

Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models

Muhammad Maaz^{1*}, Hanoona Rasheed^{1*}, Salman Khan^{1,2}, Fahad Shahbaz Khan^{1,3}

¹Mohamed bin Zayed University of AI, UAE

²Australian National University, Australia ³Linköping University, Sweden

Abstract

Conversation agents fueled by Large Language Models (LLMs) are providing a new way to interact with visual data. While there have been initial attempts for image-based conversation models, this work addresses the under-explored field of *video-based conversation* by introducing Video-ChatGPT. It is a multimodal model that merges a video-adapted visual encoder with an LLM. The resulting model is capable of understanding and generating detailed conversations about videos. We introduce a new dataset of 100,000 video-instruction pairs used to train Video-ChatGPT acquired via manual and semi-automated pipeline that is easily scalable and robust to label noise. We also develop a quantitative evaluation framework for video-based dialogue models to objectively analyze the strengths and weaknesses of video-based dialogue models. Code: <https://github.com/mbzuai-oryx/Video-ChatGPT>.

1 Introduction

The surge of deep learning applications for video understanding has led to major advancements in video-related tasks. However, the current video understanding models are still unable to hold an open-ended conversation about the video content in a coherent manner. A video-based dialogue model can revolutionize video search, surveillance operations and help summarize key events and abnormal event detection. Above all, it can provide a unified human-understandable interface to video-related tasks such as action recognition, localization, detection, segmentation, retrieval, and tracking. Further, such a capability is of great interest as it will demonstrate the model’s ability to encode temporal and spatial cues, contextual relationships and long-term dependencies.

Recent advancements in multimodal understanding are largely based on the combination of pre-

trained *image* models with Large Language Models (LLMs) but generally do not consider video inputs (Liu et al., 2023; Zhu et al., 2023; Li et al., 2022, 2023a; Dai et al., 2023). It is therefore interesting to leverage the vast capabilities of LLMs for video understanding tasks in a way that would not only maintain the temporal and spatial characteristics but also be adept at generating human-like conversations about videos. In this paper, we introduce Video-ChatGPT, a novel multimodal model that merges the representational abilities of a pre-trained visual encoder and the generative powers of an LLM, capable of understanding and conversing about videos.

Video-ChatGPT leverages an adapted LLM (Liu et al., 2023) that integrates the visual encoder of CLIP (Radford et al., 2021) with Vicuna (Chiang et al., 2023) as a language decoder, fine-tuned on generated instructional image-text pairs. Our approach further adapts the design for spatiotemporal video modeling and fine-tunes the model on video-instruction data to capture temporal dynamics and frame-to-frame consistency relationships available in video data. In contrast to other concurrent works for video-based conversation (Li et al., 2023b; Zhang et al., 2023; Su et al., 2023), Video-ChatGPT excels at temporal understanding, spatial consistency and contextual comprehension as demonstrated by our extensive evaluations.

A fundamental contribution of this work is the creation of a dataset of 100,000 video-instruction pairs using a combination of human-assisted and semi-automatic annotation methods. Each pair consists of a video and its associated instruction in the form of a question-answer. This provides Video-ChatGPT with a large and diverse dataset to learn from, increasing its video-specific understanding, attention to temporal relationships and conversation capabilities.

Moreover, we introduce the first quantitative video conversation evaluation framework for bench-

¹Equal contribution.

marking, allowing for a more accurate evaluation of the performance of video conversation models. This framework evaluates models on a variety of capabilities, such as correctness of information, detail orientation, contextual understanding, temporal understanding, and consistency.

The contributions of this work are as follows,

- We propose Video-ChatGPT, a video conversation model capable of generating meaningful conversations about videos. It combines the capabilities of LLMs with a pretrained visual encoder adapted for spatiotemporal video representations.
- We introduce 100,000 high-quality video instruction pairs together with a novel annotation framework that is scalable and generates a diverse range of video-specific instruction sets.
- We develop the first quantitative video conversation evaluation framework for benchmarking video conversation models. We demonstrate Video-ChatGPT to perform well compared to concurrent conversational engines for videos such as Video Chat (Li et al., 2023b).

2 Related Works

Vision Language Models: Significant advancements in the field of computer vision have recently been observed due to the development of many foundational vision-language models. These models represent a significant leap towards creating general-purpose vision models capable of tackling various tasks simultaneously (Radford et al., 2021; et al, 2022; Gupta et al., 2022; Maaz et al., 2022). A prime example is CLIP (Radford et al., 2021), which is trained on 400M image-text pairs and has demonstrated impressive zero-shot performance on numerous benchmarks. It has been employed in various downstream applications, from image-based object detection and segmentation (Rasheed et al., 2022; Liang et al., 2023) to 3D applications (Rozenberszki et al., 2022; Ni et al., 2022). Numerous attempts have also been made to adapt CLIP for video applications (Wang et al., 2021; Ni et al., 2022). Similar to our design, ViFi-CLIP (Rasheed et al., 2023) suggests employing temporal pooling across video frames to adapt the image-based CLIP model for video-based tasks.

Large Language Models: The field of natural language processing has witnessed a paradigm shift

with the advent of pretrained Large Language Models (LLMs) such as GPT (Brown et al., 2020), LLaMA (Touvron et al., 2023), OPT (Zhang et al., 2022), and MOSS (OpenLM Lab, 2023). These models exhibit extraordinary abilities like language generation and in-context learning, and their knack for understanding intricate tasks given user prompts in a zero-shot manner reflects their impressive adaptability and generalization. The proven capabilities of LLMs have encouraged researchers to fine-tune them to maximize their proficiency.

A key strategy in this pursuit is instruction tuning. This approach focuses on improving the model’s alignment with user intentions and optimizing its output quality. For instance, InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2023) significantly benefit from this technique, showcasing improvements in diverse conversational interaction capabilities and their aptitude to answer a broad range of complex questions. This effective approach has recently been employed in open-source models like Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023), both developed using the LLaMA (Touvron et al., 2023) framework, resulting in performance improvements.

