EAGLE: Egocentric AGgregated Language-video Engine

Anonymous Authors

ABSTRACT

The rapid evolution of egocentric video analysis brings new insights into understanding human activities and intentions from a first-person perspective. Despite this progress, the fragmentation in tasks like action recognition, procedure learning, and moment retrieval, etc., coupled with inconsistent annotations and isolated model development, hinders a holistic interpretation of video content. In response, we introduce the EAGLE (Egocentric AGgregated Language-video Engine) model and the EAGLE-400K dataset to provide a unified framework that integrates various egocentric video understanding tasks. EAGLE-400K, the *first* large-scale instruction-tuning dataset tailored for egocentric video, features 400K diverse samples to enhance a broad spectrum task from activity recognition to procedure knowledge learning. Moreover, EAGLE, a strong video-based multimodal large language model (MLLM), is designed to effectively capture both spatial and temporal information. In addition, we propose a set of evaluation metrics designed to facilitate a thorough assessment of MLLM for egocentric video understanding. Our extensive experiments demonstrate EAGLE's superior performance over existing models, highlighting its ability to balance task-specific understanding with comprehensive video interpretation. With EAGLE, we aim to pave the way for novel research opportunities and practical applications in real-world scenarios.

CCS CONCEPTS

• **Computing methodologies → Planning with abstraction and generalization**; **Planning for deterministic actions**; *Planning under uncertainty*; **Neural networks**; Search with partial observations; **Image and video acquisition**; **Natural language generation**; **Information extraction**; **Discourse, dialogue and pragmatics**; **Temporal reasoning**; **Spatial and physical reasoning**; **Computer vision problems**; **Computer vision representations**; **Computer vision**.

KEYWORDS

Augmented Reality, Egocentric Video Analysis, Integrated Video Understanding Framework, Egocentric Video Dataset, Spatial and Temporal Information Processing, Multimodal Large Language Models (MLLMs), Comprehensive Video Interpretation, Performance Evaluation Metrics

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Unpublished working draft. Not for distribution.

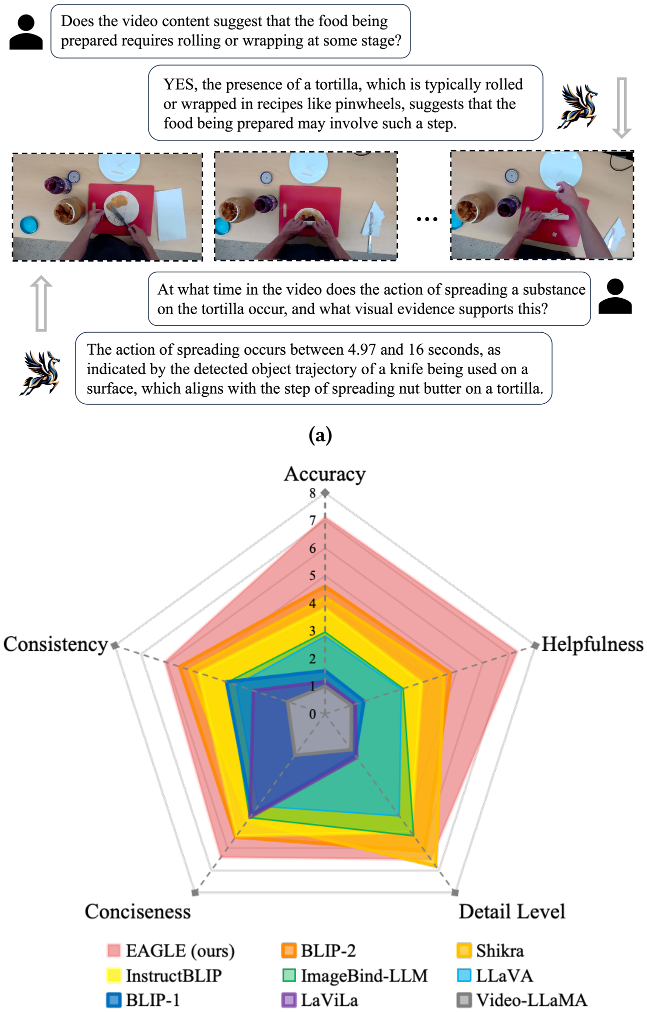
ACM MM, 2024, Melbourne, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Understanding human activities and intentions in videos is a key challenge for intelligent systems, requiring advanced reasoning capacities. While there have been advancements in computer vision, the most notable breakthroughs are seen in the evolution of Large Language Models (LLMs) [14, 63]. These models benefit from increased data and model size, resulting in enhanced



Does the video content suggest that the food being prepared requires rolling or wrapping at some stage?

