



Video-LLaMA

An Instruction-tuned Audio-Visual Language Model for Video Understanding

Hang Zhang^{1 2} Xin Li^{1 2*} Lidong Bing^{1 2}

¹ DAMO Academy, Alibaba Group

² Hupan Lab, 310023, Hangzhou, China

{zh401075, xinting.lx, l.bing}@alibaba-inc.com

Abstract

We present Video-LLaMA¹ a multi-modal framework that empowers Large Language Models (LLMs) with the capability of understanding both visual and auditory content in the video. Video-LLaMA bootstraps cross-modal training from the frozen pre-trained visual & audio encoders and the frozen LLMs. Unlike previous works that complement LLMs to process the visual or audio signals only (Zhu et al., 2023; Liu et al., 2023; Huang et al., 2023a), Video-LLaMA enables video comprehension by tackling two challenges: (1) capturing the temporal changes in visual scenes, (2) integrating audio-visual signals. To counter the first challenge, we propose a Video Q-former to assemble a pre-trained image encoder into our video encoder and introduce a video-to-text generation task to learn video-language correspondence. For the second challenge, we leverage ImageBind (Girdhar et al., 2023), a universal embedding model aligning multiple modalities, as the pre-trained audio encoder and introduce an Audio Q-former on top of ImageBind to learn reasonable auditory query embeddings for the LLM module. To align the output of both visual & audio encoders with LLM’s embedding space, we first train Video-LLaMA on massive video/image-caption pairs and then tune our model with visual-instruction datasets of moderate amount but higher quality. We found Video-LLaMA shows the ability to perceive and comprehend video content and generate meaningful responses grounded in the visual and auditory information presented in the videos.

1 Introduction

Large Language Models (LLMs) (Chowdhery et al., 2022; Bai et al., 2022; OpenAI, 2023) have demonstrated remarkable capability of understanding and

following user intentions and instructions²³⁴. Typically, the user requests and the corresponding responses from LLMs are all in texts, however, text-only human-computer interaction is not sufficient for many application scenarios because real-world information is usually multi-modal. In order to further explore the potential of LLMs, many researchers attempt to endow LLMs with the capability of understanding multi-modal content (Huang et al., 2023a; Zhang et al., 2023b; Yin et al., 2023).

Among these efforts, Alayrac et al. (2022b); Wang et al. (2022); Huang et al. (2023b); Xu et al. (2023b); Zhang et al. (2023b); Sun et al. (2023) pre-train multi-modal LLMs with massive interleaved image-text data or speech-text data to accommodate multi-modal input. Meanwhile, another group of works adopts a more parameter-efficient way by complementing LLMs with off-the-shelf vision or speech foundation models to achieve multi-modal understanding (Li et al., 2023b; Zhu et al., 2023; Liu et al., 2023; Ye et al., 2023; Zhang et al., 2023a; Huang et al., 2023a; Wu et al., 2023b; Su et al., 2023; Li et al., 2023a).

Despite their effectiveness, these approaches are dedicated to aligning the input from exactly one additional modality with text (i.e., image or audio), which is unsatisfactory for video understanding. Concretely, empowering LLMs to understand video requires comprehensive processing for different modalities including visual input, auditory input, and textual output, which is more challenging than image-only understanding and audio-only understanding tasks. Although there are several recent works attempt to unleash the video understanding capability of LLMs (Li et al., 2023c; Maaz et al., 2023; Luo et al., 2023), their primary objective is to comprehend only the visual content of the video, with the auditory content remaining unused.

*Xin Li is the corresponding author.

¹The video demonstration is available at <https://youtu.be/RDNYs3Rswhc>

²<https://chat.openai.com/chat>

³<https://www.anthropic.com/product>

⁴<https://bard.google.com/>

| Model Name | Ability | | |
|-----------------------------------|--------------|--------------|-------|
| | Static Image | Silent Video | Audio |
| BLIP2 (Li et al., 2023b) | ✓ | | |
| MiniGPT4 (Zhu et al., 2023) | ✓ | | |
| LLaVA (Liu et al., 2023) | ✓ | | |
| mPLUG-Owl (Ye et al., 2023) | ✓ | ✓ | |
| VideoChat (Li et al., 2023c) | ✓ | ✓ | |
| AudioGPT (Huang et al., 2023a) | | | ✓ |
| Video-ChatGPT (Maaz et al., 2023) | ✓ | ✓ | |
| Video-LLaMA | ✓ | ✓ | ✓ |

Table 1: Comparison with popular multi-modal large language models. Video-LLaMA has the unique ability to comprehend auditory and visual information simultaneously.

