ChatGPT

开源视频语义理解模型

- **视频-LLM 模型 (Vision-Language LLM)**：典型代表如 Video-ChatGPT、Video-LLaVA 等。这类模型通常将视觉编码器（如 CLIP）与预训练语言模型（如 Vicuna）结合，对视频帧进行编码，再由 LLM 输出文本回复 [1] [2]。
- 输入要求：输入通常为预抽取的若干视频帧（例如 Video-LLaVA 训练时固定使用 8 帧 [3]，Video-ChatGPT 可接收任意数量关键帧）。视频需先切帧、调整分辨率后作为图片输入。
- 输出能力：以自然语言回答问题、描述事件或对话形式的故事。能基于视频内容进行问答和情节叙述。输出是文本（可做结构化信息抽取或直接生成连贯描述）。
- 使用方式：开放源码，可通过 HuggingFace Transformers 或社区代码库使用。需先将视频拆分成帧，通过视觉编码器提取特征，再以 Prompt 形式输入 LLM 生成结果 [3] [1]。
- 多模态支持：支持图像与文本的联合输入（有些模型可混合图像+视频输入 [4]），一般不处理音频。可扩展到场景理解、物体识别、OCR+视频等，但需额外模块。

- 场景与局限：适用于视频问答、事件描述、对话交互等任务。优点是生成灵活、可扩展；局限在于只能理解帧间显式视觉信息，不能直接处理连续长视频，帧数受限（Video-LLaVA 限 8 帧 [3]），需要显式帧抽取，实时性受限于编码和推理速度。

- **视频事件描述模型 (Vid2Seq)**：由 Google 提出的 Vid2Seq 是一种密集事件字幕生成模型 [5] [6]。它在视频帧序列上插入特殊时间标记，输出同时包含事件边界和事件文字描述。

- 输入要求：输入为一段连续视频，通常通过视频帧或对帧序列进行编码，训练时使用有转录字幕的有声视频，但推理可仅用视频。视频需切帧并传入视觉编码器。
- 输出能力：输出是一系列事件边界标记及对应的文本描述（“密集字幕”）。可生成时序化的事件列表，例如“[00:10] 人物开始跑步；[00:30] 跳下台阶，等等”。可看作结构化的故事摘要。
- 使用方式：研究原文和开源代码提供训练与推理流程 [5] [6]。Vid2Seq 可在大型数据集（如 YouTube-Temporal-1B）上预训练后，在特定任务（YouCook2、ActivityNet）上微调。使用时需加载预训练模型，输入视频帧序列，模型自动分割并输出事件-字幕序列。
- 多模态支持：主要基于视觉信息，也可利用视频的语音转文字作为辅助（训练时用转录语音标定事件边界）。原生不包含音频理解。

- 场景与局限：适合电影剪辑、运动赛事、动作事件等需要自动标注开始/结束时间和描述的场景。优势是能一并输出时间和描述；局限是生成固定格式的字幕，难以交互问答，且效果依赖于视频中隐含的语音信息（训练用了配套字幕），对完全无声的视频泛化能力可能受限。

- **视觉视频基础模型 (InternVideo 系列)**：InternVideo 是中科院等团队提出的视频基础预训练模型系列，通过自监督的生成式（如遮挡视频重建）和判别式（如视频-文本对比）预训练，学习通用视频表征 [7]。

- 输入要求：输入为整段视频的帧序列或帧分段，可随机遮挡后输入模型进行重建预训练；下游可直接输入帧或提取特征。视频可直接输入模型，无需特定预处理（除了标准的帧归一化）。通常对分辨率没有特别限制，但模型训练时可能使用 224×224 等常见大小。
- 输出能力：InternVideo 本身主要输出视觉特征，可应用于多个任务：动作识别、动作检测、视频-文本匹配、视频问答等。它不是直接生成故事的模型，但对各种视频理解任务提供强表征 [7]。例如，在 Kinetics-400 动作识别上可达 91.1% 精度 [7]。可以结合下游头或 LLM 实现问答。

