PG-Video-LLaVA: Pixel Grounding Large Video-Language Models

Anonymous ACL submission

Abstract

Extending image-based Large Multimodal Models (LMMs) to videos is challenging due to the inherent complexity of video data. The recent approaches extending image-based LMMs to videos either lack the grounding capabilities (e.g., VideoChat, Video-ChatGPT, Video-LLaMA) or do not utilize the audio-signals for better video understanding (e.g., Video-ChatGPT). Addressing these gaps, we propose PG-Video-LLaVA, the first LMM with pixel-level grounding capability, integrating audio cues by transcribing them into text to enrich video-context understanding. Our framework uses an off-the-shelf tracker and a novel grounding module, enabling it to spatially localize objects in videos following user instructions. We evaluate PG-Video-LLaVA using video-based generative and question-answering benchmarks and introduce new benchmarks specifically designed to measure prompt-based object grounding performance in videos. Further, we propose using open-source Vicuna LLM for video-based conversation benchmarking, as opposed to GPT-3.5 utilized in Video-ChatGPT, ensuring reproducibility of results which is a concern with the proprietary nature of GPT-3.5. Our framework builds on SoTA image-based LLaVA model and extends its advantages to the video domain, delivering promising gains on video-based conversation and grounding tasks. Our codes, pretrained models, and interactive demos will be made publicly available.

1 Introduction

Recent efforts on Large Multimodal Models (LMMs), spearheaded by GPT-4V (OpenAI, 2023b), allow detailed conversations about images but generally do not scale well to videos. The magnitude of video data scales far beyond other modalities due to its massive volume on social and internet media. Furthermore, extending LMMs to videos is challenging due to their complex dynamics with long temporal context that needs to be understood accurately. Although recent approaches towards video-LMMs such as VideoChat (Li et al., 2023b), Video-LLaMA (Zhang et al., 2023a), and Video-ChatGPT (Maaz et al., 2023) have demonstrated capabilities in video comprehension and dialogue, they lack the crucial feature of visual grounding. Visual grounding in videos aims to associate the LMM responses to specific objects within the video input. Addressing this gap, we introduce PG-Video-LLaVA, the first video-LMM capable of localizing objects appearing in LMM responses. This task leads to enhanced intractability and demonstrates deep understanding of video content.

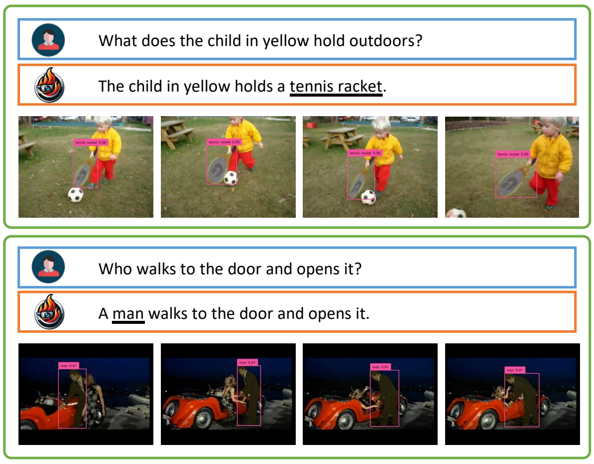

Figure 1: **Video spatial grounding** on example videos from Vid-STG (Zhang et al., 2020) (above) and HC-STVG (Tang et al., 2021) (below) datasets. PG-Video-LLaVA can generate textual responses with referred objects grounded in the video content (*tennis racket* and *man* are localized in the top and bottom examples, respectively).

In PG-Video-LLaVA, we address the unique challenges posed by video data. The model is designed to track objects within shorter video clips that maintain consistent camera views, enabling accurate visual grounding across scenes and motions. This tracking links spatio-temporal segments directly to conversational elements, enhancing the model's contextual understanding. A salient feature of PG-Video-LLaVA is its modular design, allowing for easy integration with existing grounding modules and the flexibility to adapt to future enhancements in visual grounding technology. Moreover, PG-Video-LLaVA enriches its capabilities by incorporating audio context. It achieves this by leveraging video audio in a form understandable to LLM, which is particularly useful in situations where the auditory information is essential to the conversation. This inclusion broadens the model's understanding, making it more versatile in interpreting video content.

Furthermore, this work introduces an improved framework for benchmarking video-based conversational models, pivoting from previous approaches (Maaz et al., 2023) that predominantly used the proprietary GPT-3.5-Turbo model for evaluation. Given that GPT-3.5-Turbo is subject to changes at any time and lacks transparency due to its closed-source nature, it presents challenges in terms of reliability and reproducibility. To address this, we propose the use of Vicuna, an open-source LLM for benchmarking. This shift not only enhances reproducibility but also improves transparency in the evaluation process. We evaluate PG-Video-LLaVA using our improved benchmarks and show notable improvements over existing video conversational models like Video-ChatGPT (Maaz et al., 2023) and Video-LLaMA (Zhang et al., 2023a) in ungrounded dialogues, achieving state-of-the-art (SoTA) performance.

The key contributions of this work are:

- We propose PG-Video-LLaVA, the first video-based LMM with pixel-level grounding capabilities, featuring a modular design for enhanced flexibility.
- By incorporating audio context, PG-Video-LLaVA significantly enhances its understanding of video content, making it more comprehensive and aptly suited for scenarios where the audio signal is crucial for video understanding (e.g., dialogues and conversations, news videos, etc.).
- We introduce improved quantitative benchmarks for video-based conversational models. Our benchmarks utilize open-source Vicuna LLM to ensure better reproducibility and transparency. We also propose benchmarks to evaluate the grounding capabilities of video-based conversational models.

2 Related Work

Recent advancements in Large Multimodal Models (LMMs) (Liu et al., 2023a; Zhu et al., 2023; Dai et al., 2023) and Large Language Models (LLMs) (Chiang et al., 2023; OpenAI, 2023a; Touvron et al., 2023) have significantly transformed the artificial intelligence landscape, particularly in natural language processing and multimodal tasks. These breakthroughs have enhanced machine learning models' ability to understand and generate human-like text, while also enabling more effective integration of various data types like images, sounds and videos with textual information. This progress represents a major leap in creating AI systems that can accurately interpret and interact with a diverse range of content.

Large Language Models (LLMs): The natural language processing (NLP) field has undergone a revolution with the advent of LLMs such as GPT (Brown et al., 2020), LLaMA (Touvron et al., 2023), OPT (Zhang et al., 2022), and MOSS (OpenLMLab, 2023), particularly noted for their zero-shot learning abilities and adaptability. The development of models like InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2023a) has further propelled advancements in conversational AI and complex query handling, chiefly through instruction tuning. Within the LLaMA framework, the emergence of open-source models such as Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023) exemplifies how instruction tuning can significantly boost model performance. This shift towards open-source initiatives in language modeling, highlighted by models like Alpaca and Vicuna, indicates a growing trend towards more accessible and collaborative approaches in the field. In this work, we build on the open-source Vicuna LLM and extend it with multimodal capabilities. We also propose an open-source benchmark for video conversation and reasoning tasks using Vicuna LLM that is reproducible for fair evaluations.

Large Multimodal Models (LMMs): The field of AI has witnessed significant advancements with the development of vision-language models like CLIP (Radford et al., 2021), renowned for their impressive zero-shot capabilities using extensive image-text pairs during training. These models have proven effective in a variety of applications, from image detection and segmentation (Liang et al., 2023; Bangalath et al., 2022) to more complex tasks such as 3D modeling and video analysis (Rozenberszki et al., 2022; Ni et al., 2022; Wang et al., 2021; Rasheed et al., 2023a). The introduction of BLIP-2 marked a pivotal transition, pioneering the integration of image features encoded by a visual encoder with text embeddings, setting the stage for the evolution into Large Multimodal Models (LMMs). This advancement influenced subsequent models like LLaVA (Liu et al., 2023b), InstructBLIP (Dai et al., 2023), and MiniGPT-4 (Zhu et al., 2023), which further refined image-text feature alignment and instruction tuning. VideoChat (Li et al., 2023b), Video-ChatGPT (Maaz et al., 2023) and Video-LLaMA (Zhang et al., 2023a) represents an extension of these LMMs, moving from image-based to video-based applications, while models such as Otter (Li et al., 2023a), mPLUG-Owl (Ye et al., 2023), LLaMa-Adapter (Gao et al., 2023), and InternGPT (Liu et al., 2023d) continue to push the boundaries of multimodal interaction. Despite these significant strides, challenges in achieving robust visual grounding in LMMs highlight key areas for ongoing research and development in this dynamic field. Further, effective integration of audio signals within LMMs for comprehensive video understanding is an open research question that this work aims to address.