Pre-trained LLMs in Vision-Language Tasks:

The recent strides in multimodal understanding have primarily been driven by the integration of image-based vision models with LLMs. Seminal contributions such as Flamingo (et al, 2022) and BLIP-2 (Li et al., 2023a) have demonstrated the power of utilizing web-scale image-text data, as well as pioneering techniques in cross-modal alignment, to exhibit dynamic abilities in conversational and few-shot learning contexts. Building on this foundation, MiniGPT-4 (Zhu et al., 2023) allows image-based conversations by integrating BLIP-2 and Vicuna for zero-shot image comprehension.

Equally significant is the emergence of LLaVA (Liu et al., 2023), a model derived from the LLaMa architecture, leveraging GPT-4’s language proficiency to generate multimodal instruction-following data. With instruction tuning applied on the derived data, LLaVA has displayed interesting multimodal chat capability, hinting at the scalability potential of such a methodology. In addition, InstructBLIP (Dai et al., 2023) has demonstrated strong image-based dialogue capabilities via vision-language instruction tuning by innovating with instruction-aware visual feature extraction.

More closely related to our work, VideoChat (Li et al., 2023b) employs selective components of

video foundational models (Wang et al., 2022) and image foundation models (Li et al., 2023a), and integrates them with LLMs (Chiang et al., 2023) in conjunction with few learnable layers, tuned using a two-stage lightweight training. Additionally, they construct a video-specific dataset using off-the-shelf vision-language models (Wu et al., 2022; Li et al., 2023a; Huang et al., 2023; Wang et al., 2022) for generating noisy detailed textual descriptions to enhance the training of video-centric conversational models.

Different from VideoChat, we propose a novel human assisted and semi-automatic annotation framework for generating high quality instruction data for videos. Our simple and scalable architecture design utilizes pretrained CLIP (Radford et al., 2021) to generate spatiotemporal features which help Video-ChatGPT in generating meaningful video conversation. Further, we are the first to propose quantitative framework for evaluating video conversation tasks (see Section "Video Instruction Data Generation" for more details).

3 Video-ChatGPT

Video-ChatGPT is a large vision-language model that aligns video representations with a Large Language Model (LLM), thus enhancing its ability to generate meaningful conversation about videos. Our approach draws from the approach employed in designing vision-language (VL) models for the video domain. Given the limited availability of video-caption pairs and the substantial resources required for training on such data from scratch, these models commonly adapt pretrained image-based VL models for video tasks (Ni et al., 2022; Wang et al., 2021; Rasheed et al., 2023). We adopt a similar approach, starting with the Language-aligned Large Vision Assistant (LLaVA)(Liu et al., 2023) as our foundation.

LLaVA is a LMM that integrates the visual encoder of CLIP (Radford et al., 2021) with the Vicuna language decoder (Chiang et al., 2023) and is fine-tuned end-to-end on generated instructional vision-language data. We fine-tune this model using our video-instruction data, adapting it for video conversation task. The video-instruction data is obtained as a combination of manual and automated pipelines in our proposed instruction generation setup. This adaptation on video-specific instructions allows for accommodating additional temporal dynamics, frame-to-frame consistency, and

long-range relationships present in video data. As a result, our Video-ChatGPT excels in video reasoning, creativity, and understanding of spatial, temporal, and action-oriented components within videos.

3.1 Architecture

We use CLIP ViT-L/14, which is pretrained using large-scale visual instruction tuning in LLaVa, as the visual encoder. However, LLaVa visual encoder is meant for images, which we modify to capture spatiotemporal representations in videos. Given a video sample $V_i \in \mathbb{R}^{T \times H \times W \times C}$ with T frames, the visual encoder generates temporal and spatial features. The visual encoder encodes the T frames independently as a batch of images and produces frame-level embeddings $x_i \in \mathbb{R}^{T \times h \times w \times D}$, where $h = H/p, w = W/p$. Here p is the patch size (*i.e.* 14 for ViT-L/14), and we represent the number of tokens as N , where $N = h \times w$. Frame-level embeddings are average-pooled along the spatial dimension to obtain a *video-level temporal representation* $t_i \in \mathbb{R}^{T \times D}$. This operation implicitly incorporates temporal learning through the aggregation of multiple frames. Similarly, the frame-level embeddings are average-pooled along the temporal dimension to yield the *video-level spatial representation* $z_i \in \mathbb{R}^{N \times D}$. The temporal and spatial features are concatenated to obtain the video-level features v_i ,

$$v_i = [t_i \quad z_i] \in \mathbb{R}^{(T+N) \times D}. \quad (1)$$

A simple trainable linear layer g , projects these video-level features into the language decoder’s embedding space, transforming them into corresponding language embedding tokens Q_v ,

$$Q_v = g(v_i) \in \mathbb{R}^{(T+N) \times K}. \quad (2)$$

Note that the function g acts as an adapter and can be implemented with more complicated architectures as well. However, we opt for a simplistic design that gives competitive performance compared to more sophisticated choices in our experiments. The text queries are tokenized to the same dimensions, $Q_t \in \mathbb{R}^{L \times K}$. Here L represents the length of text query. Finally, Q_v is concatenated with Q_t and input to the language decoder.

3.2 Video Instruction Tuning

We employ instruction-tuning of the LLM on the prediction tokens, utilizing its original autoregressive training objective. The pretrained model