- 使用方式：开放源码（GitHub）和模型权重可获取，可作为预训练编码器使用。开发者可加载InternVideo，提取视频特征，再通过简单分类层或与语言模型联合（Adapter 微调、图-查询接口）实现特定应用。无需从零训练，可fine-tune。
- 多模态支持：InternVideo 主要专注视觉，基础模型为视频编码器，可用于视频+文本对齐任务（多模态），但本身不内建OCR或语音功能。可与OCR模块、语音识别结合实现多模态理解。

- 场景与局限：适用于需要高质量视觉表征的场景，如视频分类、检索、动作检测等。优势是性能强、通用性好；局限是对“视频讲故事”这种生成式任务需要额外模型支持，且预训练资源消耗大，不直接输出自然语言描述。

- **轻量视频LLM（Mobile-VideoGPT)**：Mobile-VideoGPT 由阿布扎比AI大学等提出，是一种高效的视频理解框架 [8] 。它采用轻量图像编码器和轻量视频编码器抽取视觉信息，通过有效投影器输入一个小型语言模型。

- 输入要求：输入为拆帧后的视频图像序列（论文中使用均匀采样后选取 top-K 关键帧）。通常先用图像编码器对所有帧提取特征，再通过注意力选取重要帧，最后送入视频编码器。论文示例中帧率、帧数固定（如16帧、224×224 分辨率） [9] 。
- 输出能力：输出为回答问题或生成文本描述。通过小模型（0.5B~1.6B 参数）对视频进行提问回答。可生成简要回答或描述，适用于视频问答场景。论文中在多个基准测试（如 MVBench、EgoSchema、NextQA 等）上进行了 QA 评价 [10] [8] 。
- 使用方式：代码开源，可安装 Python 包使用 [8] 。使用时需加载预训练模型，输入视频帧或问题指令即可生成文本。设计为资源受限设备友好，可做实时推理（0.5B模型可达每秒46个token） [11] 。
- 多模态支持：仅使用视觉信息（图像+视频编码器），不含音频或OCR处理。可作为视频+文本联合理解的端到端框架。
- 场景与局限：适合对时效要求高的应用，如边缘设备的视频分析、实时监控问答等。小模型推理快、效率高；但模型能力有限，对复杂长视频或极细节推理不如大型模型。效果受关键帧提取影响，需要合适的帧采样策略。

商业API与平台

- **OpenAI GPT-4 Vision (GPT-4V)**：GPT-4V 是 OpenAI 的视觉感知版本，能够理解图像 [12] 。
- 输入要求：只能接受静态图片输入（不直接支持视频）。如需分析视频，需要先将视频拆分为多帧图片后分次输入 [12] 。对图片分辨率和大小有限制（建议短边 ⩾ 640px）。
- 输出能力：输出为自然语言描述、回答、对话等。例如可描述一张图片中的场景、回答相关问题等。可基于连续帧输出多个回答，但需自行合并总结。
- 使用方式：通过 OpenAI 提供的 API（或 ChatGPT 界面）使用视觉能力。开发者需按帧调用 API 并管理上下文。
- 多模态支持：支持图像+文本交互。不支持原生视频帧时序信息，也不解析音频或字幕。

- 场景与局限：适用于图片内容理解、图片生成辅助理解场景。可用于简单的事件描述（如电影剧照解读）。局限是不能直接处理视频，视频场景理解需复杂拼接；处理速度受API调用延时影响；对连贯故事理解需要额外手工整合。

- **Google Gemini (视频理解)**：Google Gemini 在 Vertex AI 平台上支持视频输入，可直接进行视频理解任务 [12] [13] 。