Visual-Language Grounding: Grounded Large Language Models (LLMs) have made notable progress in enhancing visual and language comprehension. A diverse array of models including Kosmos-2 (Peng et al., 2023), Ferret (You et al., 2023), All-Seeing Model (Wang et al., 2023), LISA (Lai et al., 2023), BuboGPT (Zhao et al., 2023), Shikra (Chen et al., 2023), and GLaMM (Rasheed et al., 2023b) have employed various methodologies to master complex grounding tasks. These models demonstrate proficiency in tasks like referring expression comprehension and image segmentation, showcasing the advanced image understanding capabilities of LLMs. Methodologically, Kosmos-2, Shikra, and All-Seeing focus predominantly on creating language-based context for visual grounding. In contrast, BuboGPT merges visual elements with language, and LISA leverages vision-language embeddings for producing segmentation masks. Furthermore, GLaMM is adept at generating natural language responses linked with object segmentation masks, facilitating detailed visual-textual interactions. However, challenges remain, such as LISA's constrained performance in multi-object scenarios and the limitations of BuboGPT and GLaMM to image-based applications, not extending to video processing. To this end, we introduce PG-Video-LLaVA, a video conversational model with pixel-level grounding capability. Further, PG-Video-LLaVA incorporates audio transcripts alongside visual and textual data, aiming to provide a more detailed understanding of video content.

3 PG-Video-LLaVA

3.1 Overview

In this paper, we introduce PG-Video-LLaVA, a novel Large Multimodal Model (LMM) designed to align video and audio representations with a Large Language Model (LLM) giving the capability to proficiently manage both video and audio data in conversational contexts. Additionally, our method integrates a specialized plug-and-play module for effective video grounding (see Figure 2). While PG-Video-LLaVA's foundation is based on the LLaVA-1.5 (Liu et al., 2023a) framework, its unique combination of enhanced video encoding, extensive training dataset, integrated audio processing and grounding capability marks it as a forward step in the field of LMMs.

Central to our model is an advanced CLIP-based video encoder, which has been adapted to process both spatial and temporal dimensions of video data. This adaptation enables a deeper understanding of video content, setting PG-Video-LLaVA apart from conventional image-centric models. In addition, PG-Video-LLaVA leverages audio transcription and filteraton techniques, inspired from WhisperX (Bain et al., 2023) and Whisper-AT(Gong et al., 2023) allowing the model to process and understand audio inputs effectively, enhancing its overall multimodal interpretation capabilities.

3.2 Architecture

In PG-Video-LLaVA, the spatio-temporal feature extraction is inspired by Video-ChatGPT (Maaz et al., 2023). Our architecture utilizes the CLIP ViT-L/14@336 as the visual encoder, which has been adapted for video processing. Given a video input $V_i \in \mathbb{R}^{T\times H\times W\times C}$, where T denotes the frame count, the encoder processes each of the T frames independently, treating them as a series of images. This leads to the generation of frame-level

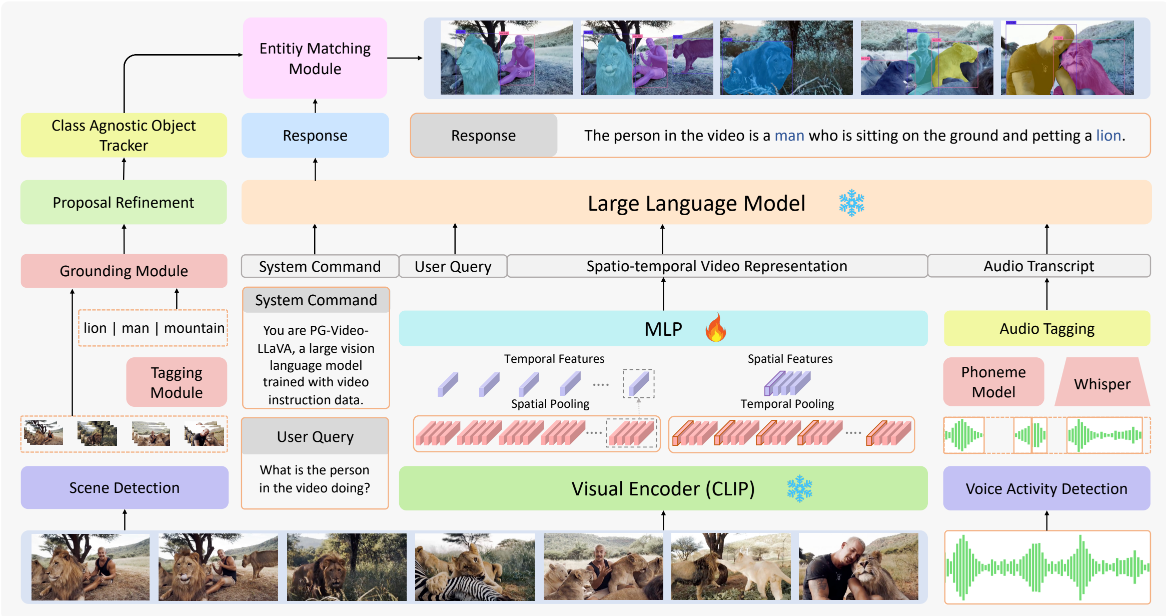

Figure 2: **Architecture of PG-Video-LLaVA:** PG-Video-LLaVA integrates a CLIP-based visual encoder with a multimodal language model for video understanding. The CLIP visual encoder extracts spatio-temporal features from videos by averaging frame-level features across temporal and spatial dimensions. These features are then projected into the LLM's input space using a learnable Multi-Layer Perceptron (MLP). The system features a grounding module for spatially locating textual descriptions within video frames, a class-agnostic object tracker, and an entity-matching module. Audio processing incorporates voice activity detection, phoneme modeling, and Whisper-based audio transcription, resulting in a multimodal pipeline that facilitates robust video-question answering. The architecture is trained on a video instruction dataset, enabling the handling of diverse conversational contexts with high accuracy.

embeddings $x_i \in \mathbb{R}^{T\times h\times w\times D}$, where $h = H/p$ and $w = W/p$, with p being the patch size (14 for ViT-L/14) and $N = h \times w$ indicating the total token count.

To construct a comprehensive video-level representation, we apply average pooling across the temporal dimension of these frame-level embeddings, resulting in a video-spatial representation $v_i \in \mathbb{R}^{N\times D}$. This temporal pooling technique effectively amalgamates information across multiple frames. Similarly, for explicit temporal information, we achieve temporal representation $t_i \in \mathbb{R}^{T\times D}$ through average pooling along the spatial dimension. The final video-level features v_i are a combination of these temporal and spatial features, as shown in the equation:

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

These video-level features are projected into the embedding space of the language decoder using, a learnable Multi-Layer Perceptron (MLP), designated as g, to serve as our cross-modal connector. This is inspired by LLaVA-1.5 (Liu et al., 2023a), and it aims to enhance the model's performance compared to using a simple linear projection as in Video-ChatGPT (Maaz et al., 2023). The process yields language embedding tokens Q_v.

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

Text queries, denoted as $Q_t \in \mathbb{R}^{L\times K}$ where L is the length of the query, are tokenized to be dimensionally compatible with these video embeddings. The combination of Q_v and Q_t is then fed into the language decoder, facilitating the seamless integration of video and textual data within the model (see Figure 2).

3.2.1 Audio Modality Integration

In PG-Video-LLaVA, we have integrated an audio processing pipeline that significantly enhances the video-question answering capabilities by incorporating audio cues from the input, drawing inspiration from the architecture of WhisperX (Bain et al., 2023). The process begins with the deployment of a Voice Activity Detection (VAD) model. This model is crucial for pinpointing speech-containing temporal segments within the audio track. Following the VAD's identification of speech segments, these segments undergo processing—cutting, merging, and padding—to align with the input specifications of the Whisper model (OpenAI, 2022). Simultaneously, a phoneme segmentation model operates in parallel, producing phone-level segmentations essential for the subsequent alignment of raw transcriptions with the audio.

