

VALLEY: VIDEO ASSISTANT WITH LARGE LANGUAGE MODEL ENHANCED ABILITY

Ruipu Luo^{1,2*}, Ziwang Zhao^{1,3*}, Min Yang^{1*}, Junwei Dong^{1,4}, Minghui Qiu^{1,†}, Pengcheng Lu¹,
Tao Wang¹, Zhongyu Wei²

¹ByteDance Inc. ²Fudan University ⁴Chongqing University

³Beijing University of Posts and Telecommunications

{luoruipu, zhaoziwang, yangmin.priv, dongjunwei}@bytedance.com

ABSTRACT

Recently, several multi-modal models have been developed for joint image and language understanding, which have demonstrated impressive chat abilities by utilizing advanced large language models (LLMs). The process of developing such models is straightforward yet effective. It involves pre-training an adaptation module to align the semantics of the vision encoder and language model, followed by fine-tuning on instruction-following data. However, despite the success of this pipeline in image and language understanding, its effectiveness in joint *video and language understanding* has not been widely explored. In this paper, we aim to develop a novel multi-modal foundation model capable of perceiving video, image, and language within a general framework. To achieve this goal, we introduce *Valley: Video Assistant with Large Language model Enhanced ability*. Specifically, our proposed Valley model is designed with a simple projection module that bridges video, image, and language modalities, and is further unified with a multi-lingual LLM. We also collect multi-source vision-text pairs and adopt a spatio-temporal pooling strategy to obtain a unified vision encoding of video and image input for pre-training. Furthermore, we generate multi-task instruction-following video data, including multi-shot captions, long video descriptions, action recognition, causal relationship inference, etc. To obtain the instruction-following data, we design diverse rounds of task-oriented conversations between humans and videos, facilitated by ChatGPT. Qualitative examples demonstrate that our proposed model has the potential to function as a highly effective multilingual video assistant that can make complex video understanding scenarios easy. Code, data, and models will be available at <https://github.com/RupertLuo/Valley>.

1 INTRODUCTION

The rapid growth of video applications and data has created a pressing need for automated technology to analyze and comprehend video content. This is particularly important for applications such as video surveillance, content-based video retrieval, and video summarization. However, existing video understanding models are often task-specific and lack a comprehensive capability of handling diverse tasks. In light of this, we endeavor to construct a foundational model that can automatically comprehend and analyze various video elements, such as actions, objects, scenes, emotions, and other pertinent components, and subsequently integrate these components to address a broad range of tasks, including video classification, object detection, action recognition, and video question-answering. Thus, it is imperative to develop more comprehensive and general video understanding models, which represent a crucial research direction for video understanding.

Large language models (LLMs) such as ChatGPT (Ouyang et al., 2022), PaLM (Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023) have demonstrated their impressive abilities in understanding and following user intentions and instructions. These language models are able to learn

* Co-first authorship. Work done during Ruipu Luo and Ziwang Zhao’s internship at ByteDance. † Corresponding author.

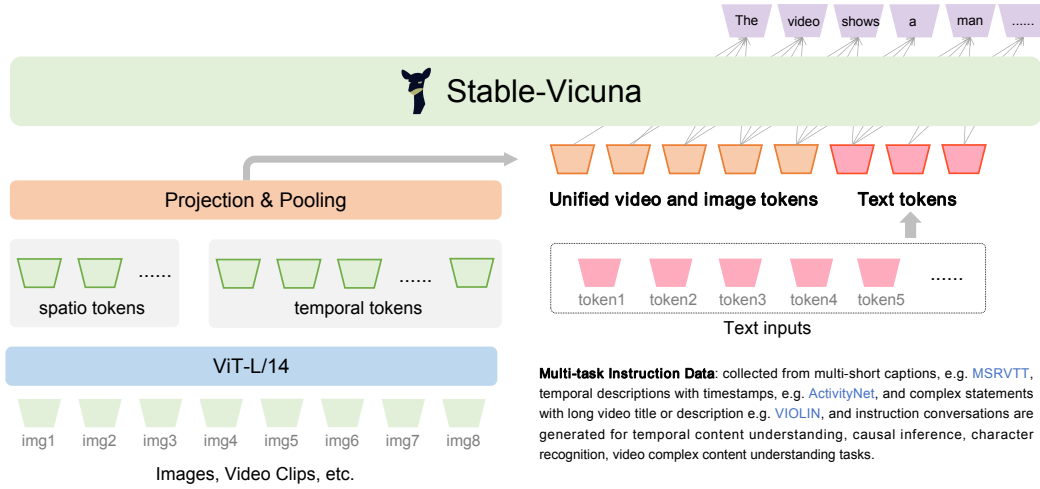


Figure 1: Valley architecture.