YES, the presence of a tortilla, which is typically rolled or wrapped in recipes like pinwheels, suggests that the food being prepared may involve such a step.

At what time in the video does the action of spreading a substance on the tortilla occur, and what visual evidence supports this?

The action of spreading occurs between 4.97 and 16 seconds, as indicated by the detected object trajectory of a knife being used on a surface, which aligns with the step of spreading nut butter on a tortilla.

(a)

(b)

Figure 1: (a) illustrates the EAGLE, a framework designed to unify egocentric video tasks, thereby facilitating inter/intra-task understanding. (b) shows evaluation results of existing methods, including our EAGLE model and BLIP-2 [43], BLIP-1 [44], InstructBLIP [15] etc., using the newly proposed metrics on the EAGLE-400K benchmark.

generalizability, which is often challenging to achieve in computer vision tasks. By leveraging the pre-trained LLMs [21, 106], MLLMs [9, 12, 15, 25, 29, 43, 44, 50, 101, 105] show superior results to a wide spectrum of multimodal tasks [22, 34, 39, 58, 59, 77, 81]. Unlike current MLLMs that predominantly focus on images, EAGLE advances to capture spatial and temporal information to enable more in-depth video analysis. To enable MLLM to achieve a more holistic and detailed examination of human activities, our work pivots from previous efforts focused on third-person view [10, 19, 37, 40, 82], towards the egocentric view, offering an unfiltered and untrimmed perspective. This perspective enhances exocentric tasks like action recognition and localization by offering insights into individual interactions with their surroundings and facilitates new unique tasks like Natural Language Queries and Action Anticipation [23]. These tasks demand an in-depth view of the video content, including activity recognition and procedure knowledge learning. Taking sandwich preparation as an example, the task requires recognizing actions like preparing ingredients and spreading condiments and understanding how these actions contribute to the overall process. Pioneering efforts like EPIC-KITCHENS-100 (EPIC-KITCHENS) [38] and Ego4D [23] have paved the way for tasks focused on activities like temporally localizing and anticipating actions. Subsequent research [5, 8, 74] has extended these concepts by introducing tasks that emphasize procedure knowledge, aiming to understand actions' intentions and contextual relevance.

While diverse tasks offer more insights, they also foster task-specific models, similar to traditional approach in the NLP field, where models are trained for specific tasks like sentiment analysis, translation, and question-answering, etc. This results in a fragmented approach, where each model specializes in a specific aspect. For instance, one model may excel in recognizing actions timestamp (e.g., identifying a *'gra b a spoon'* from seconds 5-7), while another pinpoints the timing of such actions. These two example tasks, though different in focus—action recognition versus temporal localization—essentially seek to identify the action and its temporal occurrence. Many works [33, 35, 36, 54] have attempted to mitigate these problems by employing a shared backbone [68, 75] or re-scaling labels [48, 99]. These approaches are limited by their reliance on task-specific models, highlighting the challenge in egocentric video understanding: balancing specialization with a holistic grasp of video content.

Addressing the above challenges, we introduce the EAGLE-400K dataset, the *first* large-scale instruction-tuning dataset tailored for egocentric video. We provide a unified task interface that not only integrates existing tasks but also fosters the development of new, context-rich tasks as shown in Table 3 Compared with existing large image-based instruction tuning (LLaVA-150K [50], VideoInstruct100K introduced by Video-ChatGPT [55]), our method is *3-4*× times larger to facilitate the research field. This dataset is a comprehensive collection designed to advance the understanding of activities and procedure knowledge in an egocentric view. It comprises 36k video clips sourced from three different origins: Ego4D and EPIC-KITCHENS, which are for activity recognition, and PTA, which is crucial for procedure learning, as detailed in Table 4. By employing instruction tuning, EAGLE-400K unified fragmented tasks as coherent (`VIDEO, INSTRUCTION, RESPONSE`) pairs, thereby serving as a high-quality, large-scale video instruction tuning dataset, as shown in Table 3. Moreover, EAGLE-400K leverages existing annotations to facilitate knowledge sharing across datasets, which enables the creation of novel tasks, such as Temporal Reasoning and Cross-Referencing Events as shown in Table 3that were not present in the original dataset.