The VAD model serves a dual purpose: 1) identifying speech segments and 2) aiding in filtering out non-speech audio components. To enhance the compatibility of transcriptions generated by Whisper with our model, we integrate Whisper-AT (Gong et al., 2023). This advanced version of the Whisper model specializes in audio tagging. It annotates the audio stream with labels from an extensive set of 527 audio event classes, allowing for precise temporal resolution.

The audio transcripts are then subjected to a multi-stage filtering process. Initially, a VAD-based filter is applied, followed by a phoneme-based forced alignment using the Whisper model, ensuring temporally accurate text transcriptions. Utilizing Whisper's language identification feature, we eliminate non-English speech segments at this stage. For each identified sentence segment, we apply Whisper-AT (Gong et al., 2023) for audio tagging, focusing on the top three predicted audio classes. Segments that do not predominantly feature 'speech', or where 'music' probabilities significantly exceed 'speech', are excluded from further processing.

Finally, the integration of the audio transcript with the video component is executed through a carefully designed prompt template (Appendix-A). This template is pivotal in guiding the system to understand user instructions, assimilate the video frames, and incorporate the transcriptions generated by the automatic speech recognition model. This structured approach ensures that PG-Video-LLaVA efficiently leverages all available modalities—visual and auditory—thereby enabling users to achieve task completion and query resolution based on a comprehensive analysis of both visual and auditory content (refer to Figure 2 for details).

3.2.2 Grounding Module

In PG-Video-LLaVA, our visual grounding approach starts with processing video-question pairs to generate textual descriptions. These descriptions are then used for grounding within the video frames. Key noun phrases are extracted from the generated text using Vicuna, targeting the most critical content aspects. Simultaneously, an image tagging model, RAM (Zhang et al., 2023b), tags visual elements in each frame, creating a detailed map of the video content.

The video is segmented into smaller parts using PySceneDetect (Castellano, 2023), based on changes in scene composition. This segmentation facilitates a more focused grounding process. In each segment, our grounding ensemble, composed of GroundingDINO (Liu et al., 2023c), DEVA (Cheng et al., 2023), and SAM (Kirillov et al., 2023), utilizes the image tags to create segmentation masks and tracking IDs for the identified visual elements.

The visual cues from these segmentation masks are then matched with the textual noun phrases using CLIP (Radford et al., 2021). This matching process links text to the corresponding visual elements in the video, enhancing our understanding of the content.

In quantitative analysis (Section 4.4), from the descriptive textual response to an interrogative text, a referring expression or a phrase is extracted using Vicuna. This phrase is input into our grounding module, which then generates segmentation masks and tracking IDs. We measure the spatial grounding accuracy of our model by calculating the Intersection over Union (IoU) between these segmentation masks and ground truth bounding boxes.

4 Experiments

4.1 Implementation Details

We build our strong baseline on top of LLaVA-1.5 which utilizes CLIP ViT-L/14@336 as the image encoder and Vicuna v1.5 as the LLM. We only tune the MLP projection layers during training with the VideoInstruct100K (Maaz et al., 2023) dataset, while keeping the rest of the architecture frozen. We finetune the model for 3 epochs using a learning rate of $2e^{-5}$ and an overall batch size of 32. The training of our 7B and 13B models took around 6 hours and 15 hours respectively on 4 A100 80GB GPUs.

For audio transcript extraction, Whisper-base model is used. Our grounding module is based on GroundingDINO-T variant and CLIP ViT-B/32. For the image-tagging model we use RAM Swin-Large variant (with input size 384). DEVA Tracker is applied under online-setting in our experiments.

Vicuna-13b-v1.5 model is used in performing

Model	Evaluation Metrics				
	Correctness	Detail Orientation	Contextual Understanding	Temporal Understanding	Consistency
LLaMA Adapter (Gao et al., 2023)	2.34 ± 0.03	2.44 ± 0.01	2.67 ± 0.02	2.25 ± 0.04	3.03 ± 0.04
Video Chat (Li et al., 2023b)	2.48 ± 0.02	2.81 ± 0.01	2.92 ± 0.00	2.29 ± 0.02	3.10 ± 0.02
Video-LLaMA (Zhang et al., 2023a)	2.29 ± 0.00	2.59 ± 0.02	2.68 ± 0.01	2.23 ± 0.06	2.88 ± 0.02
Video-ChatGPT (Maaz et al., 2023) (w/o audio)	2.49 ± 0.02	2.52 ± 0.03	2.85 ± 0.00	2.38 ± 0.09	3.09 ± 0.01
Video-ChatGPT (Maaz et al., 2023) (with audio)	2.63 ± 0.00	2.71 ± 0.00	2.99 ± 0.01	2.51 ± 0.05	3.23 ± 0.05
PG-Video-LLaVA (7B) (w/o audio)	2.69 ± 0.02	2.80 ± 0.02	3.10 ± 0.01	2.44 ± 0.03	3.39 ± 0.01
PG-Video-LLaVA (7B) (with audio)	2.75 ± 0.01	2.89 ± 0.00	3.16 ± 0.03	2.53 ± 0.08	3.47 ± 0.07
PG-Video-LLaVA (13B) (w/o audio)	2.80 ± 0.03	2.92 ± 0.01	3.22 ± 0.01	2.53 ± 0.03	3.44 ± 0.02
PG-Video-LLaVA (13B) (with audio)	$\mathbf{2.84 \pm 0.02}$	$\mathbf{2.97 \pm 0.01}$	$\mathbf{3.22 \pm 0.01}$	$\mathbf{2.54 \pm 0.02}$	$\mathbf{3.56 \pm 0.06}$

Table 1: **Performance benchmarking of video-based conversational models** using the benchmarking framework from Video-ChatGPT (Maaz et al., 2023) with Vicuna-13b-v1.5 (Chiang et al., 2023) as the evaluator model. Results indicate that PG-Video-LLaVA achieves favourable performance across all metrics.

video-based conversational benchmarking, zero-shot question answering evaluation, and extracting the key noun or referring expression from the model output in the quantitative evaluation of the spatial grounding task. Further, Vicuna-13b-v1.5 was used to implement the entity matching as in (Zhao et al., 2023).

4.2 Stronger Baseline

To evaluate the impact of the enhanced baseline on PG-Video-LLaVA, we apply the benchmarking framework from Video-ChatGPT (Maaz et al., 2023). This framework measures performance on several axes critical for video-based conversational agents, including correctness of information, detail orientation, contextual understanding, temporal understanding, and consistency.

In order to facilitate a reliable and reproducible evaluation, we have modified the assessment pipeline introduced in Video-ChatGPT by replacing GPT-3.5-Turbo with open-source Vicuna-13b-v1.5. This adjustment addresses the limitations in reproducibility inherent to the closed-source nature of GPT-3.5-Turbo. Subsequently, we have re-assessed both PG-Video-LLaVA and other recent models to ensure a fair and consistent comparison. Each experiment is performed 3 times and the mean and standard deviation are reported in Table 1. The results indicate that PG-Video-LLaVA outperforms the foundational Video-ChatGPT model and exhibits superior performance when compared to other recent contributions in the domain. (See Appendix-E for qualitative results.)

4.3 Effect of Audio Modality

Table 1 shows that adding the audio modality helps to improve the performance metrics. Further, in Figure 4 it can be observed that the model which takes audio transcript produces correct outputs, whereas the model without audio modality fails to capture those details from visual content alone.

Model	VidSTG	HC-STVG
Grounding DINO (Liu et al., 2023c)	25.3	19.5
Video-LLaMA (Zhang et al., 2023a)	28.6	26.1
Video-ChatGPT (Maaz et al., 2023)	32.8	20.8
PG-Video-LLaVA (7B)	34.2	**28.3**
PG-Video-LLaVA (13B)	**35.1**	27.3

Table 2: **Performance of PG-Video-LLaVA and other models on spatial grounding task**: Evaluated using the VidSTG and HC-STVG benchmarks, the results demonstrate PG-Video-LLaVA's favorable spatial grounding capabilities, as evidenced by its ability to generate accurate descriptive responses and effectively locate referring expressions within video frames. The table shows the model's progress, particularly in the 13B version, showcasing its performance among other SoTA video-conversational models.