knowledge from large-scale text corpus well and thus can tackle a wide range of challenges, such as text generation, summarization, translation, and question-answering. Additionally, one of the most significant changes brought by LLMs is the ability to support conversational interactions like human beings. By enabling natural and intuitive conversations, LLMs have paved the way for more seamless interactions between humans and computers. Users can engage in multiple rounds of conversations with LLMs to obtain information or interact with chatbots for entertainment or support. One of the most distinguished examples of such work is ChatGPT, which has become an indispensable aspect of various applications, including customer service, healthcare, and e-commerce, commonly referred to as AI assistants.

To this end, it naturally begs the question: *Can we leverage the powerful LLMs to better fuse visual and language modalities together and create a multi-modal AI assistant?* Many recent works have yielded encouraging progress in this direction. In general, these works follow similar steps: pre-training for aligning the semantics of the vision encoder and LLM, followed by fine-tuning on the generated instruction-following data. The existing methods can be generally grouped into two main categories. The first category utilizes Q-Former from BLIP-2 (Li et al., 2023b) to align visual and textual modalities, such as InstructBLIP (Dai et al., 2023), Otter (Li et al., 2023a) and Mini-GPT4 (Zhu et al., 2023). The second category involves the use of a simple projection layer to achieve modality alignment, such as LLaVA (Liu et al., 2023) and Vicuna (Chiang et al., 2023). There also exists some attempts to extend such multi-modal works to video understanding. For example, VideoChat (Li et al., 2023c) focuses on integrating video foundation models and LLMs via a learnable neural interface and Video-LLaMA (Zhang et al., 2023) proposes a Video Q-former to capture the temporal changes in visual scenes and audio-visual signals following Mini-GPT4 (Zhu et al., 2023).

In this work, we endeavor to employ a simple projection module as the bridge between video, image, and language modalities, following LLaVA’s pre-training-then-instructions-tuned pipeline. We adopt the ViT-L/14 (Dosovitskiy et al., 2021) of CLIP (Radford et al., 2021) as the vision encoder, and then propose a spatio-temporal pooling operation to unify the visual encoding for video and image inputs. Pre-training is conducted to align the unified visual embeddings with LLM by updating the projection module, where Stable-Vicuna (Chiang et al., 2023) is the selected LLM due to its multilingual capabilities. Moreover, we introduce a multi-modal instruction-following dataset collected from various video tasks including video question-answering, long description, casual relationship inference, and action recognition. With instruction fine-tuning, our multi-modal foundation model on video understanding Valley is finally presented. As an AI assistant, we anticipate that Valley will have extensive applications in various video-related tasks and assist users in better comprehending and interpreting complex videos in a manner similar to chatting with a human in the real world. In a nutshell, our key contributions are summarized as follows:

- We propose *Valley*, a multi-modal foundation model with general perception ability of video, image, and language that could be a video assistant capable of engaging in multilingual conversations. In this work, we make a nontrivial change to the original vision encoder through a spatio-temporal pooling strategy to get unified visual tokens and only use a simple projection layer to connect vision with language inspired by LLaVA.
- We collect a large multi-modal instruction-following dataset, which focuses on video understanding and comprises diverse types of tasks, including multi-short captions, temporal descriptions with timestamps, and complex statements with long videos. We also leverage ChatGPT to generate conversations between humans and video content, which further enhances the quality and diversity of the dataset.
- We will open-source all our resources, including the pre-training dataset and the generated multi-modal instruction data collected from various video tasks. In addition, prompts on how to instruct ChatGPT to design and generate the conversations based on the original content of video data will also be made public. Finally, all model weights and chat demos will be released. This will enable researchers to reproduce our experiments and facilitate further advancements in the field of multi-modal video understanding.