In this work, to fill in the blank of audio-visual LLMs, we investigate the possibility of building multi-modal LLMs that support the input of video and allow users to chat with computers around the user-uploaded video, which is usually composed of multiple video frames and audio. Instead of employing external perception models to convert visual/auditory signals to textual signals (Shen et al., 2023; Li et al., 2023c), we choose to build an end-to-end model that can handle the data from multiple modalities within one single framework. Specifically, we adopt the idea of BLIP-2 (Li et al., 2023b) to guarantee the efficiency of cross-modal pre-training. To explicitly capture the change of visual scenes in the video, we use a pre-trained visual encoder to separately compute frame representations. Then, we introduce a frame embedding layer to inject temporal information and a video Q-Former to generate visual query tokens. As for the audio signals from the video, we additionally leverage a pre-trained audio encoder as well as an audio Q-former to learn reasonable auditory query embeddings (see the right part of Figure 1).

To align textual output with video, we devise multi-branch cross-modal pre-training to learn the vision-language correspondence and the audio-language correspondence. For vision-language correspondence, we first pre-train the vision-related components on a large-scale video caption dataset with a video-clips-to-text generation task. To enhance the understanding of static visual concepts, we also add image-caption data into this pre-training stage. Then, we further fine-tune these components on a video-based conversation dataset to execute visual instruction tuning. For the alignment between the audio encoder and language decoder, we further pre-train the audio-related components on an audio caption dataset with an audio-

to-text generation task. For the audio-language correspondence, we leverage Imagebind (Girdhar et al., 2023) as an encoder, which performs exceptionally well in aligning different modalities to a common embedding space. Given the limited availability of audio-text data, we also utilize vision-text data to train the audio-related components. These components learn to align the common embedding space provided by Imagebind with the embedding space of LLMs. Despite not being explicitly trained with audio-text data, Video-LLaMA exhibits a remarkable zero-shot audio understanding capability during inference.

As shown in Table 1, our Video-LLaMA stands out from other existing multi-modal LLMs in terms of its distinctively comprehensive comprehension of audiovisual modal information in videos. In summary, our contributions are as follows:

- We propose Video-LLaMA, a multi-modal framework that enables LLM to simultaneously process both the visual and auditory content of a given video and engage in conversation with humans.
- To empower LLMs with video understanding capability, we propose a multi-branch cross-modal pre-training framework to achieve both vision-language alignment and audio-language alignment.
- We open-source the entire codebase for pre-training and fine-tuning as well as the model weights of all the variants of Video-LLaMA⁵. We also prepared the demos for video-grounded conversation^{6,7}.

2 Method

Video-LLaMA aims to empower frozen LLMs with the capability of understanding both visual and auditory content in videos. As shown in Figure 1, we design two branches, namely Vision-Language Branch and Audio-Language Branch, to respectively transform the video frames and audio signals into query representations that are compatible with the textual inputs of LLMs. In this section, we first introduce the overall architecture and the building blocks of each branch. Then, we delineate the procedures of the proposed multi-branch cross-modal pre-training and audio-visual instruction tuning.