Complementing the dataset, we propose EAGLE, a video-based MLLM, we augment its capacity for spatial and temporal reasoning through the integration of the Adapter [28] We conducted a systematic evaluation to demonstrate the efficacy and adaptability of the proposed dataset and model, comparing EAGLE with leading MLLMs, including BLIP-2 [43], BLIP-1 [44], InstructBLIP [15], LaViLa [105], LLaVA [50], ImageBind-LLM [25], Shikra [12], Video-LLaMA [102]. The results, as illustrated in Figure 1b, EAGLE outperforms all models on the proposed benchmark. We summarize our main contributions as follows:

- **EAGLE-400K Dataset:** Our work introduces the pioneering large-scale video instruction-tuning dataset for egocentric video understanding [83], providing a unified task interface to alleviate the models and task fragmentation. At *4*× times the size of the previously largest video instruction-tuning dataset, EAGLE-400K is expected to greatly benefit the community by encouraging further novel research and serving as a benchmark for evaluation.
- **PTA dataset** To fill the gap in procedural understanding within current egocentric video datasets, we have collected and annotated the Perception-driven Task Assistance (PTA) dataset.This dataset contains 268 egocentric videos, each recorded with certain recipe scripts to provide a rich, detailed insight into specific procedural tasks, advancing our understanding of egocentric procedures.
- **EAGLE Model.** We introduced the EAGLE model, a novel video-based MLLMs designed to excel at capturing both spatial and temporal information with the advantage of EAGLE-400K. To the best of our knowledge, we are the first to incorporate fine-grained object trajectories, temporal boundaries, and scripted procedure videos for *video instruction tuning*.
- **Evaluation Metrics.** we provide an comprehensive analysis of current state-of-the-art MLLMs to highlight their limitations and the challenges of applying them to egocentric video understanding.e proposed a novel metric designed to offer a more comprehensive assessment, further demonstrating our model's superior performance on the proposed benchmark.

2 RELATED WORK

2.1 Egocentric Video Understanding

Egocentric Video Understanding began with pioneering datasets [17, 46, 66] that demonstrated the unique potential of first-person video analysis. The field expanded with EPIC-KITCHENS [16], featuring 100 hours of videos, and further with Ego4D [24], which boasts an impressive 3,000 hours of data. These expansions inspired a wide range of research tasks, including human-object interactions [60, 97], activity recognition [38, 69, 87, 93], sounding object localization [1, 31, 32, 56, 96, 107], pose estimation and prediction [6, 62, 89], procedure knowledge learning [5, 26], and social understanding [76]. However, various tasks have resulted in specialized, fragmented model development. EAGLE-400K aims to consolidate these tasks for a more holistic video understanding.

2.2 LLMs for Multimodal Understanding

Recent advancements have extended LLMs to multimodal domains, resulting in MLLMs [2, 15, 20, 42, 44, 47, 102, 109] that excel in various tasks. Fine-grained multimodal Understanding involves a detailed understanding of visual content, including spatial details [11, 12, 41, 65, 90, 94, 98, 101, 104], temporal sequences [48, 48, 64, 84, 85, 91, 92], or a combination of both [7, 51, 88]. Models like [15, 44, 109] use a two-stage Q-former to align vision and language models. [102] aligns video and audio modalities with LLMs by training adapters, showing its ability to integrate multiple modalities effectively. Video-ChatGPT [55] and VideoChat [45], combining LLMs with video foundation models, are tailored for coarse-grained video-based conversations. However, few MLLMs are designed to tackle both spatial and temporal video tasks [83], and our work emphasizes interpreting 16 seconds videos, which are *2-4*× longer compared with other video MLLMs.