- 输入要求：支持多种视频格式（MP4、AVI、WebM 等），输入时无需预先抽帧。当前支持最长 45 分钟（带音频）到 1 小时（无音频）视频长度 [13] 。分辨率不限于 768×768 等整倍数（如 Gemini 1.5 录入 768×768，表中示例）；帧率可自动处理。
- 输出能力：输出为文本，可以是对视频内容的总结、关键事件描述、问答回答等。用户可通过自然语言提示（“请描述视频中的主要事件”）与 Gemini 交互。模型能考虑视频的时序和场景变化生成回答。

- 使用方式：通过 Google Cloud Vertex AI 的 Generative AI API 使用。可在 Google Cloud 控制台或 API 端点上传视频并获取响应。也可按需求自定义任务提示。
- 多模态支持：支持视频内的视觉和音频信息（如果有）。原生具备音视频合并理解能力（音频可用来辅助语言理解）。同时输入可混合图片、文本和视频，实现复杂查询。
- 场景与局限：适用于视频内容分析、摘要、问答、监控等场景。模型容量大、训练广泛，理解能力强。局限在于依赖云服务，调用成本和延时较高；对隐私有要求的视频需慎用；对极长视频仍有时长限制。

- **Runway AI**：Runway 提供视频生成和编辑的 AI 模型，目前主要面向创作领域。

- 输入要求：Runway 主要模型以文本、图像或视频作为提示生成视频（如 Gen-2、Gen-3）。不专门针对视频理解，输入通常为文本描述或图像种子，或者对已有视频做编辑输入。
- 输出能力：输出为新生成或编辑后的视频（动态图像序列）。不输出语义文本描述。模型擅长视觉特效、场景转换、图像到视频的生成等。
- 使用方式：提供 Web UI 和 API，可通过简单请求构造视频。开发者主要用其创作功能，不用自己训练模型。
- 多模态支持：支持文本、图像、视频作为生成条件。但不输出文本回答。
- 场景与局限：适用于内容创作者的视频生成/编辑场景。并不用于视频事件理解或语义提取。若需要语义分析需额外处理生成的视频和相关提示。

- **Synthesia**：Synthesia 专注于生成虚拟演讲者视频。

- 输入要求：输入为文字脚本和选定的虚拟角色。视频源主要是静态头像，不处理输入视频。
- 输出能力：输出为由虚拟演员朗读输入文本的短视频，支持多语言配音。无视频内容理解功能。
- 使用方式：通过 Web 平台上传文本脚本并选择角色，自动生成视频。
- 多模态支持：支持文本→视频（含合成语音）生成。
- 场景与局限：主要用于营销、教育、内部培训等领域的内容生成。完全不支持输入视频分析，无助于场景理解。

- **Pika (Pika Labs)**：Pika 提供图像/文本到视频的生成工具。

- 输入要求：输入为文本描述或参考图像。Pika 不做视频分析。
- 输出能力：生成短片段动画视频。
- 使用方式：主要提供网页版，用户输入描述后生成视频。
- 多模态支持：以文本和静态图像为输入，输出为视频。
- 场景与局限：用于创意内容制作，与视频理解任务无关。

- **Sora (OpenAI)**：Sora 是 OpenAI 用于视频创作的模型。

- 输入要求：接受文本提示或示例视频，生成新视频片段。
- 输出能力：生成逼真视频，时长可达数秒到几十秒。
- 使用方式：目前仅作为研究和产品功能演示，未来可能集成 API。
- 多模态支持：可混合文本、图像和视频提示生成新视频。
- 场景与局限：面向视频生成创作，不提供视频内容理解。