⁵<https://github.com/DAMO-NLP-SG/Video-LLaMA>

⁶<https://huggingface.co/spaces/DAMO-NLP-SG/Video-LLaMA>

⁷<https://modelscope.cn/studios/damo/video-llama/summary>

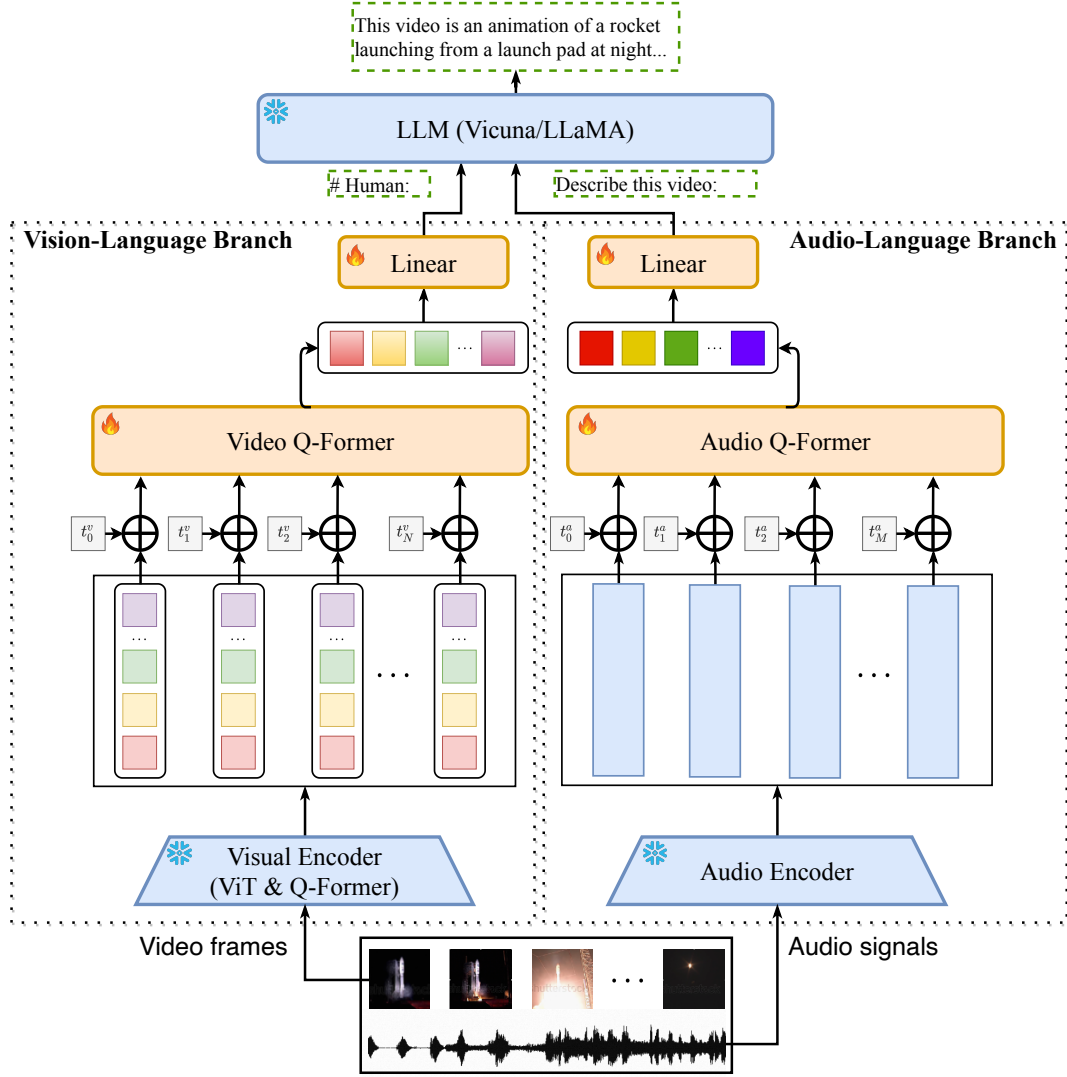


Figure 1: Overall architecture of Video-LLaMA.

2.1 Architecture

2.1.1 Vision-Language Branch

The Vision-Language Branch is designed for enabling the LLMs to understand visual inputs. As shown in the left part of Figure 1, it is composed of a frozen pre-trained image encoder to extract features from video frames, a position embedding layer to inject temporal information into video frames, a video Q-former to aggregate frame-level representations and a linear layer to project the output video representations into the same dimension as the text embeddings of LLMs. Given one video consists of N frames, the image encoder will first map each frame/image into K_f image embedding vectors, yielding video frame representations $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N]$ where $\mathbf{v}_i \in \mathbb{R}^{K_f \times d_f}$ is the set of d_f -dimensional image embeddings corresponding to the i -th frame.

Since the frame representations \mathbf{v}_i from the frozen image encoder are computed without considering any temporal information, we further apply position embeddings as the indicator of temporal information to the representations from different frames. Then, we feed the position-encoded frame representations to Video Q-former, which shares the same architecture with Query Transformer (Q-Former) in BLIP-2 (Li et al., 2023b), to obtain k_V video embedding vectors of dimension d_v as the representation $\hat{\mathbf{v}} \in \mathbb{R}^{k_V \times d_v}$ of the video.