2.3 Fine-grained Multimodal Comprehension

Fine-grained multimodal comprehension involves a detailed understanding of image or video content, including spatial [11, 12, 41, 65, 90, 94, 98, 104], temporal [48, 48, 64, 79, 84, 91, 92], or both spatial and temporal [7, 88] information. The multimodal models for fine-grained spatial understanding like [12] and [98] are utilizing LLMs trained on an instruct-tuning dataset which is produced by the language-only GPT-4 and include the coordinates of objects' bounding boxes. They can handle multiple location-related multimodal tasks like REC, PointQA, dense image captioning, and VQA. In [65, 104], special tokens representing the regions are used, while [11] and [3] adopt both special tokens and coordinates. [41, 90] implemented irregular pixel-level region segmentation, generating descriptive captions for any object within an image. The multimodal models for fine-grained temporal video understanding, including [48, 64, 85], are leveraging the capabilities of LLMs. There are seldom multimodal models designed to handle both spatial and temporal video understanding tasks.

3 EAGLE-400K DATASET AND BENCHMARK

Egocentric video understanding [61, 67] involves two primary aspects: *activity recognition*, which identifies individual actions like picking up objects, and *procedure knowledge learning*, which models sequential action relations to understand their contribution to achieving a goal. We aim to *consolidate multiple datasets* with different focuses and provide a comprehensive dataset. We start with two popular egocentric datasets, EPIC-KITCHENS [38] and Ego4D [23], featuring long-term, untrimmed videos of daily tasks. These datasets are annotated with natural human actions and object interactions without predefined procedures, focusing on solely on identifying actions.

However, these existing datasets provide only action labels without encapsulating procedure knowledge. To bridge this gap, we've also gathered the PTA dataset, consisting of 268 egocentric videos recorded in laboratory settings. This dataset is specifically designed to enhance procedure knowledge learning through detailed visualization of three distinct recipes: pinwheel, mug cake, and broil coffee. Unlike previous approaches [5, 78] which prioritized task diversity but lacked depth within individual tasks, our approach focuses on providing extensive variation and a higher number of samples within a select few tasks. This approach enables a more comprehensive analysis of procedural steps, making the PTA dataset a valuable resource. Representative examples from PTA dataset are shown in Figure 2.

Table 1: The table compares vision-language instruction-tuning datasets, including EAGLE-400K and MIMIC-IT. MIMIC-IT generates questions from visual descriptions but often produces questions not closely related to the visual content due to noisy narration. VideoInstruct is generated from ActivityNet-200 [27] and serves as popular video instruction tuning dataset, featuring short clips paired with QA-style data without spatial-temporal understanding.

Dataset	Video	#Clip	#Ins.	#Ins./clip	Duration
MiniGPT-4 [109]	×	-	5K	-	-
Shikra-RD [12]	×	-	5.9K	-	-
LLaVA [50]	×	-	345K	-	-
MIMIC-IT [42]	✓/×	400K	2.4M	6	4-8 frames
VideoInstruct [45]	✓	13k	100k	7	5 s
EAGLE-400K	✓	36K	400K	11	16-76s

We split the data into training and validation sets according to established splits for Ego4D and EPIC-KITCHENS. For PTA, we used a 70/30 split, excluding videos from one lab to serve as a novel testing environment. The remaining testing videos were randomly sampled as detailed in Table 4.

3.1 Annotation

For EPIC-KITCHENS split, we utilized official annotations that include action-object labels with temporal boundaries as shown in Figure 2. Additionally, we integrated spatial annotations from the EPIC-KITCHENS-VISOR dataset [18], an extension of EPIC-KITCHENS, providing object segmentation trajectories covering one-third of the original EPIC-KITCHENS dataset. In the case of Ego4D, the initial ~3.8 million narrations underwent refinement to generate various subsets, as outlined in [24]. Our focus lies on the Episodic Memory and Forecasting Benchmark, which includes tasks such as Natural Language Queries, Moment Queries, and Long-term Action Prediction tasks, all tailored for activity understanding. In the PTA subset, each video depicts the process of making a recipe, with timestamps marked for key procedure steps.