4.4 Spatial Grounding in Videos

To quantitatively assess PG-Video-LLaVA's spatial grounding capability, we conducted quantitative evaluations of PG-Video-LLaVA's spatial grounding capabilities using two benchmarks that are derived from the test set of the VidSTG (Zhang et al., 2020) and HC-STVG (Tang et al., 2021) datasets. Due to the novelty of integrating spatial grounding within video-conversational models, we highlight the modular nature of our grounding pipeline, which can be incorporated with other state-of-the-art video conversation models. For the VidSTG dataset, we selectively processed interrogative prompts to assess the grounding accuracy. The model generates descriptive textual responses to these prompts, from which Vicuna-13b-v1.5 extracts relevant referring expressions. These expressions are then spatially grounded in the video frames using our grounding pipeline. For the HC-STVG dataset, interrogative prompts are first mined from the text captions using Vi-

Figure 3: **Qualitative Results for Video Grounding**: Visual representation of the grounding capability of advanced video-conversational capabilities of PG-Video-LLaVA. The highlighted regions in each video frame indicate the model's ability to identify and spatially locate key subjects mentioned in the textual description, such as the giraffe, the statue, and the gymnast on a balance beam.

Model	**MSVD-QA** (Xu et al., 2017)		**MSRVTT-QA** (Xu et al., 2016)		**TGIF-QA** (Li et al., 2016)		**Activity Net-QA** (Yu et al., 2019)	
	Accuracy	**Score**	**Accuracy**	**Score**	**Accuracy**	**Score**	**Accuracy**	**Score**
FrozenBiLM (Yang et al., 2022)	32.2	–	16.8	–	41.0	–	24.7	–
LLaMA Adapter (Gao et al., 2023)	53.7	3.3	45.6	3.2	54.3	3.3	37.3	3.2
Video LLaMA (Zhang et al., 2023a)	48.6	3.2	32.8	2.8	51.4	3.4	27.1	2.9
Video-ChatGPT (Maaz et al., 2023)	62.6	3.6	50.0	3.3	66.5	3.7	**40.8**	**3.3**
PG-Video-LLaVA	**64.1**	**3.7**	**51.6**	**3.3**	**66.8**	**3.8**	39.9	**3.3**

Table 3: **Zeroshot video-based question-answering:** Comparison of PG-Video-LLaVA with other video generative models. The latest available models are used for all the approaches and the benchmarks are calculated using open-source Vicuna LLM. PG-Video-LLaVA performs better than the previously proposed video-based conversational methods.

cuna and then used similarly to VidSTG prompts. (Appendix-B)

The results shown in Table 2 position PG-Video-LLaVA alongside alternative methods using the same benchmarks, demonstrating our model's enhanced ability to accurately answer questions, thereby leading to improved spatial grounding performance.

The qualitative results shown in Figure 3 emphasize the model's refined spatial grounding precision. The accurate overlay of masks on the subjects within the videos confirms the model's adeptness at correlating textual descriptors with visual elements, a critical aspect of contextual comprehension. This refined ability is crucial for applications that integrate visual data with language, improving the model's utility in environments that demand rich, interactive visual and linguistic processing.

4.5 Zero-Shot Visual Question Answering

For PG-Video-LLaVA, zero-shot question-answering (QA) capabilities were evaluated quantitatively using several established open-ended QA datasets: MSRVTT-QA (Xu et al., 2016), MSVD-QA (Xu et al., 2017), TGIF-QA (Li et al., 2016), and ActivityNet-QA (Yu et al., 2019). These datasets are benchmarks for assessing a model's ability to generate accurate answers without any dataset-specific fine-tuning. We adopted the zero-shot evaluation methodology introduced in Video-ChatGPT (Maaz et al., 2023), with Vicuna-13b-v1.5 as the evaluator model to assess the model's understanding and predictive

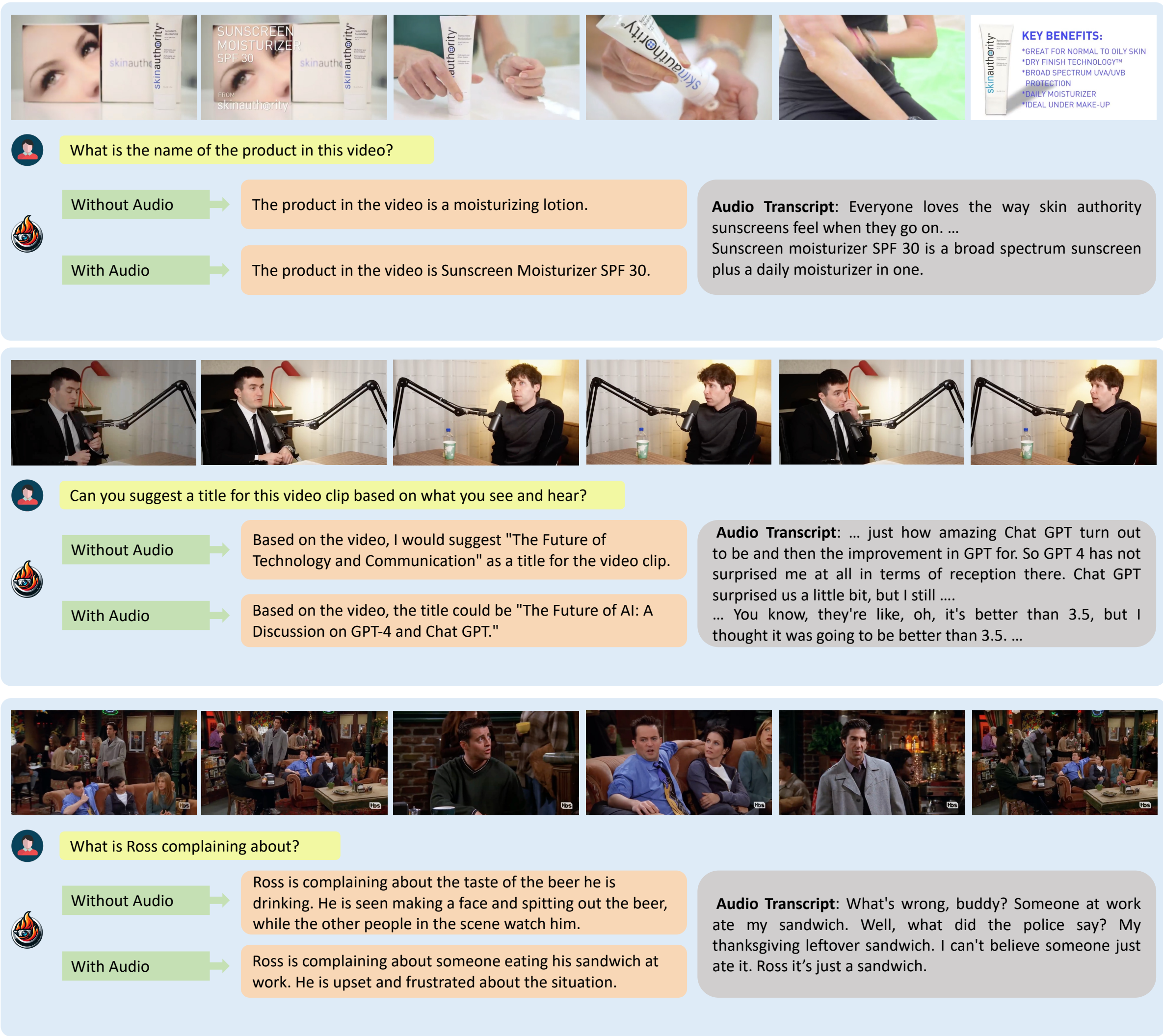

Figure 4: **Qualitative Results for Including Audio Modality**: The figure illustrates the integrated audio processing pipeline that augments video-question answering with audio cues. It provides side-by-side comparisons showing how audio cues offer additional context, leading to a more accurate interpretation of the video content.

accuracy, with scores assigned on a scale from 1 to 5. The results are presented in Table 3.

In comparison to Video-ChatGPT, PG-Video-LLaVA demonstrates superior performance, surpassing not only the predecessor but also other notable models in the field, such as Frozen-BiLM (Yang et al., 2022) and Video Chat (Li et al., 2023b). The results from our evaluations indicate that PG-Video-LLaVA has significantly enhanced its ability to comprehend video content and generate contextually relevant answers, thus establishing a new state-of-the-art in zero-shot VideoQA.