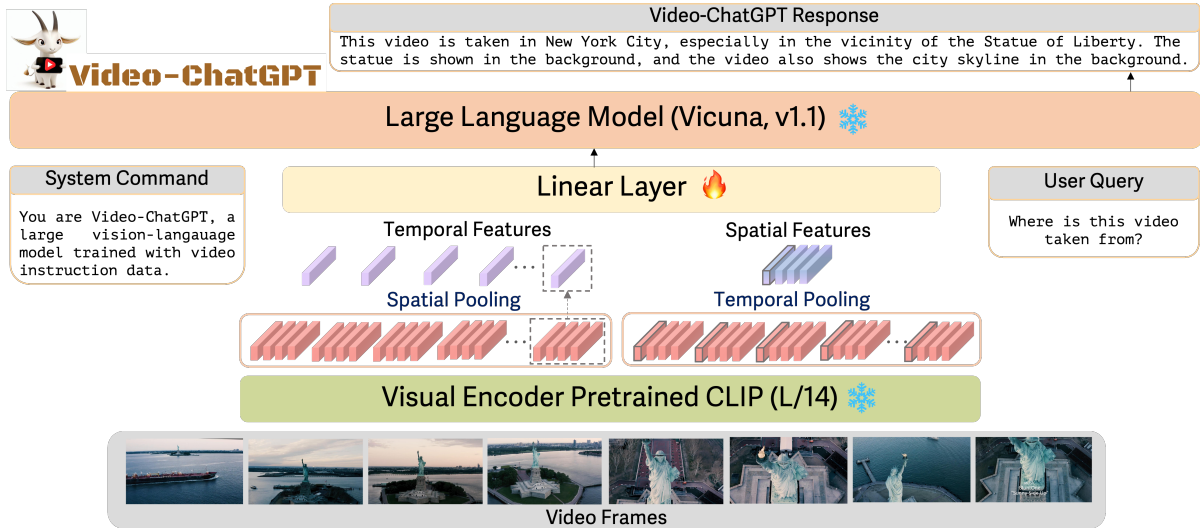


Figure 1: **Architecture of Video-ChatGPT.** Video-ChatGPT leverages the CLIP-L/14 visual encoder to extract both spatial and temporal video features. This is accomplished by averaging frame-level features across temporal and spatial dimensions respectively. The computed spatiotemporal features are then fed into a learnable linear layer, which projects them into the LLMs input space. In our approach, we utilize the Vicuna-v1.1 model, comprised of 7B parameters, and initialize it with weights from LLaVA (Liu et al., 2023).

is finetuned with curated, high-quality video-text pairs. During the finetuning phase, we use predefined prompts based on the following template:

USER: <Instruction> <Vid-tokens>
 Assistant:

Using the notations, we can represent it as,

USER: < Q_t > < Q_v > Assistant:

In this prompt, the <Instruction> represents a question pertaining to the video, randomly sampled from the training set of video-question-answer pairs. Questions can be general, asking to describe the video, or they may relate to specific temporal, spatial, or creative aspects of the video content. The prediction answer <Answer> corresponds to the specific question asked. Throughout the training, the weights for both the video encoder and LLM remain frozen, and the model maximizes the likelihood of predicting tokens representing the answer by adapting the linear layer. Consequently, the video features Q_v become aligned with the pre-trained LLM word embeddings, equipping Video-ChatGPT with the ability to produce more natural and dependable responses.

4 Video Instruction Data Generation

In this section, we discuss our data-focused approach, which uses both human-assisted and semi-automatic annotation methods to generate high-quality video instruction data. This data is cru-

cial for training Video-ChatGPT, ensuring accurate and meaningful responses. Our data collection involves two key methods. The *human-assisted annotation*, involves expert annotators analysing video content and providing detailed descriptions. This process generates data rich in context and detail, which helps our model understand complex aspects of video content. On the other hand, the *semi-automatic annotation framework* is more cost-effective and scalable. Leveraging state-of-the-art vision-language models, this method generates broad, high-volume annotations, thus increasing the quantity of data without compromising the quality substantially. Through these combined methods, we have successfully accumulated a robust set of 100,000 video-instruction pairs. This extensive dataset is crucial in fine-tuning our model to comprehend video content effectively, integrating both spatial and temporal cues into its understanding.

Our instructional data is both diverse and comprehensive, incorporating a wide range of data types. These include detailed descriptions, summarizations, question-answer pairs, tasks that stimulate creativity or generation of new ideas, and conversational tasks. The data spans a broad spectrum of concepts, ranging from visual appearance and temporal relations to complex reasoning tasks and beyond, providing a diverse training ground for our model to learn from.



Original Dataset Annotation:

An intro leads into various pictures of horse tools and a woman speaking to the camera holding a brush. She then brushes the horse's mane thoroughly while still looking to the camera and speaking.

Human Annotated:

The video begins with an intro slide, where there is a picture of the woman standing next to a brown and white horse. The woman is wearing a green jacket over a white shirt. Then an intro leads into various pictures of horse tools and a woman speaking to the camera holding a brush. The video is about "how to groom your horse to shine in four easy steps with a finishing brush. She then brushes the horse's mane thoroughly while still looking at the camera and speaking. The horse is black and the video is shot inside a stable. She wears a green jacket over a long-sleeved purple t-shirt. She is explaining how to use a finishing brush. She shows that the brush has a raised edge, which will help remove more dirt. She uses long strokes to brush the hair of the horse in the direction of hair growth. She cleans from the top to the legs. She then explains that the brush needs to be cleaned with a tool named curry. Finally, the video ends with a picture of the black horse all groomed up and credits to the video.



Original Dataset Annotation:

A close up of a christmas tree is shown followed by close ups of ornaments. Two people are then seen moving around the tree decorating as well as turning the lights off. They finish decorating the tree and playing with one another and laughing. In the end close ups of the trees are shown as well as a bear.

Human Annotated:

In the video, we see a beautifully decorated Christmas tree with lush green branches adorned with bright and colorful ornaments. As the camera pans over the ornaments, they glisten in the light, reflecting the colors of the rainbow. Two people are then shown moving around the tree, hanging ornaments and stringing lights, carefully placing each ornament in its designated spot. As they work, they chat and joke around, enjoying each other's company and the festive spirit. After they finish hanging the ornaments, they step back and admire their work, giggling and hugging each other. The camera captures close-ups of the finished tree, showing off all of its beauty and sparkle. The lights twinkle and glow, casting a warm and cozy ambiance around the room. The scene is one of holiday cheer and joy, a perfect representation of the Christmas spirit. Finally, the camera zooms out and we see a cute little bear toy dancing under the tree, surrounded by beautifully wrapped presents. Overall, the video showcases the joy and warmth of the festive season, with a beautifully decorated tree, love, and the spirit of giving and sharing.

Figure 2: **Examples of data enrichment via human-assisted annotation.** Human annotators augment video descriptions from video-caption datasets. The captions are enriched by integrating detailed information about spatial and temporal aspects, object relationships, reasoning, scene descriptions, and the chronological sequence of events.

4.1 Human-assisted Annotation

In this process, we leverage datasets containing video-caption pairs and utilize the expertise of human annotators to enrich the original ground truth annotations. Specifically, we use a subset of ActivityNet-200 (Fabian Caba Heilbron and Niebles, 2015) which provides concise ground truth descriptions of various activities in distinct video segments.