To adapt the video representations to the input of LLMs, we add a linear layer to transform the video embedding vectors into the video query vectors. The video query vectors are of the same dimension as the text embeddings of LLMs. In the forward pass, they will be concatenated to text embeddings as a *video soft prompt* and guide the frozen LLMs

to generate text conditioned on video content.

As for the implementation of the Vision-Language Branch, we utilize the pre-trained vision component of BLIP-2 (Li et al., 2023b) as the frozen visual encoder, which includes a ViT-G/14 from EVA-CLIP (Fang et al., 2022) and a pre-trained Q-former. The remaining components, including the position embedding layer, Video Q-former, and Linear layer are randomly initialized and optimized to well connect the output of the frozen visual encoder to frozen LLMs.

2.1.2 Audio-Language Branch

To deal with the auditory content of the given video, we introduce the Audio-Language Branch. Concretely, it consists of a pre-trained audio encoder to compute features given a short segment of origin audio, a position embedding layer to inject temporal information to audio segments, an audio Q-former to fuse the features of different audio segments, and a linear layer to map the audio representation into the embedding space of LLMs.

In practice, we utilize the pre-trained Imagebind (Girdhar et al., 2023) as the audio encoder. We first uniformly sample M segments of 2-second short audio clips from the video, then convert each 2-second audio clip into spectrograms using 128 mel-spectrogram bins. After obtaining the spectrogram list of input audio, the audio encoder will map each spectrogram into a dense vector. So the generated audio representation of the given video can be denoted as $A = [a_1, a_2, \dots, a_M]$.

Similar to Video Q-Former, the Audio Q-former injects temporal information by adding learnable positional embeddings to audio segments. It then generates fixed-length audio features by computing the interaction across the position-encoded audio segments. Audio Q-Former adopts the same architecture as Q-Former. It projects the variable-length audio representation list A into a fixed-length sequence $\hat{\mathbf{A}} \in \mathbb{R}^{K_a \times d_a}$, where the K_a is the number of audio embedding vectors and d_a is the dimension of each vector. Finally, we employ a linear layer to map audio features to the embedding space of the LLM.

2.2 Multi-branch Cross-Modal Training

We train the vision-language and audio-language branches separately. In the first stage, large-scale vision-caption datasets are used for training, and in the second stage, high-quality instruction-following datasets were used for fine-tuning. The

image is treated as a one-frame video.

2.2.1 Training of Vision-Language Branch

For pre-training vision-language branch, we utilized Webvid-2M (Bain et al., 2021), a large-scale dataset of short videos with textual descriptions sourced from stock footage sites. Moreover, we employed the image caption dataset CC595k, which is sourced from CC3M (Sharma et al., 2018) and filtered by Liu et al. (2023). We adopt a video-to-text generation task during the pre-training stage, i.e., given the representation of a video, prompting the frozen LLM to generate the corresponding text description. We find that a significant portion of textual descriptions are insufficient to reflect the entire content of the videos. Therefore, the visual semantics in the videos are not fully aligned with the textual semantics in the video descriptions. Nevertheless, this stage aimed to utilize a vast amount of data and enable video features to contain as much visual knowledge as possible. We left the abilities of vision-text alignment and instruction-following for the next stage.

After the pre-training stage, the model can generate content about information in the video, but its ability to follow instructions has decreased. Therefore, in the second stage, we fine-tuned the model using high-quality instruction data. We integrated the image-detail-description dataset from MiniGPT-4 (Zhu et al., 2023), the image-instruction dataset from LLaVA (Liu et al., 2023), and the video-instruction dataset from Video-Chat (Li et al., 2023c). After fine-tuning, Video-LLaMA exhibited remarkable abilities in following instructions and comprehending images and videos.