2 RELATED WORK

Large Language Models. Large language models (LLMs) have achieved tremendous success in the field of natural language processing (Chowdhery et al., 2022; Ouyang et al., 2022; Hoffmann et al., 2022), with its excellent language understanding and reasoning abilities. LLMs can handle various complex tasks by comprehending prompts in a few-shot or zero-shot manner and have thus garnered widespread attention. Further, the development of a series of open-sourced LLMs, including LLaMA (Touvron et al., 2023), GLM (Zeng et al., 2022), and BLOOM (Scao et al., 2022), has fueled interest in this field and inspired a number of works such as Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and ChatGLM. Due to their large number of parameters, LLMs can not only acquire notable task transfer generalization ability but also complete actual tasks conversationally by aligning with human language instructions and preferences. Based on these, our goal is to extend the capability of LLMs to video-grounded conversation scenarios.

LLMs for Multimodal Understanding. As LLMs have demonstrated strong general capabilities in linguistic tasks, enabling LLMs to understand multimodal content has been increasingly studied. Existing methods can be divided into two technical routes, one is to employ LLMs as schedulers to schedule the existing multimodal models, and the other is to train a multimodal model based on LLMs. After receiving user instructions and the functions of each foundation model, the former treats the LLMs as a controller to call corresponding models step by step and integrates the output content of each model to generate results (Wu et al., 2023; Shen et al., 2023; Yang et al., 2023). For example, HuggingGPT (Shen et al., 2023) utilize the ChatGPT to select appropriate models in Hugging Face¹ according to their function description and summarizes their execution results. The latter equips LLMs with auxiliary modules to help them understand multimodal contents through end-to-end training (Li et al., 2023c; Zhu et al., 2023; Zhang et al., 2023; Liu et al., 2023; Su et al., 2023; Dai et al., 2023). For instance, LLaVA (Liu et al., 2023) and MiniGPT-4 (Zhu et al., 2023) connected LLaMA (Touvron et al., 2023) with a visual encoder through a projection layer, endowing it the ability to understand images. Video-LLaMA (Zhang et al., 2023) empowered LLaMA (Touvron et al., 2023) with visual and audio information via Q-Former to endow it with video-grounded conversation ability.

3 APPROACH

In this section, we first introduce the overall architecture of the proposed Valley model. And then we present the instruction-tuning datasets and our training procedure.

¹<https://huggingface.co/models>

Context type 1: Multi-Short-Captions

- (1) Two people abseil down a rope suspended from a cage hung below an extended arm of a crane.
- (2) Two men repel down lines from a crane high in the air.
- (3) A couple of men are on a device in the sky and jump out on a string.
- (4) two people are gliding down a string from a giant crane.
- (5) Two people are quickly lowered to the ground from a tall crane.
- (6) Two people are going down with harnesses from a crane.
- (7) Two people using ropes lower themselves down from a high crane.
- (8) Two people repelling down from a crane while someone plays the drums in the background.
- (9) Two men are abseiling from the top of a crane.
- (10) Two people use a wire to slide down to the ground from a box suspended by a crane.

Context type 2: temporal description with timestamps

- [0s - 7.49s] : We see a hallway with a wooden floor.
- [7.49s - 18.09s] : A dog in socks walks slowly out onto the floor as a lady films him.
- [19.37s - 36.55s]: The dog turns around and goes back to the other room.

Context type 3: Complex Statement with Long Videos

- (1) The man in the red shirt was really proud to see the girl in the blonde come out of her shell and have fun.
- (2) The blonde lady wants to know if man in the red shirt loves his new style.
- (3) The man in the red shirt explains to the blonde lady that he is in love with her and it doesn't matter if she has blonde hair or makeup on.

Table 1: Examples to illustrate 3 types of context used to create instruction following data.

3.1 ARCHITECTURE

In order to allow pre-trained LLM to understand videos and adapt videos of different lengths together with individual images, we add a spatio-temporal pooling module to the vision encoder to aggregate each frame's grid features as unified vision tokens, while keeping the rest structures the same with LLaVA (Liu et al., 2023) using a simple yet effective projection layer to connect the vision tokens to LLM. We choose Stable-Vicuna as the language interface since it exhibits superior multilingual chat abilities. The overall architecture is shown in Figure 1.