The annotators further enrich the captions by adding comprehensive information about physical appearances and spatial and temporal localization, among other critical contextual details. Figure 2 shows an example of how a ground truth caption is enriched using human-assisted annotation.

4.2 Semi-automatic Annotation Framework

In addition to the rich human-assisted annotations, we also harness the capabilities of advanced dense image vision-language models, developing a semi-automatic annotation framework. This approach is cost-effective and scalable, thereby increasing the

quantity of data without substantially compromising the quality.

Similar to the human-assisted process, this framework also leverages datasets containing video-caption pairs. We enrich these datasets using contextual information drawn from off-the-shelf dense prediction and captioning image-based vision-language models. These models provide predictions that deliver additional contextual information, thereby enriching the video captions. We developed a comprehensive method that combines these predictions, and utilize specific models for the purpose of eliminating noisy or irrelevant context from the data. This ensures that the data maintains its accuracy and relevance.

Building on the use of off-the-shelf models, we apply pretrained models like BLIP-2 (Li et al., 2023a) and GRiT (Wu et al., 2022) for key-frame analysis in the videos. The BLIP-2 image-captioning model generates frame-level captions, while the GRiT dense captioning model provides detailed captions for scene objects. Additionally,

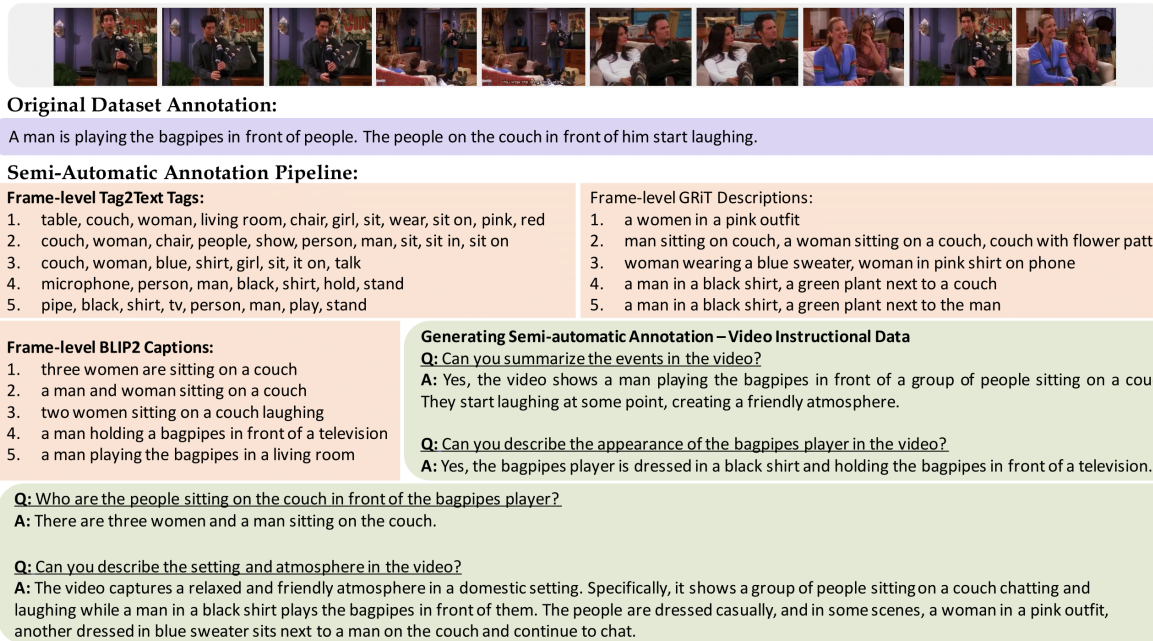


Figure 3: **Examples of generating instructional data using our proposed semi-automatic annotation pipeline.** We employ off-the-shelf dense prediction and captioning models to augment video descriptions. BLIP-v2 (Li et al., 2023a) generates frame-level captions, while GRiT (Wu et al., 2022) is utilized for dense frame captions. Tag2Text (Huang et al., 2023) generates tags for each key-frame, aiding in eliminating noise (e.g. the GRiT descriptions containing *flower pattern* and *on phone* would be discarded as there are no corresponding tags detected). Finally, we query GPT-3.5 with in-context examples to generate video-instructional data.

the pretrained Tag2Text (Huang et al., 2023) model is used to generate tags for each key-frame of the video. Despite their utility, these models can introduce noise into the data.

To ensure high-quality data and mitigate noise, we implement three key steps. *First*, we maintain a high prediction threshold for all off-the-shelf models to uphold accuracy. *Second*, we employ a specialized filtering mechanism that removes any frame-level caption from BLIP-2 or GRiT not matching with the Tag2Text frame-level tags. This process involves extracting words from the frame-level captions that are within the predefined Tag2Text tags vocabulary and eliminates any captions that contain words not in the tags for a given frame. This strategy acts as an additional filtering layer and enriches the captions by integrating predictions from multiple models.

In the *third* step, we merge frame-level captions and use the GPT-3.5 model to generate a singular, coherent video-level caption. This step augments the original ground truth caption with context from these models. We also direct GPT-3.5 to discard inconsistent information across frames, ensuring a precise, contextually rich video instruction dataset. Figure 3,4 illustrates how a ground truth caption

is enriched using this process after all three refinement stages to generate instructional data and detailed descriptive caption. All of our designed prompts for in-context learning along with the curated dataset will be made publicly available.

4.3 GPT-Assisted Postprocessing

Lastly, we implement a GPT-Assisted Post-processing mechanism that refines and optimizes the enriched annotations, in order to generate high-quality video instructional data. We prompt GPT-3.5 model to create question-answer pairs from the enriched and detailed captions that cover a wide variety of aspects using in-context learning. These aspects include detailed descriptions, summarizations, question-answer pairs, tasks that stimulate creativity or the generation of new ideas, and conversational tasks.

Each of these elements plays a crucial role in our data-centric approach. Our ultimate goal is to create a video-based conversation model that is accurate, capable of understanding video content from both spatial and temporal cues, and adept at engaging in conversations.



Original Dataset Annotation:

A countdown is shown followed by a woman catching balls thrown at her. The number counts down more and shows more clips of women playing dodgeball. Several more numbers and clips are shown ending with people performing impressive throws.