2.2.2 Training of Audio-Language Branch

Training the audio-language branch directly using audio-text data is highly challenging due to the rarity of such data. The objective of the learnable parameters in the audio-language branch is to align the output embedding of the frozen audio encoder with the embedding space of LLM. Given the scarcity of audio-text data, we employ a workaround strategy to achieve this objective. ImageBind, which is used as our audio encoder, has a remarkable ability to align different modalities' embeddings to one common space, demonstrating impressive performance on cross-modal retrieval and generation tasks. In light of the scarcity of audio-text data and the abundance of visual-text data, we train the audio-language branch using visual-text

data, following the same data and process as the vision branch. Thanks to the shared embedding space provided by ImageBind, Video-LLaMA exhibits the ability to comprehend audio during inference, even though the audio interface has never been trained on audio data.

3 Related Works

Large Language Models: Large language models (LLMs) (Black et al., 2022; Scao et al., 2022; OpenAI, 2023; Tsimpoukelli et al., 2021) have demonstrated remarkable language understanding and reasoning abilities, enabling the generation of high-quality natural language text across various domains, including articles, conversations, stories, and poetry. LLMs have already sparked a technological revolution and have been widely applied in different applications. Moreover, a series of open source large models, such as LLaMA (Touvron et al., 2023), BLOOM (Scao et al., 2022) and OPT (Zhang et al., 2022), have greatly promoted technological advancement and made outstanding contributions to the NLP community. Building upon these LLMs, researchers have further extended their capabilities and developed excellent models for various NLP tasks. Examples include Vicuna (Chiang et al., 2023) and Baize (Xu et al., 2023a). Our work is based on these LLMs and provides plug-and-play plugins that empower them with the capability of comprehending both visual and auditory content in videos.

Multi-modal Large Language Models: Researchers have been actively exploring the use of LLMs for processing multi-modal inputs (Gao et al., 2023; Li et al., 2023c). Existing approaches can be categorized into two main groups. The first category involves employing LLMs as controllers and utilizing existing multi-modal models as tools. In this approach, when receiving the user’s text instruction, the LLM recognizes the user’s intention and makes decisions about which tools to call. It then generates comprehensive responses by incorporating the results obtained from these off-the-shelf multi-modal models. Examples include ChatGPT (Wu et al., 2023a), HuggingGPT (Shen et al., 2023), and AudioGPT (Huang et al., 2023a). The second category focuses on training fundamental large-scale multi-modal models. The key idea of this line of work is to align the pre-trained foundation models for other modalities to textual LLMs. For instance, Flamingo (Alayrac et al.,

2022a) utilizes a perceiver resampler and a gated cross-attention layer to connect a frozen image encoder and LLM. BLIP2 (Li et al., 2023b) introduces a Q-Former to map learned image queries to the textual embedding space of LLMs. (Liu et al., 2023), mPLUG-owl (Ye et al., 2023) and MiniGPT4 (Zhu et al., 2023) develop instruction-following image-LLMs using image-instruction-following dataset. Video-Chat (Li et al., 2023c) and Video-ChatGPT (Maaz et al., 2023) extend image encoders to video encoders and connect them with LLMs to understand visual content in videos. PandaGPT (Su et al., 2023) utilizes multi-modal encoders from ImageBind, trained exclusively on image-instruction pairs, to enable large models to understand six modalities. Our work falls into the second category, where we train fundamental models to comprehend both the visual and auditory content in videos.

4 Examples

In this section, we show some cases to demonstrate Video-LLaMA’s multi-modal instruction-following capability in video/audio/image-grounded conversations. Figure 2 shows some of these examples, while additional cases can be found in Appendix A.

(1) Audio-visual integration perception ability.

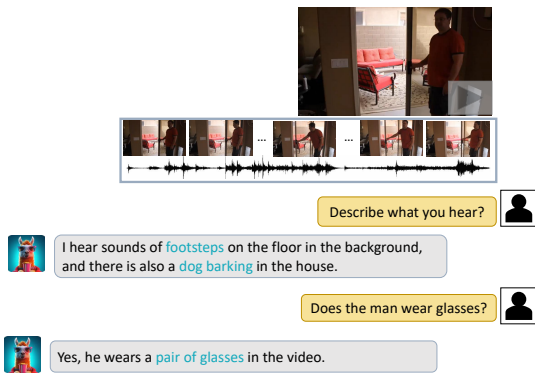
Figure 2(a) and Figure 3 show Video-LLaMA’s unique ability to comprehend auditory and visual information simultaneously. The videos in both cases contain audio. In each conversation, we pose two questions related to visual and auditory content respectively. If the model could only receive one modal, it would be unable to answer both of these questions. However, we can observe that Video-LLaMA accurately responds to both visual and auditory questions in both cases.