To enrich the annotation with object information, we first fine-tuned the DINO [52] using the EgoObject dataset [108] without its class head, significantly improving its object proposal accuracy to over **90%** on the test set. Next, we integrated this enhanced DINO model with the latest DEVA [13] tracker, achieving reliable object tracking from an egocentric viewpoint. Lastly, we employed the LLaVA-13B model to interpret the semantic meanings of the proposed object regions. As shown in Figure 2, while this approach may not always reach the accuracy level of human annotation—occasionally mistaking a tortilla for flatbread, for instance—it marks a considerable leap forward, especially given the scarcity of zero-shot vision models capable of achieving high accuracy in grounding.

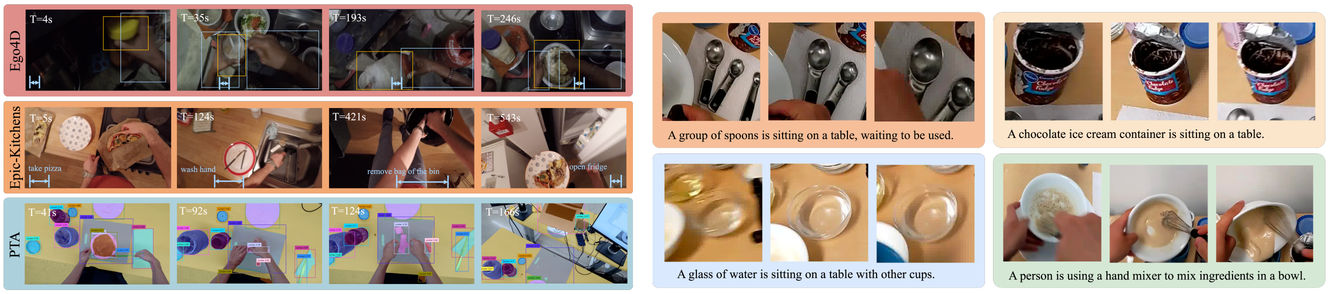

Ego4D
T=4s
T=35s
T=193s
T=246s
Epic-Kitchens
T=5s
take pizza
T=124s
wash hand
T=421s
remove bag of the bin
T=543s
open fridge
PTA
T=41s
T=92s
T=124s
T=166s
A group of spoons is sitting on a table, waiting to be used.
A chocolate ice cream container is sitting on a table.
A glass of water is sitting on a table with other cups.
A person is using a hand mixer to mix ingredients in a bowl.

Figure 2: Left: Representative frames from the Ego4D [23], EPIC-KITCHENS [16], and PTA datasets, showcasing the intricate capture of task-oriented activities. Right: Visualizations of trajectories and object interactions within the EAGLE-400K dataset, emphasizing the tasks' complexity and diversity.

Table 2: An example task on EAGLE-400K: the preparation of a dish called "Pinwheel" from PTG data. It details the dish preparation process and featured objects. The task involves placing various ingredients on a tortilla with a knife, performed by a participant wearing a camera. The top block presents prompts for GPT, including captions and object boxes, while the bottom block shows question types and responses. Notably, the visual image does not prompt GPT.

Context type 1: Task Description

Pinwheels with steps 1: Place the tortilla on the cutting board., 2: Scoop nut butter and spread it on the tortilla, leaving a margin at the edge., 3: Clean the knife with a paper towel., 4: Using the knife, scoop jelly from the jar and spread it over the nut butter., 5: Clean the knife with a paper towel., 6: Roll tortilla into a tight, 1.5-inch thick log without squeezing out the filling, 7: Secure the roll with 5 toothpicks spaced 1 inch apart., 8: Trim tortilla roll ends, leaving a 1/2 inch margin near the last toothpick; discard the ends., 9: Place floss under the roll, halfway between two toothpicks, perpendicular to its length, 10: Cross floss ends over the roll and pulls in opposite directions to slice., 11: Continue slicing with floss to create 5 pinwheels., 12: Place the pinwheels on a plate..