5 Conclusion

In this work, we introduced PG-Video-LLaVA, a novel video-based conversational model equipped with pixel-level grounding capabilities. PG-Video-LLaVA enhances image-based conversational models by extracting spatio-temporal features essential for comprehensive video understanding. It incorporates filtered audio transcripts to enrich the interpretation of visual scenes where audio cues are pivotal. Additionally, we developed a novel grounding module capable of tracking and generating pixel-level grounding of objects within videos. To promote reproducibility, we propose quantitative benchmarks for video-based conversational models, utilizing the open-sourced Vicuna LLM instead of GPT-3.5, as employed by previous approaches. These benchmarks are specifically designed to evaluate grounding capabilities. In summary, this work represents the first effort to integrate grounding capabilities into video-based LMMs.

6 Limitations

Though we present a novel large multimodal model for video understanding, with unprecedented capabilities in multimodal fusion and visual grounding, we would like to acknowledge some of the limitations it encompasses, which points to open research directions. Especially, the adaptability of the proposed model's video understanding capabilities for extremely varied or uncommon real-world scenarios remains untested. Due to the inherent complexity of understanding long and diverse video content, and the lack of high-quality diverse human-annotated training data, the performance gains of our work demonstrated on standard video understanding datasets, might not always generalize into special cases such as egocentric videos. Though the proposed architecture sets the baseline for conversational grounding in videos and serves as a proof-of-concept, its capability derives mainly from the clever amalgamation of large pretrained foundational models. Embedding these abilities into the large multimodal model remains an open research problem and will be addressed in future work.

7 Ethical Considerations

The key potential risk of our work being misused lies in the possibility of fake textual content generation based on video prompts. This adds to the already existing risk associated with large language models which are prone to be exploited by users with malicious intent to generate articles that appear as if generated by a human.

In this work, we utilize multiple open-source source code repositories, models, and datasets intended and licensed for research use only. They are also restricted to use cases that follow the license agreement of CLIP, LLaMA, Vicuna and GPT-4. Our work will be made publically available subject to a non-commercial license, and it should not be used outside of research purposes.

8 Use of AI Assistants

We acknowledge that LLMs were used as AI assistants in benchmarking conversational performance, spatial grounding, and zero-shot video-based question-answering as mentioned in Section 4. Further, the VideoInstruct100K dataset used to train our model contains AI-generated text, which resulted from the semi-automatic annotation involving ChatGPT/GPT-3.5.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, , Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Brandon Castellano. 2023. Pyscenedetect: Video scene cut detection and analysis tool. https://github.com/Breakthrough/PySceneDetect.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. 2023. Tracking anything with decoupled video segmentation. In *ICCV*.

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. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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. *arXiv preprint arXiv:2305.06500*.

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:2304.15010*.

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-at: Noise-robust automatic speech recognizers are also strong audio event taggers. In *Proc. Interspeech 2023*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

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*.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv:2305.06355*.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A New Dataset and Benchmark on Animated GIF Description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*, pages 7061–7070.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*.

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. 2023d. Interntchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*.

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 *European Conference on Computer Vision*, pages 1–18. Springer.

OpenAI. 2022. Whisper. https://openai.com/research/whisper.

OpenAI. 2023a. Chatgpt: Large language model for human-style conversation. https://chat.openai.com.

OpenAI. 2023b. Gpt-4v(ision) system card. https://openai.com/research/gpt-4v-system-card.

OpenLMLab. 2023. MOSS: Codebase for MOSS Project. https://github.com/OpenLMLab/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.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

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.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023a. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. 2023b. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*.

David Rozenberszki, Or Litany, and Angela Dai. 2022. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*.

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.

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*.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. 2023. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*.

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 *ACM Multimedia*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

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 *AAAI*, pages 9127–9134.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *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*.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2023b. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where does it exist: Spatio-temporal video grounding for multi-form sentences.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*.

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*.

Appendix

A Audio Modality Integration

Here, we outline the implementation details of audio modality integration in PG-Video-LLaVA.

A.1 Audio Transcript Filtering

To generate audio transcripts, we first experimented with using the state-of-the-art Whisper (OpenAI, 2022) directly. However, the obtained transcripts were too noisy, contained hallucinations, and unwanted text such as lyrics from songs. Passing these raw audio transcripts directly to the LLM without any filtering can negatively affect the overall model's performance. Therefore, a preprocessing method is required to filter out noisy text and keep only the parts of the audio that carry meaningful information.

The following steps combining WhisperX(Bain et al., 2023) and Whisper-AT(Gong et al., 2023) are used to refine the original Whisper transcripts to be usable as inputs to the video LMM.

1. We first apply VAD-based preliminary filtering to the audio, and then use the Whisper model with Phoneme-based forced alignment to get temporally aligned text transcriptions.

2. As Whisper is able to identify the language spoken, all non-English speech can be ignored at this point since PG-Video-LLaVA generates responses in English.

3. For each sentence segment obtained, slice the original audio at the corresponding timestamps and pass to Whisper-AT to produce audio-tagging output.

4. For each sentence segment, consider the top 3 audio classes predicted.
    (a) If "speech" is not among the top 3 predictions, the segment is ignored.
    (b) If $P[\text{music}] > P[\text{speech}]$ and $P[\text{music}] - P[\text{speech}] > threshold$, the segment is ignored (the $threshold$ is set empirically to 1.1).

Figure 6 shows the effectiveness of our audio transcript preprocessing method in filtering out hallucinations, music, and garbage characters from the raw audio transcript.

A.2 Integrating Audio Transcript into the LLM

The following prompt template is used when combining the spatiotemporal video features and audio transcript with the user instruction text.

```
SYSTEM:
    You are PG-Video-LLaVA,
    a large vision-language
    assistant.
    You are able to understand
    the video content that the
    user provides, and assist
    the user with a variety
    of tasks using natural
    language.

USER:
    <Instruction>
    <Video-Tokens>
    The noisy audio transcript
    of this video is:
    <Audio-Transcript>

ASSISTANT:
```

B Visual Grounding: Quantitative Evaluation

B.1 Overview

We introduce novel benchmarks for quantitatively evaluating conversation-based video spatial grounding, based on two existing spatio-temporal video grounding datasets, VidSTG(Zhang et al., 2020) and HC-STVG(Tang et al., 2021).

In conversation-based spatial grounding, the objective is to localize interrogative sentences with unknown objects in the given video (e.g. "What is caught by the squatting boy on the floor?"). Unlike grounding for declarative sentences where the explicit characteristics of objects (e.g. the class "toy" and visual appearance "yellow") are present within the sentence itself, grounding for interrogative sentences is challenging due to the fact that it can only depend on relationships between the unknown object and other objects (e.g. the action relation "caught by the squatting boy" and spatial relation "on the floor") (Figure 5). A benchmark based on this task can be regarded as a measure of the sufficient relationship construction and cross-modal relation reasoning ability of the video-language model.

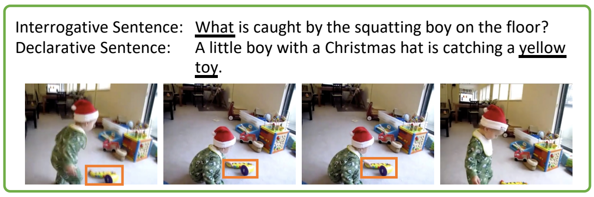

Figure 5: **Interrogative vs declarative sentences**

To evaluate our model for conversation-based video spatial grounding, we pass interrogative prompts to the model. It then generates descriptive textual responses to these prompts, from which Vicuna-13b-v1.5 extracts relevant referring expressions. These expressions are then passed into the GroundingDINO-based spatial grounding and tracking module. For the obtained object tracks, bounding box IoU is calculated by comparing them with the ground truth annotations.

From the two spatiotemporal grounding datasets, to form a spatial-only grounding benchmark, we crop the video in the temporal axis to contain only the segment where the target object is present, and the mean spatial IoU is reported as the metric for comparison.

It should be noted that we evaluate our model in these benchmarks only in the zero-shot setting, without any training on these datasets.

1. Benchmark based on the VidSTG Dataset: VidSTG dataset consists of videos paired with multiform sentences (both interrogative and declarative). To form a benchmark to quantitatively evaluate the performance of conversation-based video spatial grounding, we leverage the 5693 video and interrogative sentence pairs in its test set.

