PR #47872 完整报告

vllm-project/vllm

[Model][HunyuanVL] Use native transformers processor and adapt to transformers 5.13

合并时间：2026-07-08 15:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47872

执行摘要

- 一句话：使用 transformers 原生 HunYuanVLProcessor，删除 vLLM 本地副本
- 推荐动作：本 PR 值得精读，尤其是 vllm/model_executor/models/hunyuan_vision.py 中对处理器调用的适配模式（包装 token、清理废弃参数），它展示了一种低成本的兼容策略。对于维护多模态模型的工程师有参考价值。

功能与动机

transformers 5.13 改变了 image-processor 的 backend 选择 API（use_fast 废弃）和 AutoImageProcessor.register 签名，导致现有注册在导入时崩溃。vLLM 只需从 processor 获取 pixel_values 和 image_grid_thw，token 扩展和 xdrope 位置由 vLLM 自身计算，因此切换到 HF 原生处理器可减少与 transformers 保持同步的维护成本。

实现拆解

1. 删除本地处理器文件：移除 vllm/transformers_utils/processors/hunyuan_vl.py（226 行）和 hunyuan_vl_image.py（477 行），它们分别定义了 HunYuanVLProcessor、HunYuanVLImageProcessor 及 smart_resize 等辅助函数。
2. 更新导入路径：在 vllm/model_executor/models/hunyuan_vision.py 中，将原本从本地 processors 模块的导入改为从 transformers 原生包导入：from transformers import HunYuanVLProcessor 和 from transformers.models.hunyuan_vl.image_processing_hunyuan_vl import HunYuanVLImageProcessor, smart_resize。
3. 适配处理器调用：
   - get_hf_processor：弃用 use_fast 参数，改为 backend='pil'，并清除 use_fast 避免传递给 HF。
   - _call_hf_processor：缓存 processor 实例；如果 prompt 中包含裸的 image_token，则自动包裹为 image_start_token + image_token + image_end_token，以满足原生处理器对占位符格式的要求。
4. 更新测试注册信息：在 tests/models/registry.py 中，将 HunYuanVLForConditionalGeneration 的 is_available_online=False 替换为 min_transformers_version='5.13'，明确此模型要求 transformers>=5.13。

关键文件：

- vllm/transformers_utils/processors/hunyuan_vl_image.py（模块 图像处理；类别 source；类型 deletion；符号 smart_resize, HunYuanVLImageProcessor, init, _preprocess）：删除 vLLM 本地实现的 HunYuanVLImageProcessor 和 smart_resize 函数，共 477 行，是精简代码的核心。
- vllm/transformers_utils/processors/hunyuan_vl.py（模块 视觉处理器；类别 source；类型 deletion；符号 HunYuanVLProcessor, init, call, batch_decode）：删除 vLLM 本地实现的 HunYuanVLProcessor，共 226 行，是精简代码的另一核心。
- vllm/model_executor/models/hunyuan_vision.py（模块 视觉模型；类别 source；类型 data-contract；符号 get_hf_processor, _call_hf_processor）：修改了处理器获取和调用逻辑，适配 transformers 5.13 的 API 变化，是实际功能适配的桥梁。
- tests/models/registry.py（模块 测试注册；类别 test；类型 test-coverage）：更新了 HunYuanVL 模型注册的 transformers 版本要求，确保 CI 在正确版本下测试。

关键符号：get_hf_processor, _call_hf_processor

关键源码片段

vllm/model_executor/models/hunyuan_vision.py

修改了处理器获取和调用逻辑，适配transformers 5.13 的API 变化，是实际功能适配的桥梁。

```
# vllm/model_executor/models/hunyuan_vision.py (head 版本关键改动 )

from transformers import BatchFeature, HunYuanVLProcessor
from transformers.models.hunyuan_vl.image_processing_hunyuan_vl import (
    HunYuanVLImageProcessor,
    smart_resize,
)

class HunYuanVLProcessingInfo(BaseProcessingInfo):
    # ...
    def get_hf_processor(
        self,
        **kwargs: object,
    ) -> HunYuanVLProcessor:
        # transformers >= 5.13 使用 `backend` 代替 `use_fast`;
        # 固定 PIL 后端以匹配已发布 HunyuanOCR 检查点的打包方式。
        kwargs.pop("use_fast", None)
        kwargs.setdefault("backend", "pil")
        return self.ctx.get_hf_processor(
            HunYuanVLProcessor,
            **kwargs,
        )

    def _call_hf_processor(
        self,
        prompt: str,
        mm_data: Mapping[str, object],
```

```
        mm_kwargs: Mapping[str, object],
        tok_kwargs: Mapping[str, object],
    ) -> BatchFeature:
        hf_processor = self.info.get_hf_processor(**mm_kwargs)
        # HunYuanVLProcessor 要求图片占位符被开始 / 结束标记包裹。
        if mm_data.get("images") is not None and prompt:
            img_tok = hf_processor.image_token
            wrapped = (
                f"{hf_processor.image_start_token}{img_tok}"
                f"{hf_processor.image_end_token}"
            )
            if img_tok in prompt and wrapped not in prompt:
                prompt = prompt.replace(img_tok, wrapped)
        return self.info.ctx.call_hf_processor(
            hf_processor,
            dict(text=prompt, **mm_data),
            dict(**mm_kwargs, **tok_kwargs),
        )
```

评论区精华

No substantial review discussion occurred. The PR was directly approved by Isotr0py without comments, indicating the changes were straightforward and correctly handled the adapter logic. The only automated comments were from welcome bot and mergify regarding pre-commit checks.

- 审核批准 (other): 无需修改，直接合并。

风险与影响

- 风险：
  1.transformers版本依赖：模型现在要求transformers>=5.13，低于此版本将无法加载。测试注册中已明确标注 min_transformers_version。
  2. 占位符包装逻辑：_call_hf_processor 中的 token 包裹假设 prompt 格式总是裸 image_token，若未来 transformers 原生 processor 对输入格式有进一步要求可能需再次适配。
  3. 删除文件影响：本地处理器文件仅被 hunyuan_vision.py 引用，直接删除无副作用。其他模型未依赖这些文件。
  4. 测试覆盖：仅有的测试变更是 registry 中的版本标注，没有新增集成测试验证处理器返回的字段是否正确。手工 E2E 测试已通过。 - 影响：影响范围：仅影响 HunYuanVLForConditionalGeneration（如 tencent/HunyuanOCR）模型用户。使用 transformers 5.13+ 的用户将获得原生处理器支持，且代码量减少约 700 行。影响程度：中等——模型加载路径产生变化，但保持输入输出兼容。其他模型和系统组件不受影响。团队：减少了 vLLM 需要同步的 transformers 内部代码，后续升级 transformers 时对 HunYuanVL 部分的维护负担降低。
- 风险标记：transformers 版本依赖 ，占位符格式假设 ，无新增集成测试

关联脉络

- 暂无明显关联 PR