Bridging the Gap: Adapting Video Language Models for Egocentric Understanding

Michael Cruz Christopher Lee Saurabh Aggarwal Robert Martin

University of Central Florida
Orlando, FL

mi484725@ucf.edu chris.lee@ucf.edu sa181349@ucf.edu ro134250@ucf.edu

https://github.com/mbcruz96/Video-EC4

Abstract

Recent developments in Large Video Language Models (LVLMs) have shown their ability to understand diverse video content [1, 2]. However, the first-person perspective of egocentric videos remains a major challenge. Egocentric data presents unique challenges such as rapid camera movement, high scene complexity, and the need to deduce context from subtle visual cues. To overcome these challenges, we propose Video-EC4, aimed at benchmarking and adapting state-of-the-art LVLMs (specifically Video-LLaVA and Video-ChatGPT) for resilient egocentric video understanding. We improve the models' ability to handle hand movements and withstand rapid motion and spatial understanding by fine-tuning them on Ego4d [12]. Our results on benchmarking Video-EC4 on the EgoSchema dataset demonstrates the significant potential of fine-tuning to improve VideoLLM capabilities, opening new applications in disciplines wherein the first-person perspective is critical [16].

1. Introduction

Large Language Models (LLMs) have completely revolutionized Natural Language Processing (NLP), making it capable of understanding and generating human like text. This encouraged researchers in the field to develop Large Video Language models (LVLMs), which were developed to align text and image modalities to extend this understanding as well as extending the functionalities to image generation. These state-of-the-art models demonstrate potential on several tasks such as video captioning, long-context Video QA, action recognition, and REC. Unfortunately, these LVLMs have been trained on third person perspective video datasets, curated to offer a consistent viewpoint which does not capture a complete scene context. This focus leaves a significant gap in understanding of videos which are captured from a first-person perspective, known as egocentric video. Such Ego-Centric data offer valuable insight into human behaviour, context, and their daily experiences. In this digital era, various Ego-Centric devices exist which produce a huge amount of Ego-Centric data such as GoPro's, wearable cameras, dashcams, Virtual Reality cameras etc... Understanding of such visual data would be extremely valuable for fields like assistive technology, augmented reality (AR), and human-computer interaction (HCI) [11, 13].

The inherent properties of egocentric video pose a significant challenge for traditional LVLMs, hampering their comprehension. Challenges such as unusual viewpoints, rapid camera movement, and the strong focus on hands and objects in the setting of action requires the model to go beyond the conventional video comprehension approach of merely depending on object recognition. These tasks require complex reasoning abilities regarding context, spatial dynamics, and the holistic motivations behind the captured scenes.

To close this gap, this paper introduces Video-EC4. We aim to benchmark existing state-of-the-art LVLMs (specifically Video-ChatGPT and Video-LLaVA) on the EgoSchema dataset and then finetuned the models on Ego4D to enhance the performance on such egocentric data [16, 12]. Subsequently, the models were benchmarked on EgoSchema once more to compare the performance of the fine-tuned models.

2. Related Works

2.1. Aligning Exocentric Models to Egocentric Data

Ego4D's data is split into several tasks like episodic memory or hand tracking. Some models like GroundNLQ use a custom pretraining phase to ensure the model is trained specifically on egocentric data. However, this technique does not retain the information gained from third-person perspective image data. An additional challenge to using egocentric data is that existing models are used to learning from data that is agnostic of the user's role as a fulcrum. The emphasis is on the captured scene rather than the position of the user; there is little overlap in the pretraining done on LVLMs. Because of the limited scope, our work examines the viability of pretrained models to accept egocentric data in a similar structure and output generation format that aligns with the existing model weights. This is accomplished via finetuning on Ego4d using a variation of Video-ChatGPT's annotated formatting [2].