Semi-Automatically Annotated:

The video starts with a countdown, and as the numbers decrease, we see a woman standing in an arena. Suddenly, multiple balls are thrown at her from different directions, and we see her successfully catch them one by one with great agility and precision. As the countdown continues, we see more clips of women playing dodgeball in the same arena, demonstrating their skills in avoiding being hit by the balls. The countdown continues to show more numbers and clips of people playing dodgeball. We see players on different teams with colorful jerseys, throwing balls at their opponents and trying to avoid getting hit themselves. In some of the clips, we see players making impressive throws, spinning and turning before they release the ball, which adds to the excitement of the game. Towards the end of the video, we see people performing some incredible throws, bouncing the ball off the floor or walls before hitting their targets. The video ends with a final impressive move, where someone jumps from a trampoline and throws the ball with great power and accuracy. Overall, the video showcases the excitement, skills, and competitiveness of the game of dodgeball, emphasizing the agility and precision required to play it successfully.

Figure 4: **Examples of data enrichment using our proposed semi-automatic annotation.** We employ off-the-shelf dense prediction and captioning models (Li et al., 2023a; Wu et al., 2022; Huang et al., 2023) to augment video descriptions. All additional context elements are combined with the video captions and undergo a GPT-assisted post-processing stage, generating the final detailed description.

Evaluation Aspect	Video Chat	LLaMA Adapter	Video-LLaMA	Video-ChatGPT
Correctness of Information	2.23	2.03	1.96	2.40
Detail Orientation	2.50	2.32	2.18	2.52
Contextual Understanding	2.53	2.30	2.16	2.62
Temporal Understanding	1.94	1.98	1.82	1.98
Consistency	2.24	2.15	1.79	2.37

Table 1: **Performance benchmarking of text generation models.** An in-depth comparative analysis of Video-ChatGPT and Video Chat (Li et al., 2023b) across five key evaluation aspects we propose in our benchmark. For a fair comparison, 7B variants are used for all the models. Video-ChatGPT shows competent performance across all key aspects.

5 Experiments

5.1 Implementation Details

We use LLaVA (Liu et al., 2023) as our baseline model and finetune it on our 100K video instruction pairs. We only update the linear layer projecting the video features to the LLMs’ input space, while the rest of the architecture is kept frozen. We finetune the model for 3 epochs using a learning rate of $2e^{-5}$ and an overall batch size of 32. We use 7B parameter model in all the experiments and its training took around 3 hours on 8 A100 40GB GPUs. During inference, for memory efficiency, we load the models in FP16 mode.

In our semi-automatic annotation framework, we use Katna (KeplerLab, 2019) to extract video keyframes. For off-the-shelf Tag2Text (Huang et al., 2023) model, we use the Swin-B variant with an input size of 384×384 and a confidence threshold of 0.7. For GRIT (Wu et al., 2022), we use ViT-B version with CenterNet2 (Zhou et al., 2021).

5.2 Quantitative evaluation

In this section, we highlight a key contribution of our work: the quantitative evaluation of Video-ChatGPT using advanced metrics and comparative evaluations with existing state-of-the-art models. We conduct two types of quantitative evaluations: i) Video-based Generative Performance Benchmarking and ii) Zero-Shot Question-Answer Evaluation.

Video-based Text Generation Performance Benchmarking:

We introduce a benchmark to evaluate the text generation performance of video-based conversation models. To do this, we curate a test set based on the ActivityNet-200 dataset (Fabian Caba Heilbron and Niebles, 2015), featuring videos with rich, dense descriptive captions and associated question-answer pairs from human annotations. We also develop an evaluation pipeline using the GPT-3.5 model. This pipeline assesses various capabilities of the model and assigns a relative score to the generated predictions

Model	MSVD-QA		MSRVTT-QA		TGIF-QA		Activity Net-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	32.2	–	16.8	–	41.0	–	24.7	–
Video Chat	56.3	2.8	45.0	2.5	34.4	2.3	26.5	2.2
LLaMA Adapter	54.9	3.1	43.8	2.7	-	-	34.2	2.7
Video LLaMA	51.6	2.5	29.6	1.8	-	-	12.4	1.1
Video-ChatGPT	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.8

Table 2: **Zeroshot question-answering** comparison of Video-ChatGPT with other video generative models. For a fair comparison, 7B variants are used for all the models. Video-ChatGPT performs competitively across all datasets.

on a scale of 1-5, in the following five aspects:

- (i) *Correctness of Information:* We verify the accuracy of the generated text, ensuring it aligns with the video content and does not misinterpret or misinform.
- (ii) *Detail Orientation:* We evaluate the depth of the model’s responses, looking for both completeness, meaning the model’s response covers all major points from the video, and specificity, denoting the inclusion of specific details rather than just generic points in the model’s response.
- (iii) *Contextual Understanding:* We assess the model’s understanding of the video’s context, checking if its responses align with the overall context of the video content.
- (iv) *Temporal Understanding:* We examine the model’s grasp of the temporal sequence of events in the video when answering questions.
- (v) *Consistency:* We evaluate the model’s consistency across different but similar questions or different sections of the video.

We present the evaluation results of our proposed model, Video-ChatGPT, using the quantitative benchmarking framework in Table 1. The results reveal its competent performance across all key aspects compared with the recently introduced contemporary video conversation models, Video Chat (Li et al., 2023b), LLaMA Adapter (Gao et al., 2023) and Video-LLaMA (Zhang et al., 2023). Video-ChatGPT shows good performance, largely due to the instruction tuning we perform and its straightforward architecture that leverages LLMs with a pretrained visual encoder fine-tuned for video data. This provides it with the robust ability to generate contextually relevant, detailed, and temporally accurate text from video input.

Zero-Shot Question-Answer Evaluation: We conducted a comprehensive quantitative evaluation using several commonly used open-ended question-answer datasets: MSRVTT-QA (Xu et al., 2017), MSVD-QA (Xu et al., 2017), TGIF-QA FrameQA (Jang et al., 2017), and ActivityNet-QA (Yu et al., 2019). These evaluations were carried out in a zero-shot manner, employing GPT-assisted evaluation to assess the model’s capabilities. This evaluation process measures the accuracy of the model’s generated predictions and assigns a relative score on a scale of 1-5.