(2) The ability to capture temporal dynamics in videos.

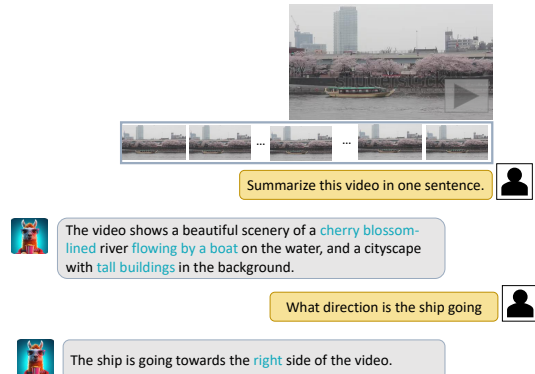
Figure 2(b) and Figure 4 illustrate the capability of Video-LLaMA to identify actions over time. It successfully describes the actions of the girl and the moving direction of the boat.

(3) The ability to perceive and understand static images.

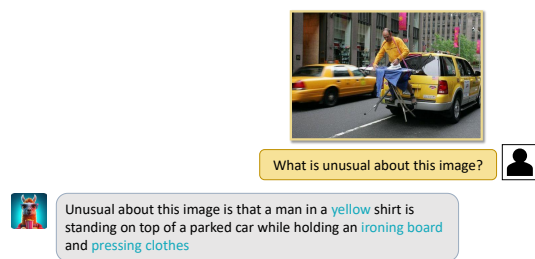
Figure 2(c) and Figure 5 show Video-LLaMA’s ability to perceive and understand pictures. Figure 2(c) demonstrates Video-LLaMA’s ability to understand the concept of "unusual" and specifically describe the unusual scene. In Figure 5, not only does Video-LLaMA accurately describe



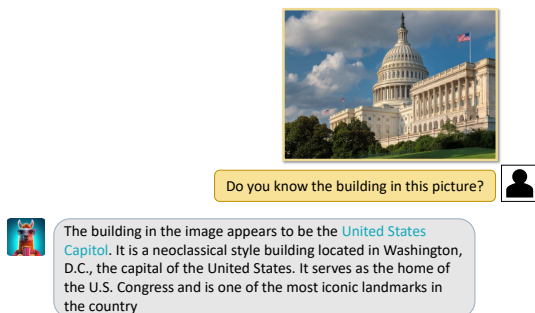
(a) A case that Video-LLaMA answers the questions based on the background sound and visual content of the video



(b) A case showing Video-LLaMA's ability to identify actions over time.



(c) A case demonstrating Video-LLaMA's ability to comprehend static images.



(d) A case demonstrating Video-LLaMA's ability to recognize famous landmarks.

Figure 2: Some examples generated by Video-LLaMA.

the main content, but it also associates it with the friendly interaction between a dog and a human.

(4) The ability of common-knowledge concept recognition. Figure 2(d) and Figure 6 demonstrate Video-LLaMA's remarkable capacity for recognizing common-knowledge concepts in visual signals. Video-LLaMA successfully recognizes famous landmarks and characters and can engage in common-sense question-answering.

5 Conclusion

In this paper, we present Video-LLaMA, a cutting-edge multi-modal framework that empowers large language models with both audio & video understanding capabilities. Our experiments demonstrated the impressive abilities of Video-LLaMA in audio and video-grounded conversations, highlighting its potential as a promising prototype for audio-visual AI assistants. We have open-sourced the entire training code and various model weights, along with detailed instructions to assist developers in utilizing our code for further development. In addition, we have provided online demo websites and offline demo deployment guides for users to experi-

ence Video-LLaMA's capabilities firsthand. We are committed to constantly maintaining and improving Video-LLaMA, and will continue to contribute to the open-source community.