2.2. Unified Dataset for Generalized Task Solving

The collected egocentric data is annotated via a task-detail-culmination style. To better utilize the assigned tasks to videos, the paper, Egocentric Video Task Translation, determines a synergy strategy among the tasks to create a more general task [7]. However, each video is annotated for a specific task, making artifacts of these intentions appear downstream. For pre-built models like Video-ChatGPT and Video-LLaVA, which have been trained on a broader range of videos, this approach still assumes the pretrained model has existing context for such tasks. Other papers like [Encode Store Retrieve] use models like Vicuna and finetune upon the dataset and perform well for the specific tasks present within the video, like episodic memory, but performs poorly on general QA tasks. Our approach sweeps across the Ego4d dataset and treats each video as task-agnostic to all but long-context Video QA.

3. LVLMs

The main goal of this work is to adapt two state-of-the-art Large Video Language Models (LVLMs) for comprehending egocentric video, namely Video-LLaVA and Video-ChatGPT [1, 2].

3.1. Video-LLaVA

The core innovation is the LanguageBind encoder, which creates a unified visual representation for both videos and images [1]. Learning from a variety of visual inputs if made possible by this technique, which aligns various visual modalities with a textual feature space. The model is primarily trained in two key stages namely understating training and instruction tuning.

- Understanding Training: The model picks up the ability to decipher textual descriptions combined with images or videos.

- Instruction Tuning: Video-LLaVA acquires the ability to generate answers that are aligned with complex instructions. For egocentric tasks, where the model must comprehend actions, context and the intents underlying captured events, this step is especially crucial.

3.2. Video-ChatGPT

Video-ChatGPT expands the functionalities of large language models (LLMs) by enabling them to synthesize in-depth conversations about videos [2]. Video-ChatGPT extends the Language-aligned Large Vision Assistant (LLaVA) framework for video specific conversational tasks, building upon image-based visual-language (VL) models [2]. Video-ChatGPT leverages the Vicuna language decoder and the CLIP visual encoder [2]. When it comes to video reasoning and comprehending spatial, temporal, and action-oriented video elements, Video-ChatGPT is exceptional.

3.3. Adapting LVLMs for Egocentric Tasks

Video-ChatGPT and Video-LLaVA were selected based on their capabilities to adjust to data, which in this scenario is egocentric data. Their emphasis on comprehending instructions and producing meaningful interactions aligns with the challenges associated with egocentric video, where action and context both are crucial. We hypothesize that finetuning these given models on egocentric datasets, such as Ego4d, will improve their ability to comprehend the distinct dynamics of first-person video content even more.

4. EgoCentric Data

4.1. Overview

Egocentric video, or first-person perspective video, provides a unique view into the wearer's perspective, activities, environment, and interactions [12]. In contrast to conventional third-person perspective videos, egocentric data exhibits several unique features:

- Wearer-centric Viewpoint: The camera follows the wearer's line of sight to record their perception of the environment.
- Hands-on Activities: Egocentric videos frequently highlight the hands and objects used in particular actions or daily duties.
- Rapid & Unpredictable: The wearer's head and body movements are reflected in the camera's movement, creating dynamic and occasionally shaky film.
- Contextual Importance: Interpreting egocentric video requires an understanding of the wearer's behaviors and the larger environment.

These features pose difficulties for conventional video interpretation models, which are intended for carefully selected, third-party datasets. Though it is these vary features that have enticed researchers to start exploring egocentric data more [3-7]. By fine-tuning LVLMs to manage egocentric data efficiently, new application opportunities may arise.

4.2. EgoSchema

The goal of the EgoSchema dataset is to advance egocentric video comprehension [16]. EgoSchema is based on the large-scale Ego4D dataset and performs multiple choice long-context Video QA [16]. This is done by asking multiple-choice questions pertaining to long-form videos that demand a better comprehension of complex visual information across extended time periods [8, 9].

4.3. Ego4D

Ego4D serves as a cornerstone for egocentric video research [12]. It provides a never-before-seen volume of video data, capturing a variety of scenes, people, and activities. Furthermore, the multimodal annotations offered by Ego4D—which include audio, gaze tracking, and other features—offer insightful information for creating strong models of egocentric understanding.