To benchmark Video-ChatGPT, we compared its performance with other significant models, such as FrozenBiLM (Yang et al., 2022) and the generative video model, Video Chat, LLaMA Adapter and Video-LLaMA. FrozenBiLM is a model that adapts frozen bidirectional language models pretrained on Web-scale text-only data to multi-modal inputs, showing promising results in zero-shot VideoQA settings. Despite the solid foundation established by these models, Video-ChatGPT consistently outperformed them, achieving state-of-the-art (SOTA) performance across all datasets. These results indicate Video-ChatGPT’s ability to understand video content and generate accurate, contextually rich answers to questions.

5.3 Ablations

Impact of Semi-Automatic Annotations: We train Video-ChatGPT on two subsets: one with human annotations (30% of our data) and one with semi-automatic annotations (70%). The results in Table. 3 indicate that training solely with human-annotated data or semi-automatically generated data yields good performance. The overall performance when using only human-generated data is the lowest due to the limited number of labels (30% of all data) available in this scenario. However, the optimal results are achieved when utilizing a combined dataset for training.

Metric	Human only	Automatic only	Combined
Correctness	2.27	2.35	2.40
Detail Orientation	2.49	2.49	2.52
Contextual Understanding	2.50	2.56	2.62
Temporal Understanding	1.85	1.92	1.98
Consistency	2.21	2.38	2.37
Average	2.28	2.34	2.38

Table 3: **Human Annotated vs Semi-automatically Annotated Data:** Training using both human annotated and semi-automatically annotated data achieves best performance.

Quantitative Evaluation with GPT-3.5: Considering the limitations posed by the use of GPT-3.5, which is accessed via API and is not open-source, we perform evaluations using the open-source LLM, Vicuna-1.5 (13B) (Chiang et al., 2023). The results in Table. 4 show similar trend in correctness, detail, contextual and temporal understanding and consistency compared with the initial GPT-3.5 evaluation. This ensures our evaluation method remains accessible and replicable.

Metric	Video Chat	Video-LLaMA	Video-ChatGPT
Correctness	2.32	2.10	2.49
Detail Orientation	2.50	2.18	2.52
Contextual Understanding	2.76	2.41	2.85
Temporal Understanding	2.27	2.17	2.38
Consistency	2.95	2.67	3.09

Table 4: **Evaluation using Vicuna-1.5 (13B) Model:** We observe similar trend when evaluating using open-source Vicuna-1.5 (13B) model versus GPT-3.5-Turbo.

Ensuring Automatic Annotation Pipeline Consistency: To ensure consistency between our automatic evaluation pipeline and human assessments, we conducted a blind test comparing QA pairs from human and semi-automatically annotated sources using 50 randomly sampled videos. A 52% accuracy rate in distinguishing between the two demonstrated the reliability of our semi-automatic data, confirming that our quality control effectively aligns automatic evaluations with human judgment standards.

6 Conclusion

In this work, we presented Video-ChatGPT, a multimodal model that merges a pretrained visual encoder with a large language model (LLM) to enable video understanding and conversations based on videos. Video-ChatGPT leverages an adapter on top of pretrained LLM and vision backbones and is fine-tuned on video-instruction data to capture temporal dynamics and spatial consistency relationships in spatiotemporal sequences. A dataset of 100,000 video-instruction pairs is created to en-

hance Video-ChatGPT’s video-specific understanding and conversation capabilities. The work also introduced a quantitative video conversation evaluation framework for benchmarking, evaluating models on a diverse set of capabilities including conventional video question answering as well as open-ended descriptions.

7 Limitations

While the model performs competitively in several scenarios, we note it finds it challenging to understand subtle temporal relationships in long videos (> 2 min), which can compromise its predictive performance. Additionally, it has difficulty recognizing the details of small objects, often missing additional information embedded in these details.

8 Potential Risks

Video-ChatGPT, like any other AI model, must be handled with due caution to prevent misuse and to ensure it upholds the principles of fairness, transparency, and respect for user privacy.

We made a concerted effort to minimize bias during the dataset creation phase for Video-ChatGPT. Despite these efforts, it is important to recognize the possibility of residual bias persisting. The use of our model should be mindful of these potential biases, which may subtly influence the model’s understanding and response to visual content. We encourage all users to consider these limitations in their application of Video-ChatGPT and to strive for ethical and responsible use in all contexts.

9 Use of Data and AI Assistant

We curate our dataset based on a subset of the ActivityNet-200 dataset (Fabian Caba Heilbron and Nibbles, 2015), distributed under MIT LICENSE, available for use in research. Further, the use of GPT models abides by (OpenAI). Respecting source license information, we will release all datasets created in this work under MIT LICENSE.

10 Human Annotations

The semi-automatic dataset curation involves human annotation. Annotators are provided with concise video caption ground truths. Specific instructions are given to enrich the caption with comprehensive descriptions of the video content, with specific attention to temporal and spatial details. They are given specific instructions to neutralize the tone and biases during the correction process.

References

- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- 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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.
- Jean-Baptiste Alayrac et al. 2022. Flamingo: a visual language model for few-shot learning.
- Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. 2022. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2023. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*.
- Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.
- KeplerLab. 2019. Katna: Tool for automating video keyframe extraction, video compression, image autocrop and smart image resize tasks. <https://github.com/keplerlab/katna>.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.
- Kunchang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.
- Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. 2022. Class-agnostic object detection with multi-modal transformer. In *The European Conference on Computer Vision*. Springer.
- Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *The European Conference on Computer Vision*.
- OpenAI. Openai terms of use. <https://openai.com/policies/terms-of-use>.
- OpenAI. 2023. Chatgpt. Large Language Model for human style conversation <https://chat.openai.com>.
- OpenLM Lab. 2023. Moss: Codebase for moss project. An open-sourced plugin-augmented conversational language model, <https://github.com/OpenLM Lab/MOSS>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Finetuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. In *Advances in Neural Information Processing Systems*.
- David Rozenberszki, Or Litany, and Angela Dai. 2022. Language-grounded indoor 3d semantic segmentation in the wild. In *The European Conference on Computer Vision*.
- 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*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.
- Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.
- Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*.
- Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.
- Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*.
- Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.
- Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2021. Probabilistic two-stage detection. In *arXiv preprint arXiv:2103.07461*.
- 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*.