6 Limitations

Although Video-LLaMA has demonstrated impressive abilities in understanding both visual and auditory content in videos, it is still an early-stage prototype and has some limitations, including: (1) Limited perception capacities: Video-LLaMA's performance is hindered by the quality and scale of the current training dataset. We are actively constructing a high-quality audio-video-text alignment dataset to enhance the model's perception capabilities. (2) Limited ability to handle long videos. Long videos (such as movies, and TV shows) contain a large volume of information and impose higher demands on computational resources. This challenge remains a crucial issue that the research community is actively working to address. (3) Hallucination. Video-LLaMA inherits the hallucination problem from the frozen LLMs. We will continue to address these challenges and develop more powerful versions for video understanding.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022a. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022b. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*.
- Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2022. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, W. Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Jiao Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Rohit Girdhar, Alaeldin El-Nouby, Zhuang Liu, Manmeet Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.
- Rongjie Huang, Mingze Li, Dongchao Yang, Jiaotong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023a. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023b. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Kunchang Li, Yanan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Ming-Hui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual*

- Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565. Association for Computational Linguistics.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugging-gpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.
- Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.
- Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yuezhe Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yilun Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linqun Liu, and Yu Wu. 2023b. On decoder-only architecture for speech-to-text and large language model integration. *arXiv preprint arXiv:2307.03917*.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023a. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.
- Haiyang Xu, Qinghao Ye, Mingshi Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qichen Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Feiran Huang, and Jingren Zhou. 2023b. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Chao Zhang, and Feiyan Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2023a. Transfer visual prompt generator across llms. *arXiv preprint arXiv:23045.01278*.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Peng Peng Wang, Yaqian Zhou, and Xipeng Qiu. 2023b. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher De-wan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

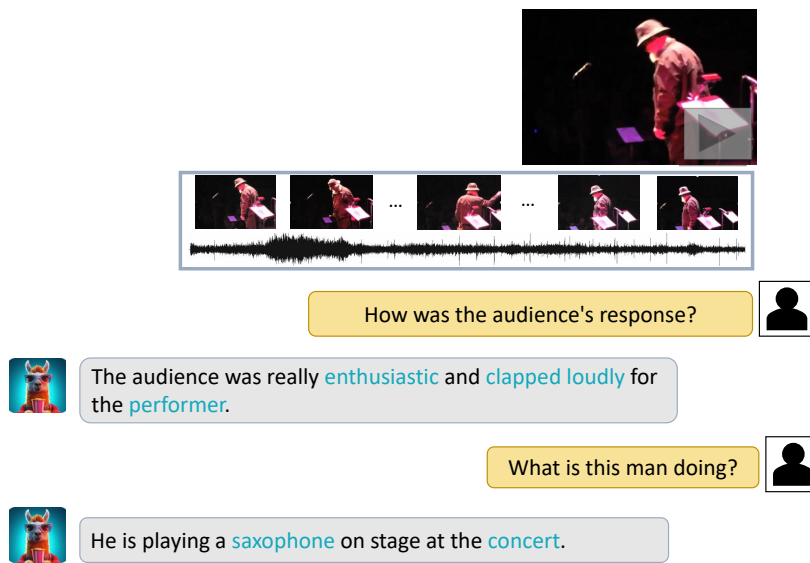


Figure 3: A case showing Video-LLaMA’s ability to identify the sound of applause in a video and infer the positive response from the audience. Additionally, it infers that a man is playing the saxophone on stage based on the visual content.

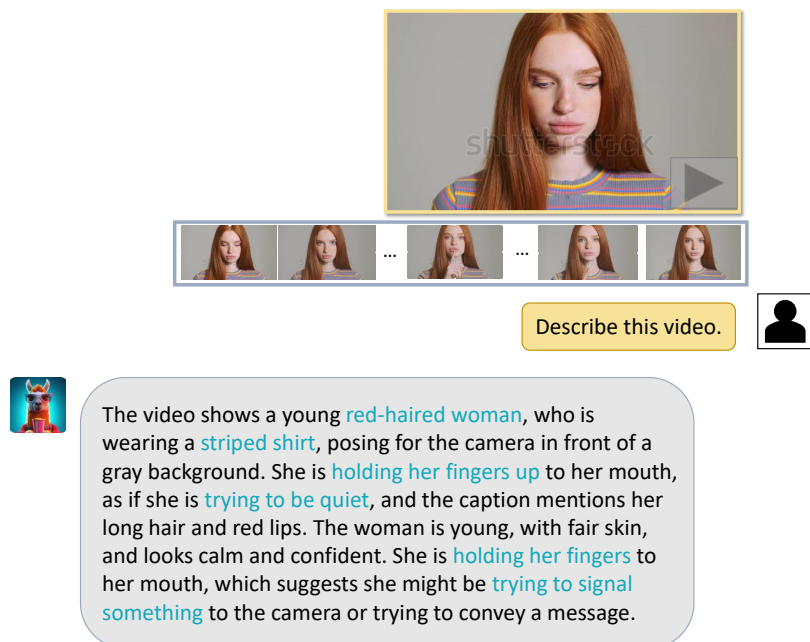
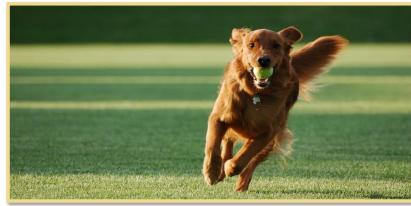