5. Experiments

5.1. Implementation Details

Ego4d is a massive dataset and the act of storing and accessing each index and corresponding video was out of scope due to time limitations and server capacities [12]. The methodology of the work given this constraint was benchmarking and training on a subset of the data rather than the entire length. A few possible routes were explored, such as the shorter length set of benchmarking clips or task-specific filtered versions of the full video dataset. The final choice was to use the full videos due to their easily accessible dictionary lookup IDs within the annotation file and their ease of use with the Video-ChatGPT annotation schema.

5.2. EgoSchema Benchmark

For Video-LLaVA and Video-ChatGPT, EgoSchema benchmarking was done through respective custom evaluation scripts that retrieved each question list and answer pair. The model was fed the question and numerical choices and asked to choose a best answer. The model returned a number choice representing its chosen response which was compared to the correct answer from its answer file. Due to the models' inability to do multiple choice long-context Video QA zero-shot out of the box, multiple prompts were evaluated to determine the best prompt to guide the model predictions.

5.3. Fine-Tuning

Due to Video-ChatGPT's deficient performance in zero-shot multiple choice long-context Video QA, Video-LLaVA was chosen to be fine-tuned on Ego4d. Annotations for the model were created utilizing the Video-ChatGPT annotation style, using two-actor conversations between the user and the model. We modified the script with the annotations from narrations of Ego4d [12, 10]. Ego4d has two passes of narrations; one correlating to a particular segment of video, and one proving one summary for the entire video [10]. To best

utilize this data the entire segments were combined into a summary and used as annotation, sans final summary to avoid duplicates. Alternative methods explored were to attempt to use the benchmark clips corresponding full-length video ID to match the clip to the narration file, and then compare the clip's timestamps with a segment from the full-length video's narration. Unfortunately, the timestamps did not match up and creating a meaningful narration was not feasible. Our approach that uses the full-length video also has the advantage of interfacing with LLaVA's format of 8 frames of video chosen from a distribution of frames to perform a generation. By combining the data across segments, this gives it a rich feature set to draw upon when making predictions.

6. Results

6.1. Benchmarking

Method	Accuracy
VIOLET	19.9
mPLUG-Owl	31.1
InternVideo	32.1
InternVideo2-6B	41.1
Video-ChatGPT	27.6
Video-LLaVA	37.4
Gemini 1.5 (1st Frame)	54.3
Gemini 1.5 (16 frames)	**64.5**
Gemini 1.5 (150 frames)	63.6

Table 1: Comparison of benchmarks on EgoSchema dataset.

The results of benchmarking were commendable. The zero-shot capabilities of comparable models such as InternVideo [14] are on par with other models of their tier. However, the performance for innovative models like Gemini by Google [15] is expectedly not up to par. With more time and prompt tuning it seems that the zero-shot capabilities of the Video-ChatGPT and Video-LLaVa could potentially achieve an accuracy score upwards of forty percent before finetuning.

6.2. Results of prompt tuning

The results of benchmarking initially tended to be biased towards choosing option 0, skewing the results towards that area. To combat this, additional information was included in the prompt. These changes not only reduced this bias but also led to gains in accuracy.

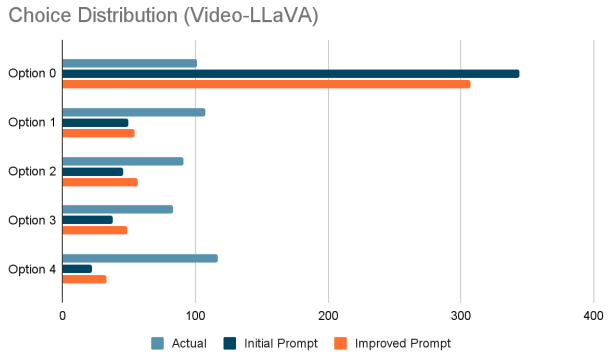

Figure 1: Egocentric performance for Video-LLaVA and Video-ChatGPT comparision with Gemini for zero-shot capability.

The first passthrough with a typical "summarize this" prompt yielded the dark blue line distribution, adding additional requirements to the question prompted the model's second attempt of output generation, in orange, to be more in line with the actual results, in light blue.