We input a video V and sample T frames by 1 FPS, which can be denoted as $V = [V_1, V_2, \dots, V_T]$. Each image obtains visual features through the pre-trained CLIP visual encoder (ViT-L/14), denoted

as $\mathbf{V}_T = \text{ViT}(\mathbf{V}_T)$. Each feature contains 256 patch features and 1 global feature ("[CLS]" token), denoted as:

$$\mathbf{V}_T = [\mathbf{V}_T^{[\text{CLS}]}, \mathbf{V}_T^1, \mathbf{V}_T^2, \dots, \mathbf{V}_T^{256}].$$

We use the average pooling method to aggregate patch features of T frames in the time dimension, denoted as:

$$\bar{\mathbf{V}}^i = \text{AvgPool}([\mathbf{V}_1^i, \mathbf{V}_2^i, \dots, \mathbf{V}_T^i]),$$

where i is the index of patches. However, this method will cause confusion in the time dimension as it averages features in different time stamps. In order to alleviate the vanishing of temporal features caused by pooling, we obtain the representation \mathbf{Z}_V of the entire video by concatenating patch features after pooling and global features of T frames, the mathematical form is expressed as follows

$$\mathbf{Z}_V = [\bar{\mathbf{V}} \oplus \mathbf{V}_1^{[\text{CLS}]} \oplus \mathbf{V}_2^{[\text{CLS}]} \oplus \dots \oplus \mathbf{V}_T^{[\text{CLS}]}],$$

where \oplus means concatenation.

We also consider a projection layer to connect visual features into the word embedding space, since it already shows its effectiveness in LLaVA (Liu et al., 2023). We utilize a trainable matrix to transform the video patch features and global features into the language feature space, which has the same dimension. Finally, the projected visual features and text embedding are input into LLM for response generation.

$$\hat{\mathbf{Z}}_V = \text{LLM}(\mathbf{Z}_V).$$

3.2 INSTRUCTION TUNING DATA COLLECTION

We build a video-centric multimodal instruction data based on **MSRVTT** (Tan et al., 2020), **VATEX** (Wang et al., 2019), **ActivityNet** (Fabian Caba Heilbron & Niebles, 2015) and **VIOLIN** (Liu et al., 2020). The corresponding question-answer generations are produced by stable-vicuna (Chiang et al., 2023) based on three types of context as shown in table 1. Context type 1 data is collected from **MSRVTT** (10k) and **VATEX** (22k) datasets, each data containing several captions for one video. Context type 2 data is from **ActivityNet** (10k), and each video in this dataset is described by activity sentences with timestamps. Context type 3 is from **VIOLIN** (5.8k), and contains 3 complex statements in the video.

Using these three types of contexts and open source LLM (Stable-Vicuna), we generate instruction tuning data by referring to the prompt motivated by LLaVA (Liu et al., 2023) and VideoChat (Li et al., 2023c), the prompt is shown in the appendix. The generated conversation and question-answer pairs are shown in the Table 2. we generate 42k conversations and 5.8k QA pairs in total. The generated conversation data covers basic video content description (object information, etc.) and temporal content understanding, and the QA data includes causal inference, character recognition, and video complex content understanding.

3.3 TRAINING

Inspired by LLaVA (Liu et al., 2023), we adopt a two-stage training framework. The first stage pre-trains the projection layer for feature alignment, and the second stage fine-tunes the language model and projection layer.

Valley supports the input of any number of images, so in the pre-training phase, we use image-text pairs and video-text pairs for pre-training. The pre-training data includes 595K CC3M image-text pairs provided by LLaVA (Liu et al., 2023), and 702K WebVid2M (Bain et al., 2021) video-text pairs filtered by us refer to the filtering method in LLaVA (Liu et al., 2023). Both images and videos are input into the model in a unified way, and the prompt is as follows:

```
### Xsystem message
### Human: Xinstruction < patch1 > ... < patch256 > < frame1 > ... < frameT >
### Assistant:
```

If a single image is input, the number of frames is 1. The image-text pair and video-text pair are constructed as a single-round dialogue, using various questions to inquire about the video content and answering with the corresponding caption.