Supplemental Material

A Qualitative Evaluation

We performed an extensive evaluation of our model on a variety of open-ended video question-answering tasks, utilizing diverse videos sourced from ActivityNet and YouTube. The evaluation tasks included video reasoning (Figure 5), creative and generative tasks (see Figure 6), spatial understanding (Figure 7), action recognition (Figure 8), video conversation (Figure 9), question answering (Figure 10) and temporal understanding (Figure 11). Our model demonstrates proficiency in comprehending the content of the videos and generating accurate responses across multiple video-based tasks. Our model can effectively understand the visual information present in the videos and provide precise answers (see Figures 5 to 11).





Why is this video funny?

The video is funny because it shows a baby girl who is learning to read. She is sitting on a bed and looking at a book, and she is trying to read it. The baby girl is wearing glasses, and she is holding the book with both hands. The baby girl is trying to read the book, and her actions are adorable and endearing. The video captures the innocence and curiosity of a young child as she learns to read, which is a joyful and amusing sight for the viewer.






What is strange about this video?

The strange aspect of this video is that the lions are seen playing with a man in the wild. It is not common to see lions interacting with humans in the wild, as they are wild animals and can be unpredictable. The man is seen petting the lions and playing with them, which is quite unusual and could potentially be dangerous for both the man and the lions.





Can you identify any safety hazards in this video?

Yes, there is a safety hazard in the video. A child is playing with a toy car on the floor, and there is a fireplace nearby. The child is sitting on the floor in front of the fireplace, which could potentially be a safety hazard if the child gets too close to the fireplace. It is important to keep children away from fireplaces and other heat sources to prevent burns or other injuries.





What is unusual about this video?

The unusual aspect of this video is that it shows a giraffe performing a flip in the air on a diving board. This is not something that one would typically see in a regular diving pool.

Figure 5: **Video Reasoning Task.** This figure illustrates an example from Video-ChatGPT’s demonstration showcasing its performance in video reasoning tasks.



KEY INGREDIENTS:
 *VITAMINS B5 & E
 *ALOE
 *ESSENCE EXTRACTS

What is this video about?

The video is about a women applying sunscreen on her face and body.

Generate a creative advertisement feature about this product that can be used on the product website.

"Say goodbye to sunburn and hello to a glowing, healthy complexion with our new sunscreen product. With its lightweight formula and SPF 50+ protection, it's perfect for everyday use. Apply it daily to your face and neck for a long-lasting barrier against harmful UV rays. Don't let sun damage ruin your skin - choose our sunscreen and enjoy the benefits of a healthy, radiant complexion. Get your glow on with our sunscreen today!"



Can you write a short poem inspired from the video.

Sure, here's a short poem inspired by the video:

In a field of flowers,
 Blooms of purple and yellow,
 Their beauty shines so bright,
 As they dance in the sun.
 Their petals sway in the breeze,
 As they open-up to the sky,
 A symphony of colors and fragrance,
 A sight to make you smile.
 Their beauty is not just skin deep,
 For they hold a secret within,
 A hidden treasure of life and love,
 A gift from nature's kin.
 So let us cherish these flowers,
 For they remind us of the beauty,
 That lies within each of us,
 A reminder to embrace life.



Create a short fairy tale with a moral lesson inspired by the video.

Sure, here's a short fairy tale with a moral lesson inspired by the video:

Once upon a time, there was a beautiful princess named Cinderella who lived in a castle with her cruel stepmother and two stepsisters. She was known for her kindness and her love for animals, especially birds. One day, while she was out in the garden, she met a handsome prince who was lost in the woods. He was in need of help, and Cinderella offered to guide him back to the castle.

As they walked together, the prince noticed that Cinderella's stepsisters were cruel to her and her stepmother was always mean to her. He asked her about her life and how she managed to stay so kind and gentle despite the hardships she faced. Cinderella told him that she found happiness in the simple things in life, such as the beauty of nature and the love of animals. The prince was touched by her story and decided to help her in any way he could. He introduced her to the king and queen, who were impressed by her kindness and grace. They arranged for her to attend a ball at the castle, where she met the prince once again. They fell in love and got married, living happily ever after.

The moral of the story is that even in difficult situations, one should find happiness in the simple things in life and never lose sight of their values and kindness.

Figure 6: **Creative and generative tasks.** Illustrative examples from Video-ChatGPT's demonstration highlight its performance in video-based creative and generative tasks, such as crafting a story, poem, or advertisement.

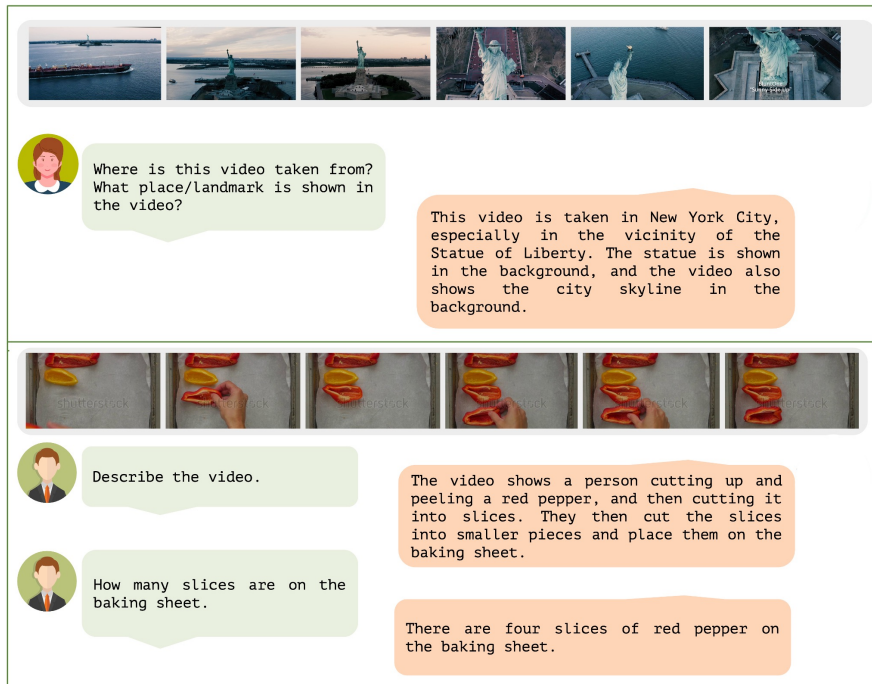


Figure 7: **Spatial understanding tasks.** The figure depicts examples from Video-ChatGPT’s demonstration, emphasizing its capability in video-based spatial understanding tasks, including identifying renowned locations or counting the number of objects in a scene.

