。
4. 在文件头新增 import re。整个变更集中在 __init__ 内部，无其他文件或外部行为变动。

关键文件：

- python/sglang/srt/multimodal/processors/gemma4.py（模块 多模态；类别 source；类型 dependency-wiring；符号 Gemma4SGLangProcessor.init）：唯一的变更文件，在 Gemma4SGLangProcessor.__init__中为 image、video、audio 三种模态添加了 token 字符串和匹配扩展块的正则表达式，修复了导致 RuntimeError 的解析错误。

关键符号：Gemma4SGLangProcessor.init

关键源码片段

python/sglang/srt/multimodal/processors/gemma4.py

唯一的变更文件，在 Gemma4SGLangProcessor.__init__中为 image、video、audio 三种模态添加了 token 字符串和匹配扩展块的正则表达式，修复了导致 RuntimeError 的解析错误。

```
# File: python/sglang/srt/multimodal/processors/gemma4.py

import re
from typing import Dict, List, Optional, Union
import numpy as np
import torch

from sglang.srt.managers.multimodal_processor import (
    BaseMultimodalProcessor as SGLangBaseProcessor,
)
from sglang.srt.managers.schedule_batch import Modality, MultimodalProcessorOutput
from sglang.srt.models.gemma4_audio import _SSCP_CONV_STRIDE_SIZES
from sglang.srt.models.gemma4_mm import Gemma4ForConditionalGeneration
from sglang.srt.multimodal.processors.base_processor import MultimodalSpecialTokens
from sglang.srt.utils.video_decoder import VideoDecoderWrapper


class Gemma4SGLangProcessor(SGLangBaseProcessor):
    """Multimodal processor for Gemma4 supporting image, video, and audio inputs."""

    models = [Gemma4ForConditionalGeneration]

    def __init__(self, hf_config, server_args, _processor, *args, **kwargs):
        super().__init__(hf_config, server_args, _processor, *args, **kwargs)

        self.IM_START_TOKEN_ID = hf_config.boi_token_id
        self.IM_END_TOKEN_ID = hf_config.eoi_token_id

        self.AUDIO_START_TOKEN_ID = hf_config.boa_token_id
        self.AUDIO_END_TOKEN_ID = hf_config.eoa_token_id

        # 修复：为三种模态添加 token 字符串和匹配扩展块的正则表达式
        # 每个正则第一个备选匹配完整块 (BOI + N×patch + EOI) 作为一个 marker
        # 第二个备选匹配单个 bare placeholder，保持向后兼容
        self.mm_tokens = MultimodalSpecialTokens(
            image_token="<|image|>",
            image_token_id=hf_config.image_token_id,
            image_token_regex=re.compile(
                r"<\|image>(?:<\|image\|>)+<image\|><\|image\|>"
            ),
            video_token="<|video|>",
            video_token_id=hf_config.video_token_id,
            video_token_regex=re.compile(
```

```
            r"<\|image>(?:<\|video\|>)+<image\|>|<\|video\|>"
        ),
        audio_token="<|audio|>",
        audio_token_id=hf_config.audio_token_id,
        audio_token_regex=re.compile(
            r"<\|audio>(?:<\|audio\|>)+<audio\|>|<\|audio\|>"
        ),
    ).build(_processor)

    # Register image-processor and video-processor outputs so they are stored on
    # MultimodalDataItem via collect_mm_items_from_processor_output.
    self.ATTR_NAME_TO_MODALITY["image_position_ids"] = Modality.IMAGE
    self.ATTR_NAME_TO_MODALITY["video_position_ids"] = Modality.VIDEO
```

评论区精华

作者在 PR 评论区指出 CI 中的失败（AMD stage-b 基础设施超时、NPU 性能回归、缺少 run-ci-extra 标签）均与其变更无关。reviewer @kpham-sgl 和 @pyc96 均已批准 PR，最终由 @ch-wan 合并，确认失败测试在主分支上已存在问题。

- CI 失败是否由本 PR 引起 (other): reviewer 和合并者确认 CI 失败是已有问题，不影响合入。

风险与影响

- 风险：风险极低。变更仅扩展了 MultimodalSpecialTokens 的配置字段，不影响原有逻辑。正则表达式包含 fallback 分支，因此旧有单占位符请求仍能正确匹配。主要风险在于正则表达式本身可能存在边界遗漏（如缺少 EOI 的异常情况），但当前实现与 gemma3 等已稳定运行的处理器模式一致。缺少对应的单元测试覆盖解析过程，但变更影响面小，可通过集成测试验证。
- 影响：影响范围：仅限使用 Gemma4 模型且发送符合 HuggingFace processor 输出格式的请求的用户。修复前这些用户会遇到 RuntimeError，修复后请求能正常处理。对其他模态组合（如纯文本请求）无任何影响，对系统整体无破坏性变更。
- 风险标记：缺少测试覆盖

关联脉络

- PR #21952 Add Gemma4 multimodal processor: 本 PR 修复了 #21952 引入 Gemma4 处理器时遗漏的 token_regex 配置，使 processor 与 gemma3.py 和 qwen_vl.py 保持一致。