Conversation for Context type1

[User]: What are the people in the video doing?

AI: The people in the video are abseiling down a rope from a crane. They are suspended from a cage that is hung below an extended arm of the crane.

[User]: How many people are in the video?

AI: There are two people in the video.

[User]: How are the people in the video descending from the crane?

AI: The people in the video are using ropes and harnesses to lower themselves down from a high crane.

[User]: Is there any background music in the video?

AI: Yes, there is background music in the video. Someone is playing the drums in the background while the people are repelling down from the crane.

[User]: Can you describe the device that the people are using to descend from the crane?

AI: Based on the video, the device the people are using to descend from the crane appears to be a cage. The cage is suspended from an extended arm of the crane and has a rope attached to it. The two people are using the rope to lower themselves down from the cage.

Conversation for Context type2

[User]: What is in the video?

AI: The video depicts a hallway with a wooden floor. At around 7.5 seconds, a dog wearing socks walks slowly onto the floor as a lady films him.

[User]: How many dogs are in the video?

AI: There is only one dog present in the video.

[User]: What does the dog do in the video?

AI: The dog walks slowly onto the wooden floor wearing socks while being filmed by the lady. After a short time, at around 19 seconds, the dog turns around and goes back to the other room.

[User]: Why is the dog wearing socks in the video?

AI: It's possible that the dog is wearing socks to prevent scratches on the wooden floor or to provide better traction for its paws. Alternatively, the dog may be wearing the socks as a fashion statement or for some other purpose not related to the floor.

QA Pairs for Context type3

Question: Why was the man in the red shirt proud? (Casual Inference)

Answer: The man in the red shirt was proud because he saw the girl in the blonde come out of her shell and have fun.

Question: Who wants to know if the man in the red shirt loves his new style? (Character Identifying)

Answer: The blonde lady wants to know if the man in the red shirt loves his new style.

Question: What does the man in the red shirt explain to the blonde lady? (Video Understanding)

Answer: The man in the red shirt explains to the blonde lady that he is in love with her and it doesn't matter if she has blonde hair or makeup on.

Table 2: Examples of instruction data generated by our prompt based on the 3 contexts mentioned in Table 1.

As introduced in Section 3.2, we construct an instruction dataset that consists of 42k conversations and 5.8k QA pairs. These data contain a lot of causal reasoning and temporal understanding content. In order to enhance the ability to describe visual content in detail, we also gather 150k image instruction data from LLaVA (Liu et al., 2023), and 11K video instruction data from VideoChat(Li et al., 2023c).

4 EXPERIMENTS

4.1 SETTINGS

In our experiments, we employ the Stable-Vicuna (Chiang et al., 2023) as the LLM backbone and the pre-trained ViT-L/14 from CLIP to encode videos and images. We first pre-train Valley for one epoch with a learning rate of $2e-3$ and then fine-tune the model for three epochs with a learning rate of $2e-5$ on the instruction dataset. All the experiments are conducted on $8 \times A100$ 80G GPUs.