7. Limitations

The largest limitation of the LVLMs benchmarked and finetuned with egocentric data is related to the typical problems that they have with counting and generating enumerated guesses. Based on the experiments done with prompt tuning, there is evidence that the models perform poorly on zero-shot multiple choice long-context Video QA related tasks. While further prompting may be employed to account for this, it may be intrinsically linked to the base LVLMs being poor at this downstream task. Specifically, for Video-LLaVA, the distribution of options selected was skewed and heavily favoured the first option, meaning when the model was unsure of the correct answer, it most like selected the first option. Regarding Video-ChatGPT, the model was unable to select an option for every question of the benchmark, regardless of the prompt used during evaluation. This is indicative of Video-ChatGPT's inadequate performance in zero-shot multiple choice long-context Video QA related tasks.

Additionally, the fine-tuning pipeline for both models was ineffective for the Ego-4D dataset. Both models process video samples by dividing the sample into equal partitions. By doing so, the models are not leveraging the fact that major events in the videos may occur at various moments. Segmenting the video into equal parts does not adequately partition these events throughout the video. Furthermore, there may be more events that occur than the constant number of scenes the videos are segmented into, which loses vital per-segment information.

In conjunction with the video segments, the fine-tuning pipeline was heavily dependent on the annotations for the videos. We chose to use the voice annotations included in the Ego-4D dataset which hindered the finetuning process. The annotations were broken down into key events that occurred in the video. Because both models processed the videos by breaking the video up into equal length segments, the individual clip data was not being sufficiently used. Instead, we appended all the individual clip annotations into a single video annotation and used that for the fine-tuning process. This may have impeded the fine-tuning process, as key clip specific information was lost in the training process.

Conclusion and Future Directions

The work performed is a combination of benchmarking on two models, demonstrating performance comparable to other similarly tiered models. To utilize Ego4D for finetuning these models, a custom annotation script was developed to generate new annotations formatted for Video-LLaVA. Given more time to allow refine the method of finetuning and to give the schema ample time to train. Routes for future work is to align LLaVA's video processing in a dynamic way. For each of the frames of video LLaVA reads, one might select a video frame found between each sub step of narration, to better select frames that represent a segment of narration. This would better suit the model to learning segments of videos and could capture a richer context of text. Other routes for future work include exploring the model's potential on task specific data for benchmarking with Ego4Ds benchmarking labels – catering the model to recognize hands.

References

[1] Lin, Bin, et al. "Video-llava: Learning united visual representation by alignment before projection." *arXiv preprint arXiv:2311.10122* (2023).

[2] Maaz, Muhammad, et al. "Video-chatgpt: Towards detailed video understanding via large vision and language models." *arXiv preprint arXiv:2306.05424* (2023).

[3] Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." *Advances in neural information processing systems* 35 (2022): 23716-23736.

[4] Shen, Junxiao, John Dudley, and Per Ola Kristensson. "Encode-Store-Retrieve: Enhancing Memory Augmentation through Language-Encoded Egocentric Perception." *arXiv preprint arXiv:2308.05822* (2023).

[5] Pramanick, Shraman, et al. "Egovlpv2: Egocentric video-language pre-training with fusion in the backbone." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[6] Qinghong Lin, Kevin, et al. "Egocentric Video-Language Pretraining." *arXiv e-prints* (2022): arXiv-2206.

[7] Jia, Baoxiong, et al. "Egotaskqa: Understanding human tasks in egocentric videos." *Advances in Neural Information Processing Systems* 35 (2022): 3343-3360.

[8] Xue, Zihui, et al. "Egocentric video task translation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[9] Tang, Yunlong, et al. "Video understanding with large language models: A survey." *arXiv preprint arXiv:2312.17432* (2023).

[10] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.

[11] "Pre-Annotations Narrations." Annotation Guidelines Ego4D, ego4d-data.org/docs/data/annotation-guidelines/#pre-annotations-narrations.

[12] "DoMSEV Dataset." Papers with Code, paperswithcode.com/dataset/domsev.

[13] "Ego4D Dataset." Papers with Code, paperswithcode.com/dataset/ego4d.

[14] "EOAD (Egocentric Outdoor Activity Dataset)." Zenodo, zenodo.org/records/7738719.