2. Benchmark based on HC-STVG Dataset: Unlike in VidSTG dataset, in HC-STVG dataset contains only declarative form sentences for all of its videos. Therefore interrogative sentences are first generated from the declarative text captions in 3025 samples of the test set using Vicuna-13b-v1.5 model. Then the evaluation is performed in a similar manner to VidSTG.

B.2 Generating Interrogative Statements

The original text annotations in the HC-STVG dataset are in the declarative statement format. In order to make our video prompt-based grounding evaluation pipeline, we extract interrogative statements (questions) from these text annotations using Vicuna-13b-v1.5 using the following prompt template.

```
SYSTEM:
    You are an intelligent
    chatbot designed for
    generating question-answer
    pairs from sentences.
USER:
    Your task is to generate
    a question and answer from
    the given sentence.
    The question should start
    with 'Who'.
    The question should refer
    to the subject of the
    given sentence.
    The answer should include
    the subject of the given
    sentence.
    Please generate the
    response in the form of
    a Python dictionary string
    with keys 'Q' for question
    and 'A' for answer.  Each
    corresponding value should
    be the question and answer
    text respectively.
    For example, your response
    should look like this:
    {'Q': 'Your question
    here...', 'A': 'Your
    answer here...'}.
    Please note that the
    generated question and
    answer should only include
    information from the given
    sentence.
    Please process the
    following sentence:
    The man in the suit goes
    to the man in white and
    looks at him.

ASSISTANT:
    {'Q': 'Who goes to the man
    in white?', 'A':'The man
    in the suit'}

USER:
    Please process the
    following sentence:
    <DECLARATIVE_STATEMENT>

ASSISTANT:
```

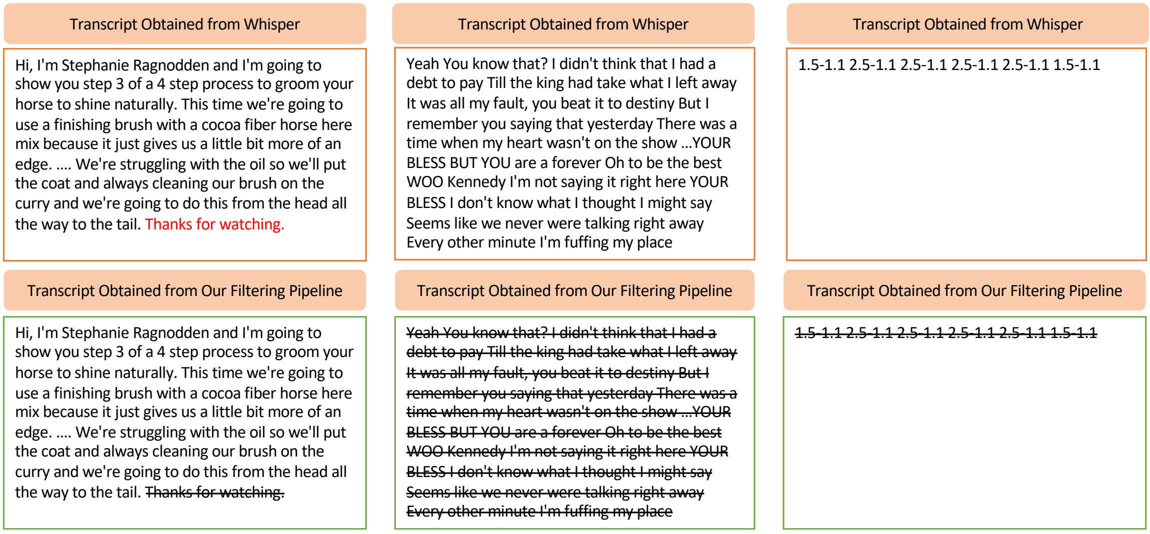

Figure 6: **Filtering the audio transcript**: to remove hallucinations (left), music (center), and garbage (right) characters from the raw audio transcript.

B.3 Extracting Referring Expression Using Vicuna

In the quantitative evaluation, we use the following prompt template with Vicuna-13b-v1.5 to extract the referring expression from the output of the video-based LMM, which is used as the input prompt to the off-the-shelf-grounding module.

```
SYSTEM:
    You are an intelligent
    chatbot designed for
    identifying the most
    relevant subject/object
    phrases in video-based
    question-sentence pairs.
USER:
    Your task is to compare
    the question with the
    sentence, and extract the
    subject or object phrase
    of the sentence that most
    accurately answers the
    given question.
    The selected phrase should
    be short and should
    contain only one noun.
    The selected phrase can
    include adjectives that
    explain the attributes of
    the subject/object.
    The selected phrase should
    not exceed 4 words.
    The selected phrase should
    not include articles ('a',
    'the', 'and').
    Please generate the
    response in the form of
    a Python dictionary string
    with keys 'OBJECT', where
    its value is the extracted
    phrase in Python string
    format.
    DO NOT PROVIDE ANY OTHER
    OUTPUT TEXT OR EXPLANATION.
    Only provide the Python
    dictionary.
    For example, your response
    should look like this:
    {'OBJECT': 'green toy'}.
    Please process the
    following video-based
    question-answer pair:
    Question: who is in front
    of the guitar at the show?
    Answer: A woman in a
    black dress is in front
    of the guitar on stage.
ASSISTANT:
    {'OBJECT': 'woman in
    black dress'}
USER:
    Question: who points to
    the window?
    Answer: The old man is
```

```
    pointing to the window.

ASSISTANT:
     {'OBJECT': 'old man'}
USER:
     Question:  who is inside
    the blue car?
    Answer:  The driver of the
    blue car.

ASSISTANT:
     {'OBJECT': 'driver'}
USER:
     Please process the
    following video-based
    question-answer pair:
    Question:
    <INPUT_TO_VIDEO_LMM>
    Answer:
    <OUTPUT_OF_VIDEO_LMM>
ASSISTANT:
```

B.4 Entity Matching with Vicuna

As shown in Figure 2, our method employs an LLM-powered entity matching module similar to (Zhao et al., 2023) to match the key phrases in the video-LMM's output with the object tracks obtained from the grounding and tracking module. We use the same prompt template as (Zhao et al., 2023).

