PR #40335 完整报告

vllm-project/vllm

[MM][Misc] Support image+video mixed inputs (per prompt) for VLM examples

合并时间：2026-04-21 11:43

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40335

执行摘要

- 一句话：为 VLM 离线推理示例添加图像 + 视频混合输入支持。
- 推荐动作：该 PR 对于想了解 vLLM 多模态集成或测试混合输入场景的开发者值得浏览，重点关注占位符拼接和 limit_mm_per_prompt 的动态配置设计，但核心逻辑较为直白，无需深度分析。

功能与动机

根据 PR body，VLM 离线推理示例之前只接受单一模态（image 或 video），用户无法测试同时包含图像和视频的提示。此变更旨在提供一种方式，使社区能通过示例脚本测试 vision-language 模型在混合模态输入上的表现。

实现拆解

1. 参数解析扩展：在示例脚本的 argparse 中添加 --modality "image+video" 选项，支持新模态。
2. 多模态限制配置：在每个模型运行函数中，将 limit_mm_per_prompt 从硬编码 {modality: 1} 改为动态计算：mm_limit = {"image": 1, "video": 1} if modality == "image+video" else {modality: 1}，确保混合模态下允许一个图像和一个视频。
3. 占位符拼接逻辑：为每个模型定义图像和视频占位符变量，并在 image+video 模态下拼接两者；同时修复 Qwen 系列等模型的占位符包装错误，将视觉边界标记移入占位符变量以保持一致性。
4. 特定模型调整：例如在 run_hyperclovax_seed_vision 中增加消息内容列表分支以插入两种内容块，并调整 max_model_len；在 run_minicpmv_base 中用显式 content_prefix 变量替换字典查找。
5. 测试配套：PR body 提供了测试命令和结果，但未包含测试文件变更，主要依赖现有测试或手动验证。

关键文件：

- examples/offline_inference/vision_language.py（模块 示例脚本；类别 source；类型 core-logic；符号 _mm_data, _mm_uuid, _mm_empty）：这是 PR 的唯一变更文件，包含了所有模型运行函数的更新，以支持图像 + 视频混合输入模态。

关键符号：run_ernie45_vl, run_exaone4_5, run_glm4_1v, run_hyperclovax_seed_vision, run_minicpmv_base, run_llava_onevision, run_qwen3_vl

关键源码片段

examples/offline_inference/vision_language.py

这是 PR 的唯一变更文件，包含了所有模型运行函数的更新，以支持图像 + 视频混合输入模态。

```
def run_ernie45_vl(questions: list[str], modality: str) -> ModelRequestData:
    model_name = "baidu/ERNIE-4.5-VL-28B-A3B-PT"

    # 动态计算多模态限制：如果模态为 image+
    video，则允许一个图像和一个视频；否则保持单一模态限制
    mm_limit = {"image": 1, "video": 1} if modality == "image+video" else {modality: 1}
    engine_args = EngineArgs(
        model=model_name,
        max_model_len=4096,
        max_num_seqs=5,
        limit_mm_per_prompt=mm_limit, # 使用动态计算的多模态限制
        trust_remote_code=True,
    )

    # 定义图像和视频的占位符字符串，便于后续拼接
    image_placeholder = "Picture 1:<|IMAGE_START|><|image@placeholder|><|IMAGE_END|>"
    video_placeholder = "Video 1:<|VIDEO_START|><|video@placeholder|><|VIDEO_END|>"

    # 根据模态选择或拼接占位符：image+video 模态下连接两者
    if modality == "image":
        placeholder = image_placeholder
    elif modality == "video":
        placeholder = video_placeholder
    elif modality == "image+video":
        placeholder = image_placeholder + video_placeholder # 拼接图像和视频占位符

    prompts = [
        (
            f"<|begin_of_sentence|>User: {question}{placeholder}\n"
            "Assistant: <think></think>"
        )
        for question in questions
    ]

    return ModelRequestData(
        engine_args=engine_args,
        prompts=prompts,
    )
```

评论区精华

- 硬编码模型路径问题：gemini-code-assist[bot] 指出 Qwen3-VL 模型名硬编码为本地路径，建议使用公共 Hugging Face ID 以确保社区可运行性。作者在后续提交中可能已调整，但评论中未显示明确解决。
- 占位符连接优化：DarkLight1337 和 Isotr0py 建议通过连接单个模态占位符来减少代码重复，作者回应并更新了实现，采用 image_placeholder + video_placeholder 的方式拼接。
  - 硬编码模型路径 (correctness): 建议使用 "Qwen/Qwen3-VL-8B-Instruct" 替代本地路径。
  - 占位符连接优化 (design): 采用 image_placeholder + video_placeholder 的方式拼接占位符，简化逻辑。

风险与影响

- 风险：
  - 配置兼容性风险：如果模型后端不支持混合模态输入，可能导致运行时错误；但示例脚本主要面向测试，风险较低。
  - 占位符拼接错误：在修复占位符包装时，若视觉边界标记处理不当，可能影响模型解析，但 PR 已统一调整。
  - 缺少自动化测试：变更覆盖多个模型函数，但未添加对应单元测试，依赖手动验证，可能存在遗漏。
- 影响：
  - 用户影响：使开发者能更方便地测试多模态模型的混合输入能力，提升示例脚本的实用性和灵活性。
  - 系统影响：仅影响示例脚本，不涉及核心推理引擎或生产代码，对系统性能无直接影响。
  - 团队影响：为多模态功能提供更完整的示例，有助于社区学习和集成。
  - 风险标记：配置兼容性风险，缺少测试覆盖

关联脉络

- PR #40355 [Doc] Update ViT CUDA graph doc for mixed (image+video) inputs: 两者都涉及多模态混合输入（图像 + 视频）的支持，此 PR 是文档更新，当前 PR 是示例实现。
- PR #40411 [Bugfix] Gemma4: fix multimodal embedder norm order to match HF reference: 都涉及多模态模型的修复和增强，属于同一功能线。