Figure 4: A case where Video-LLaMA provides a detailed description of the visual content in a dynamic video.



Describe the image in details.



The image features a beautiful and playful **golden retriever** running on a bright green lawn. The dog is **running** with a **tennis ball** in its mouth, as it is an active and lively animal that enjoys playing and having fun. **The green grass** in the background creates a vibrant contrast with the golden color of the dog's coat, adding a touch of color and texture to the scene. The dog's fur is well-groomed, shiny, and nicely groomed, highlighting its health and appearance. The brightness of the green lawn and the playful behavior of the dog suggest a joyful atmosphere, inviting viewers to engage in this lovely scenery and the dog's energetic play.

It's worth mentioning that the dog's focus and dedication to playing with the ball also showcase its natural instincts and its hunting abilities. Additionally, this image captures a fun and heartwarming moment of bonding between the dog and its human companion, as the dog's positive attitude towards the ball reflects its enjoyment of interacting with its owner and its environment. Overall, the image conveys an inviting and joyful atmosphere.

Figure 5: A case where Video-LLaMA provides a detailed description of the static image content.

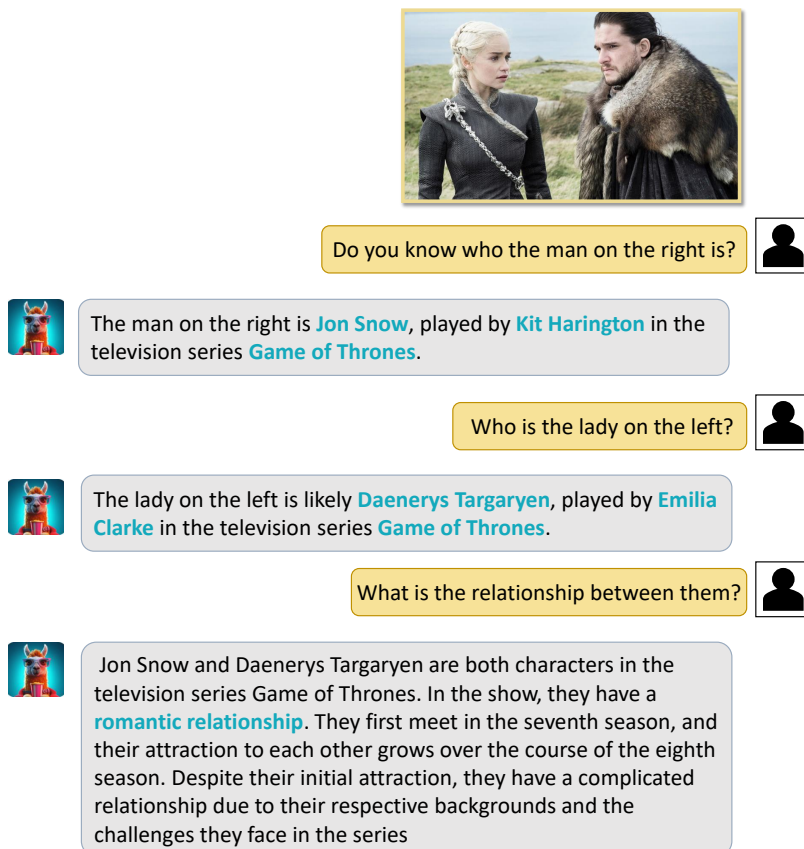


Figure 6: A case showing Video-LLaMA’s ability to recognize renowned characters and participate in video-grounded question answering.