C Qualitative Results for Visual Grounding

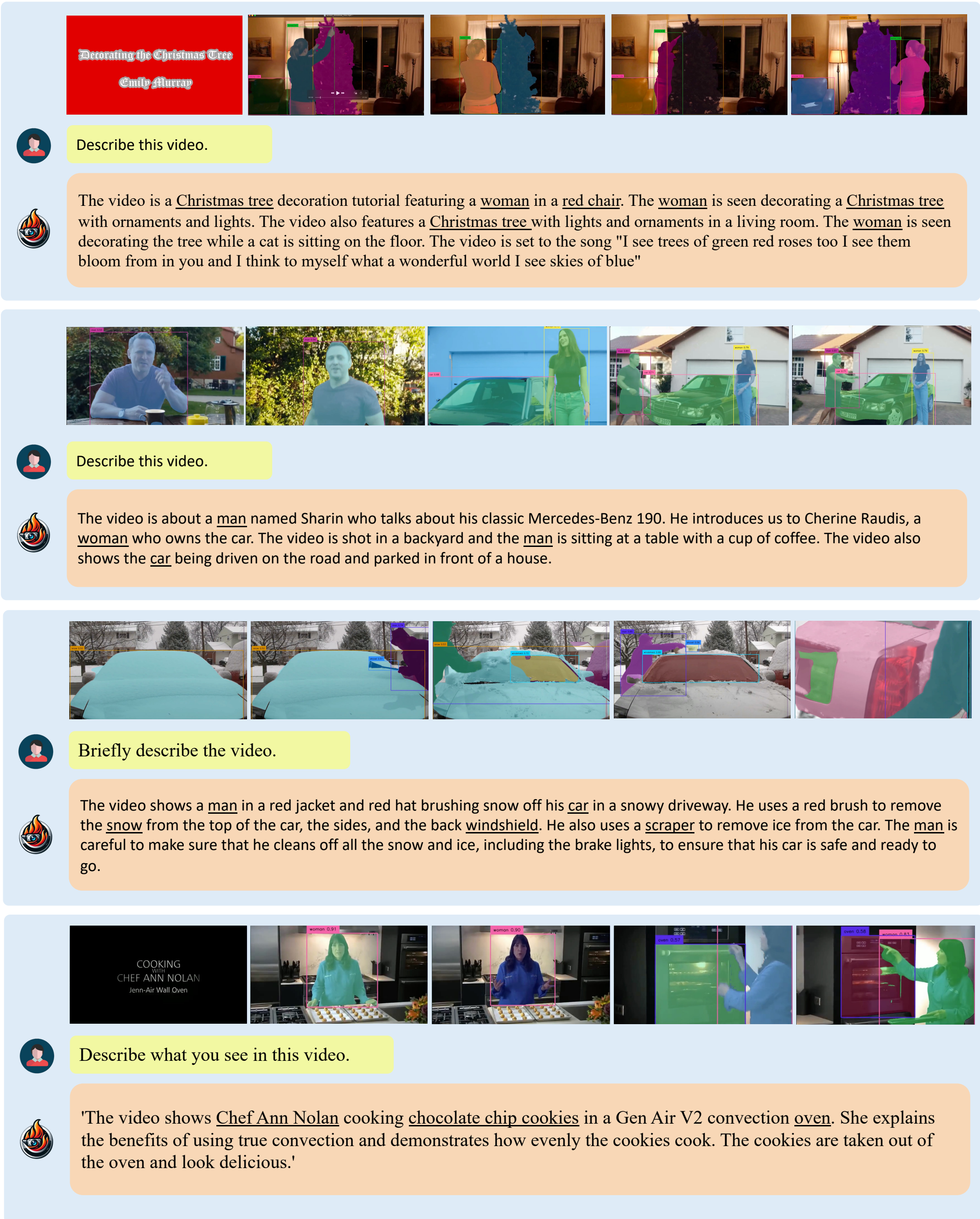

Figure 7: **Qualitative results for video grounding** obtained using image tags as the queries to the off-the-shelf grounding module and an entity matching module to match the image tags with the corresponding phrase in the LMM output. (e.g., in the 4th row, the phrase 'Chef Ann Nolan' is matched with the image tag 'woman'.)

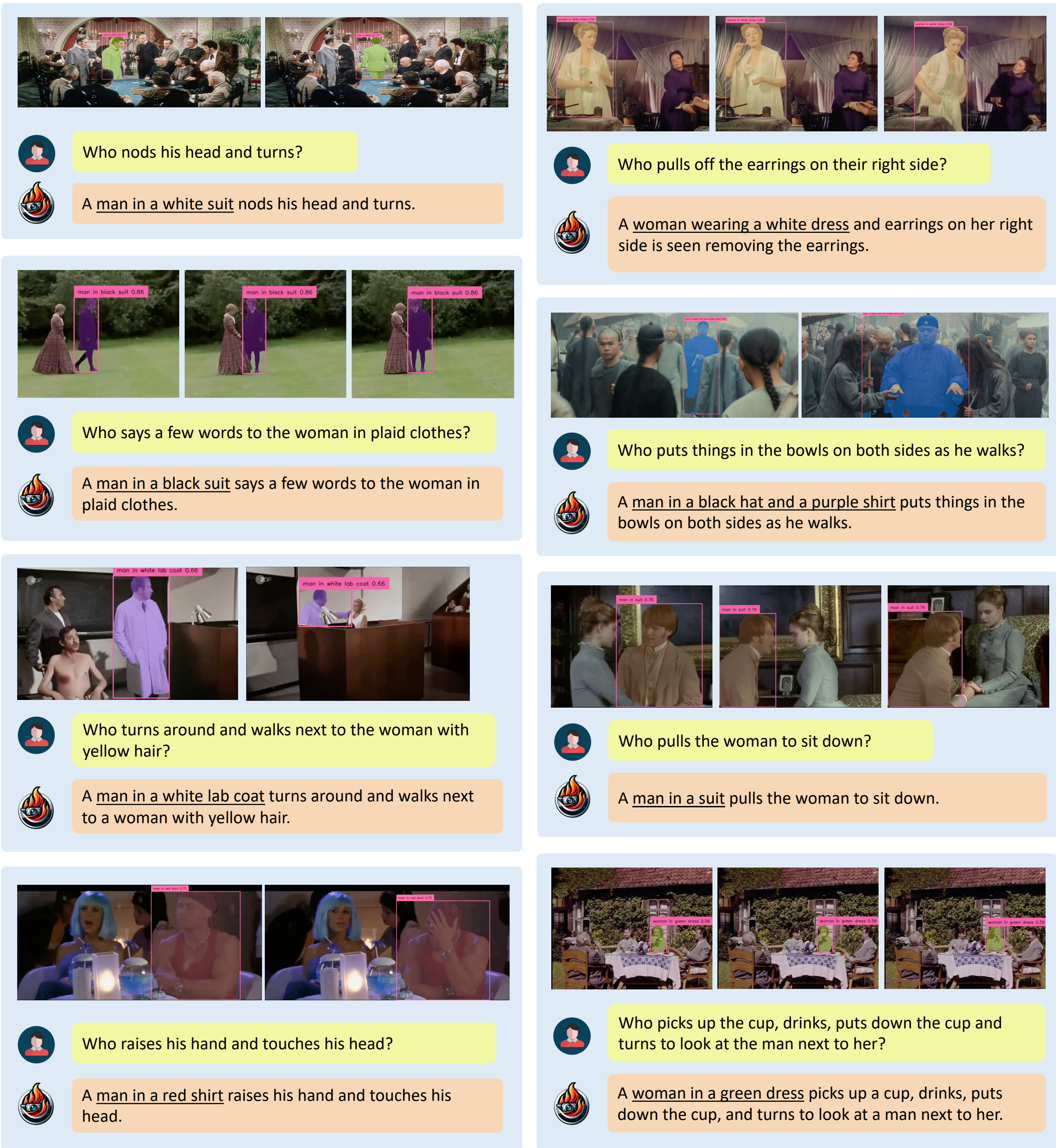

Figure 8: **Qualitative results for video grounding** on example videos from the HC-STVG(Tang et al., 2021) dataset. These results are obtained by using Vicuna with the prompt template in B.3 to extract the referring expression from the LMM output which is then passed to the off-the-shelf grounding module.

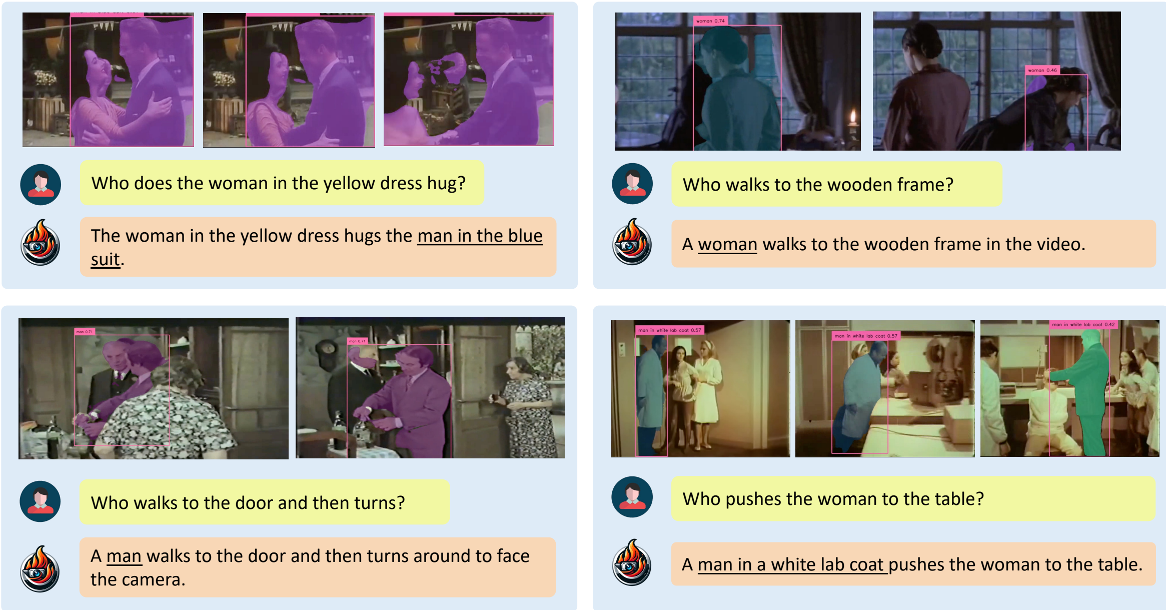

Figure 9: **Qualitative results for visual grounding on the HC-STVG dataset (failure cases):** errors in our model's output (e.g., bottom-left: our model identifies the woman as a man), incorrect localizations in the off-the-shelf grounding module (e.g., top-left), and incorrect tracking (e.g., top-right, bottom-right) result in these failure cases.