模型和服务对比

方案	输入要求	输出形式	使用方式	多模态支持	应用场景/局限
Video-ChatGPT	视频帧（切帧后图片序列）	文本（对话式问答/描述）	开源代码+API（Python）	图像+文本	视频问答、情节描述；需要显式切帧、无法处理超长视频
Video-LLaVA	固定帧数图像序列（训练时8帧）	文本（指令式答复/描述）	开源模型（HuggingFace）	图像+文本	图像视频混合理解；帧数受限（8帧），适合短视频场景
Vid2Seq	视频（及可用音频转文本的字幕）	文本（带时间戳的事件描述序列）	开源代码（Google）	视频（视觉）+文本标注	密集事件字幕、动作分段；专注于生成时间+描述列表
InternVideo	视频帧	视觉特征（用于分类、匹配等）	开源模型+微调	视频视觉（可用于多模态任务）	视频分类/检索/动作识别等；不直接生成故事，需要配合下游模型
Mobile-VideoGPT	视频帧（关键帧采样，例：16帧）	文本（问答/概要回复）	开源模型+API/PyTorch	视频视觉	视频问答/摘要；小模型实时；推理快但能力有限，适合边缘设备
GPT-4V (OpenAI)	静态图片（需分帧后逐帧调用）	文本（描述、问答）	API/ChatGPT 界面	图像+文本	图像内容理解；无视频输入接口（需自行拆帧），对视频串讲有限
Google Gemini	视频（多种格式，最长1小时）	文本（总结、问答）	云API (Vertex AI)	视频视觉+音频+文本	视频内容分析；支持音视频；适用于长视频摘要和问答；需云服务
Runway	文本/图像/视频提示（生成条件）	视频（生成或编辑后的画面）	Web/API	文本+图像+视频生成	视频生成/编辑工具；不输出文本；不用于理解任务
Synthesia	文本脚本（及选定虚拟演员角色）	视频（AI虚拟人物朗读脚本）	Web 平台	文本→视频+语音生成	生成虚拟主持人视频；不处理原视频输入，仅文本到视频
Pika	文本/图像（生成提示）	视频（短动画片段）	Web/App	文本/图像→视频生成	创意短视频生成；不具备视频语义分析功能

方案	输入要求	输出形式	使用方式	多模态支持	应用场景/局限
Sora	文本/图像/视频示例	视频（生成新片段）	研究性质（非公开）	文本+图像+视频生成	视频生成研发；不提供语义分析

说明：上述开源模型多需先将视频拆分为帧再处理，输出通常为自然语言描述或事件列表。商业平台（GPT-4V、Gemini）则可通过云端服务对视频内容进行推理，但 GPT-4V 仅限图像输入[12]；Google Gemini 支持长视频[13]。其他商业工具（Runway、Synthesia、Pika、Sora）主要用于视频生成/编辑，不用于视频理解。所有方案均不包含直接的语音理解，OCR 等需要额外集成。以上模型各有侧重，在视频摘要、问答、事件识别等不同场景下存在性能和资源上的权衡。

---

[1] Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models
https://arxiv.org/html/2306.05424v2

[2] GitHub - mbzuai-oryx/Video-ChatGPT: [ACL 2024] Video-ChatGPT is a video conversation model capable of generating meaningful conversation about videos. It combines the capabilities of LLMs with a pretrained visual encoder adapted for spatiotemporal video representation. We also introduce a rigorous 'Quantitative Evaluation Benchmarking' for video-based conversational models.
https://github.com/mbzuai-oryx/Video-ChatGPT

[3] [4] Video-LLaVA
https://huggingface.co/docs/transformers/en/model_doc/video_llava

[5] [2302.14115] Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning
https://arxiv.org/abs/2302.14115

[6] Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning
https://antoyang.github.io/vid2seq.html

[7] [2212.03191] InternVideo: General Video Foundation Models via Generative and Discriminative Learning
https://arxiv.org/abs/2212.03191

[8] [11] GitHub - Amshaker/Mobile-VideoGPT: Mobile-VideoGPT: Fast and Accurate Video Understanding Language Model
https://github.com/Amshaker/Mobile-VideoGPT

[9] [10] Mobile-VideoGPT: Fast and Accurate Video Understanding Language Model
https://arxiv.org/html/2503.21782v1

[12] 如何用GPT-4o解读视频_可以分析视频的gpt-CSDN博客
https://blog.csdn.net/xindoo/article/details/143837432

[13] Video understanding | Generative AI on Vertex AI | Google Cloud
https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/video-understanding