4.2 QUALITATIVE ANALYSIS

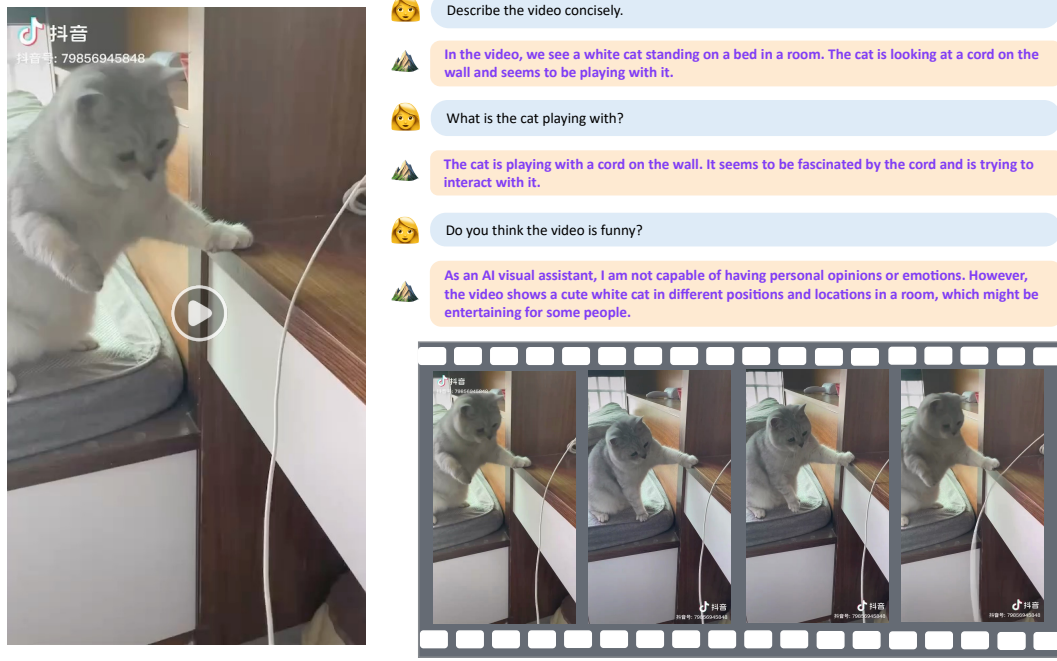


Figure 2: Video description, funny things identification, and recognition of interactions of the objects in the video.

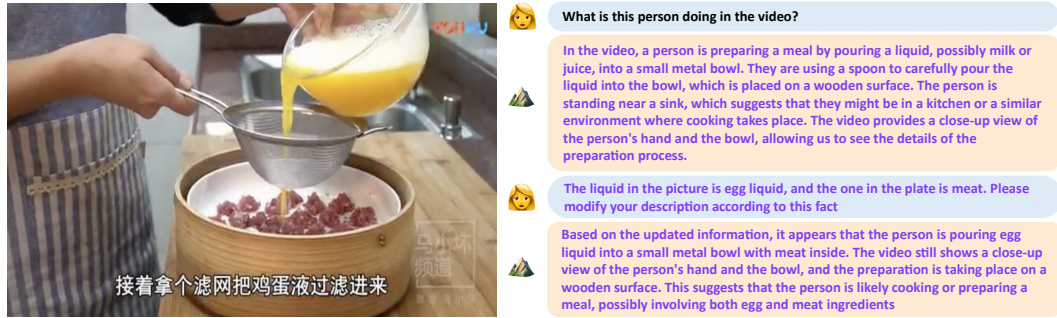


Figure 3: Image question answering and factual description with fixing mistakes.

In this section, we present some cases from our Valley to demonstrate the superior video understanding and instruction following the capabilities of Valley. As shown in Figure 2-4.

5 LIMITATION

Our current Valley relies on video and language input, however, incorporating audio input will present a significant opportunity to enhance the overall user experience. In the future, we will extend our Valley to a multi-modal assistant that leverages audio, visual, and textual information. On the other hand, the Valley still has some shortcomings in understanding Chinese. To tackle this challenge, we plan to construct more Chinese multi-modal data to iterate the model and improve the understanding ability of Chinese in future research. There also exists a hallucination problem in our model which is inherent in LLMs. Solving the language hallucination problem is one of the recent hotspots and difficulties.