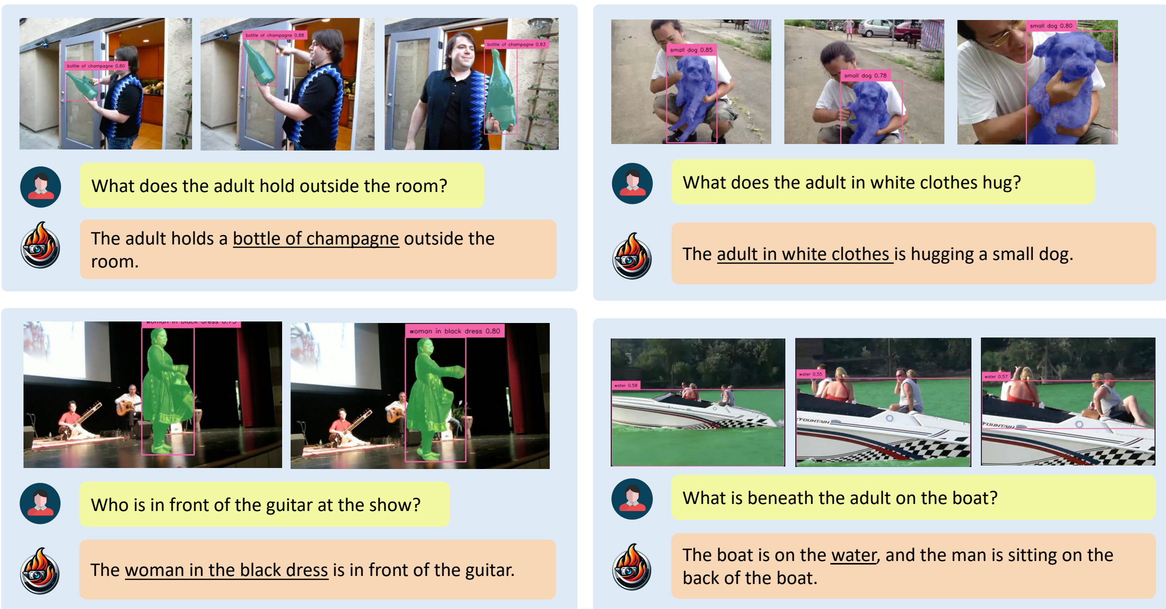

Figure 10: **Qualitative results for video grounding** on example videos from the VidSTG(Zhang et al., 2020) dataset

D Quantitative Evaluations of Video-based Conversation Performance

We leverage the video-based conversation performance benchmarks introduced in Video-ChatGPT(Maaz et al., 2023), while changing the evaluation LLM from GPT-3.5-Turbo to Vicuna-13b-v1.5 model. The prompt templates used with Vicuna are as same as with (Maaz et al., 2023).

Video-based Generative Performance Benchmarking: In this benchmark we continue to use the same test set of 500 samples curated from the ActivityNet-200(Heilbron et al., 2015) videos as in (Maaz et al., 2023).

Zero-Shot Question-Answer Evaluation: Following Video-ChatGPT, we perform zero-shot evaluation on four standard open-ended question-answer datasets: MSRVTT(Xu et al., 2016), MSVD(Xu et al., 2017), TGIF(Li et al., 2016), and ActivityNet-QA(Yu et al., 2019). No specific training is performed on these datasets, and the evaluation is performed in a zero-shot manner.

E Qualitative Analysis of Conversation Performance

The qualitative results in Figure 11 indicate the enhanced baseline's impact on PG-Video-LLaVA's performance. The PG-Video-LLaVA (13B) model exhibits improved accuracy in the information presented, a deeper level of descriptive detail, and a stronger alignment with the context and temporal progression of the videos. This advancement is particularly noticeable in the precise depiction of the child's engagement with their surroundings and the giraffe's behaviour, indicating a refined interpretation of both the activities and their settings. These qualitative insights are consistent with the quantitative results, highlighting the augmented baseline's role in advancing PG-Video-LLaVA's capacity in video understanding.

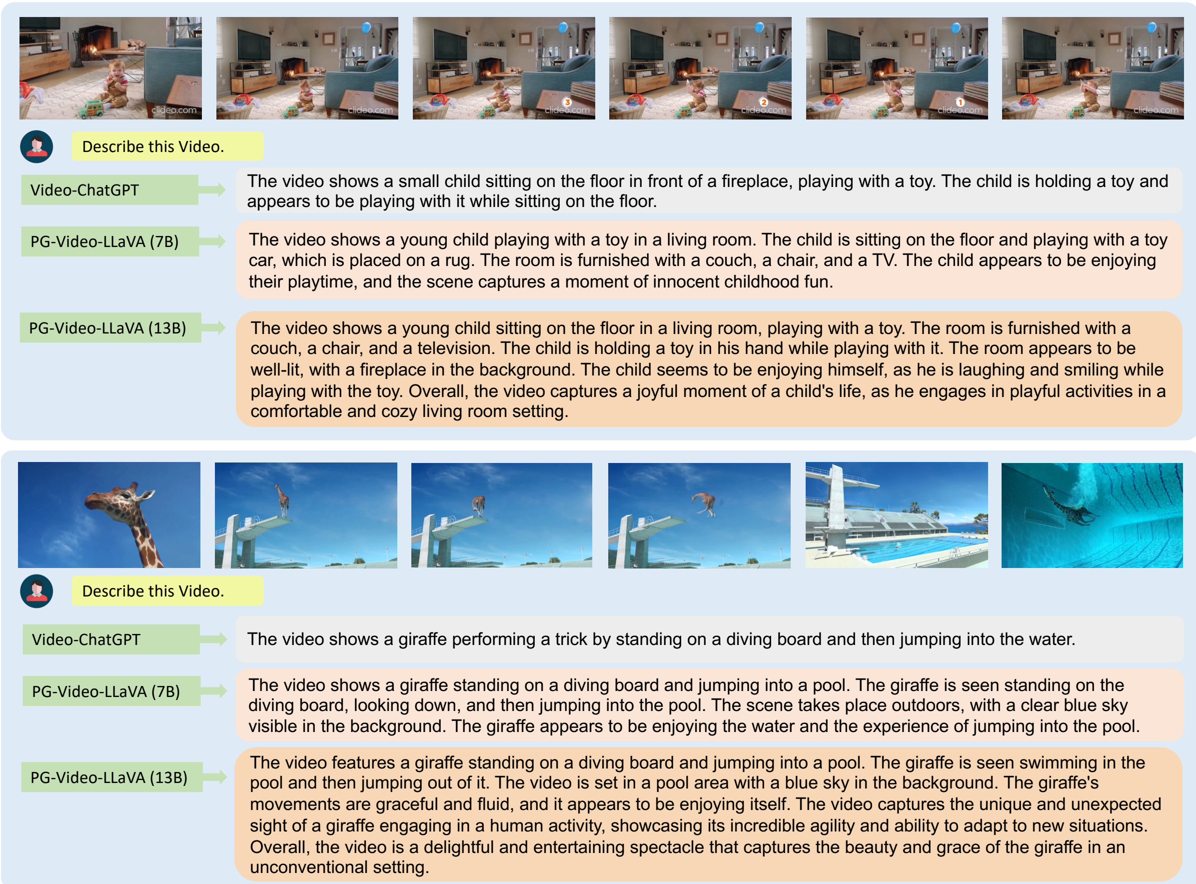

Figure 11: **Qualitative results comparison of Video-ChatGPT vs PG-Video-LLaVA (Ours)** Qualitative analysis of video descriptions generated by Video-ChatGPT, PG-Video-LLaVA (7B), and PG-Video-LLaVA (13B) models. The evolution in model performance is evident, with enhancements in the accuracy of information, richness of descriptive detail, and alignment with the video's context and sequence of events as we move from the baseline Video-ChatGPT to the more advanced PG-Video-LLaVA (13B) model.