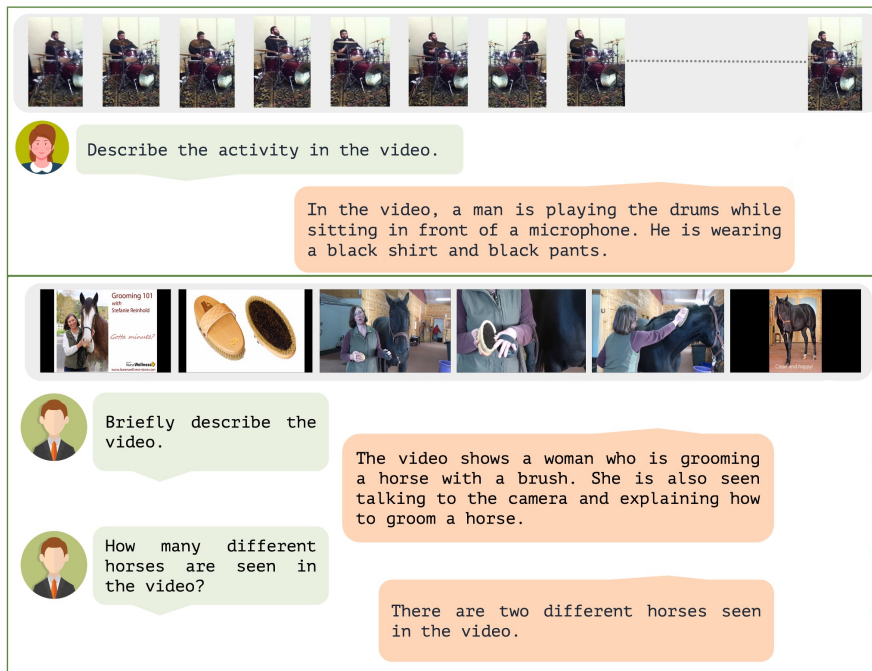


Figure 8: **Action Recognition Task.** This figure illustrates examples from Video-ChatGPT’s demonstration showcasing its performance in video action recognition tasks such as playing drums and grooming horse.



 Describe the video in detail.

The video shows a man cleaning snow off his car using a snow brush. He starts by brushing the snow off the hood of his car, then moves on to the windshield and finally the side mirrors. He uses a snow brush to remove the snow from the car's surface.



 What is the main focus of the video.

The main focus of the video is the process of drawing a tree on a white canvas. The artist starts by drawing a trunk, then adds branches and leaves.



 What does the woman use to split the logs and how does she do it?

The woman uses a large axe to split the logs. She starts by placing the axe on the log and then swings it down with a powerful motion. The axe slices through the log with ease, and the woman continues to swing it until the log is split in half.

 What is the women wearing in the video?

The woman is wearing a black jacket in the video.



 Describe the video.

The video shows a young girl performing a gymnastics routine on a balance beam. She starts by walking onto the beam and then performs a series of flips and spins, ending with a handstand. She then walks off the beam and smiles at the camera.

 Is there a flag in the background?

Yes, there is an American flag in the background of the video.

Figure 9: **Video Understanding and Conversation Tasks.** This figure illustrates examples from Video-ChatGPT's demonstration showcasing its performance in video understanding and conversation tasks.





What is the video about?

The video shows a man in a red shirt throwing a discus in a field. He is seen throwing the discus and then walking away from it.



Can you describe the discus thrower's technique in the video?

The discus thrower in the video demonstrates a powerful and smooth technique. He starts by holding the discus in his right hand and then releases it with a flick of his wrist. As the discus flies through the air, the thrower follows through with his arm, and his body rotates in a circular motion. The thrower's technique is characterized by a combination of strength, speed, and precision, which allows him to achieve a long and accurate throw.





What is the main challenge faced by the people on the boat?

The main challenge faced by the people on the boat is the rough waves and strong winds. They are trying to maintain their balance and control the boat while navigating through the rough waters.





What are the people doing in the video?

The people in the video are playing in the snow and having fun. They are also seen building a snowman and playing with a sled.





What activities are the woman and the dog engaged in the video?

The woman and the dog are playing fetch in the yard. The woman throws a ball and the dog runs to fetch it. The woman then throws the ball again and the dog runs to fetch it.

Figure 10: **Question-Answering Task.** The figure depicts examples Video-ChatGPT's demonstration showcasing its performance in question-answering tasks.





Describe the video in detail.

The video shows a man standing on a snowy field, throwing a javelin. He throws the javelin and then runs towards it, picking it up and throwing it again. He then runs towards the javelin and picks it up, throwing it once more. The man then runs towards the javelin and picks it up, throwing it for the fourth time. The video ends with the man running towards the javelin and picking it up, throwing it for the fifth time.





What is the video about?

The video is about two men who are playing a game of arm wrestling. The video shows the two men trying to push each other's hands down on the table.



What happens after the match is over?

After the match is over, the two men shake hands, showing good sportsmanship and camaraderie.





Can you describe the process for cleaning a screen window as demonstrated in the video?

The video shows a man using a squeegee to clean a window. He starts by wetting the squeegee with water and then rubbing it on the window. He then moves the squeegee across the window, removing any dirt or debris. After cleaning the window, he wipes it dry with a cloth. The man then repeats the process for the other side of the window.





Step-by-step describe the cooking recipe shown in the video.

The video shows a person cooking a dish in a pan. They add ingredients such as noodles, and spices to the pan, and then stir the mixture. The person then pours a sauce into the pan and stirs it until it thickens. Finally, they add some more ingredients and stir the mixture again. The dish is then served on a plate.

Figure 11: **Temporal Understanding Task.** The figure provides examples from Video-ChatGPT's demonstration, highlighting its performance in temporal understanding tasks, particularly in comprehending sequences of events.