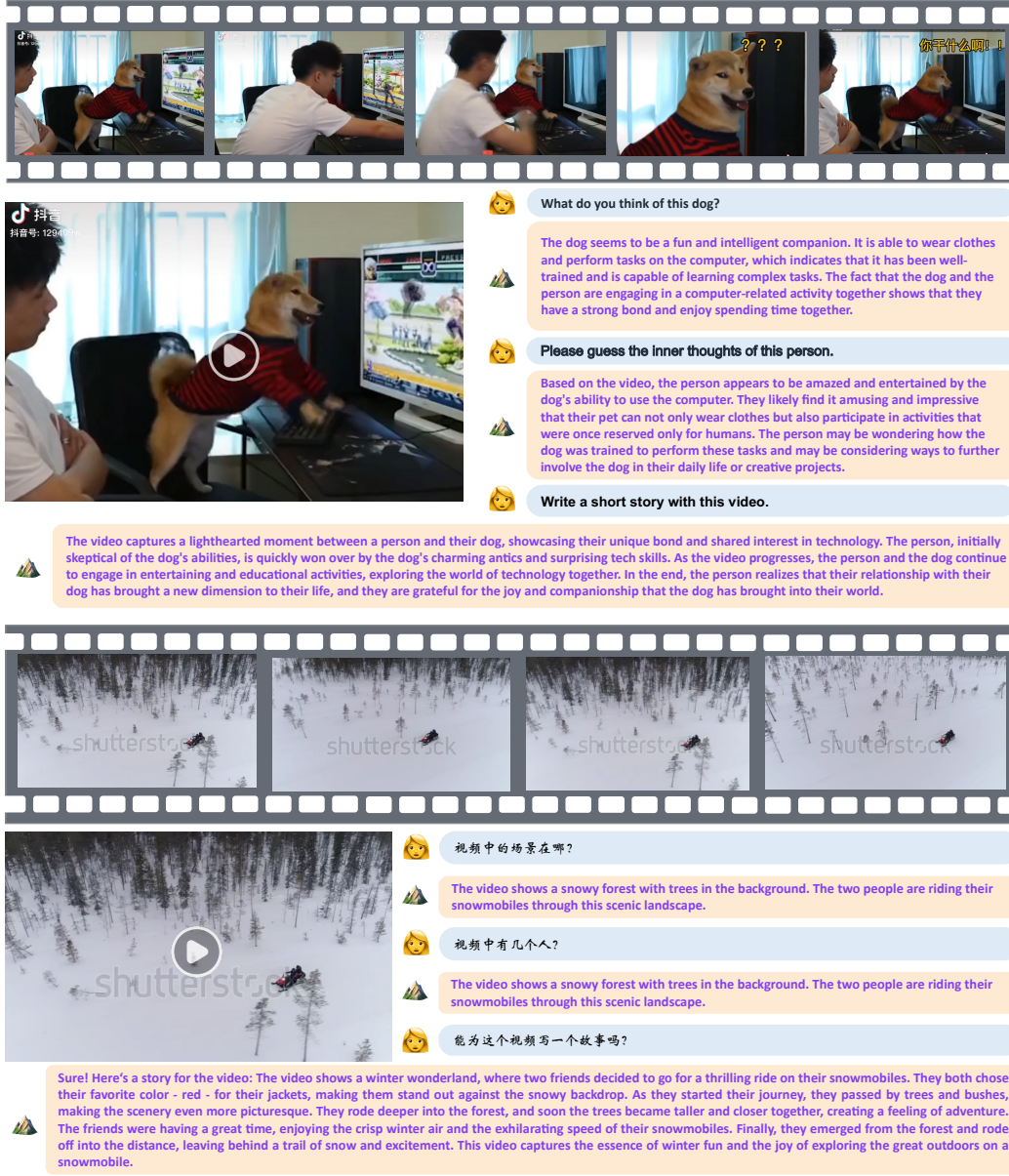


Figure 4: Capture the main content of the video, do the causal inference, and identify objects and relationships in the video. Chinese prompts are also well understood.

6 CONCLUSION

The objective of our work is to construct a foundation model that is capable of perceiving video, image, and language in a multi-modal manner. To address this issue, we propose a framework called Valley, which stands for Video Assistant With Large Language Model Enhanced Ability. We utilize a spatio-temporal pooling approach to extract a unified vision encoding from video and image inputs, gather a large set of vision-text pairs for pre-training, and then generate a multi-task instruction-following video dataset among which the conversations are designed with the help of ChatGPT. Ultimately, our goal is to create a more intuitive, personalized, and human-like interaction between humans and machines.

REFERENCES

- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations*, 2021.
- Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726, 2023a.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597, 2023b.
- Kunchang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023c.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023.
- Jingzhou Liu, Wenhui Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10900–10910, 2020.

- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580, 2023.
- Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all, 2023.
- Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. Learning to discretely compose reasoning module networks for video captioning. *CoRR*, abs/2007.09049, 2020.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.
- Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: prompting chatgpt for multimodal reasoning and action. *CoRR*, abs/2303.11381, 2023.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. *CoRR*, abs/2210.02414, 2022.
- Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023.