The current step, as ground truth, is: <0,16> 4: scoop jelly and spread jelly

Context type 2: Object Trajectory

A jar of ice cream is sitting on a table.: [12, 0.215, 0.57],[4, 0.17, 0.57],[10, 0.2, 0.56],[7, 0.185, 0.545],[6, 0.175, 0.47],[8, 0.175, 0.52],[13, 0.16, 0.53],[14, 0.185, 0.695] A person is using a knife to spread peanut butter on a plate.: [12, 0.78, 0.61],[9, 0.755, 0.59],[11, 0.75, 0.59],[8, 0.735, 0.555],[6, 0.745, 0.52],[14, 0.73, 0.55],[7, 0.75, 0.57],[10, 0.755, 0.6] A bowl of peanut butter is being held by a person.: [11, 0.765, 0.68],[0, 0.71, 0.795],[4, 0.75, 0.65],[3, 0.675, 0.645],[13, 0.755, 0.675],[8, 0.735, 0.645],[12, 0.785, 0.705],[14, 0.755, 0.66] A man is using a knife to cut a piece of cheese on a plate.: [5, 0.79, 0.825],[8, 0.855, 0.81],[6, 0.85, 0.75],[12, 0.875, 0.905],[15, 0.84, 0.94],[14, 0.83, 0.88],[11, 0.905, 0.85],[10, 0.905, 0.895] A person is holding a bowl of yellow paint.: [6, 0.76, 0.6],[1, 0.87, 0.835],[4, 0.735, 0.63],[10, 0.76, 0.67],[5, 0.725, 0.615],[7, 0.77, 0.645],[11, 0.765, 0.675],[13, 0.74, 0.645]

Response type 1: Task Verification

Question: Based on the video content, is the person in the video currently spreading something on the tortilla, and if so, what is it?

Answer: YES, the person is spreading jelly on the tortilla. This is indicated by the action of using a knife to scoop out a substance, likely jelly, from a container and then spreading it onto a flat surface, which is consistent with step 4 of the recipe.

Response type 2: Step Verification

Question: At what moment in the video can you confirm that jelly is being spread on the tortilla, and what visual clues support this?

Answer: The action of spreading jelly on the tortilla can be confirmed throughout the video, as indicated by the trajectory of the knife moving across the frame, consistent with a spreading motion. The presence of a container from which the substance is being scooped also supports this.

Question: How does the movement of the knife in the video align with the current step of spreading jelly?

Answer: The movement of the knife in the video aligns with the spreading action required in the current step. The trajectory data shows the knife moving across the frame in a manner consistent with spreading a substance on a flat surface.

Response type 3: Objects Verification

Question: At what time and location in the frame does the jar of jelly appear, and how is it relevant to the task?

Answer: The jar of jelly appears throughout the video, with its location varying slightly but generally centered around the coordinates (0.2, 0.57). It is relevant to the task as it is the source of the jelly being spread on the tortilla, which is a key step in making pinwheels.

3.2 Instruction Tuning Data Generation

As previously mentioned, diverse tasks and inconsistent annotation standard often limit the comprehensive understanding of videos. We adapt the instruction tuning [100] to unify these annotations under a cohesive framework. In our dataset, videos are segmented into 16-second clips, *3-5*× longer than common video understanding dataset, ensuring each contains a rich number of actions while maintaining a manageable length, as shown in Table 4. By comparison, our baseline model, LaViLa [**?**], which is trained specifically on egocentric videos, typically takes a 1-sec clip. Another example is EPIC-KITCHEN Action Anticipation task, although videos tend to be minutes, only a 5-second segment is used for analysis. Adopting 16-second clips allows us to capture comprehensive action details without overwhelming the model.

To determine the optimal frame rate, we draw inspiration from recent studies [73, 95] that have shown promising results in frame-based video understanding by analyzing videos frame-by-frame and using feature pooling. Building on this, we sample one frame per second, maintaining a consistent interval regardless of the video's frame rate. To enhance contextual understanding, we incorporate temporal context with 30 seconds before and after each clip. We chose a 30-second duration to balance action details and cohesive

Table 3: The table outlines activities in a kitchen video, including opening/closing cabinets, grabbing a knife, and washing vegetables, showcasing a person's kitchen work. It serves as an example of instruction-following data. The top block displays prompts like captions and boxes for GPT, while the bottom block shows response types. Notably, the visual image does not prompt GPT and is included for reference only.

Context type 1: Temporal History

Past 30 second: take container, take tofu, close fridge, open fridge, take carrots and, open drawer, close fridge, put down vegetables, open cupboard, take cutting board, put down cutting board Current <0,0.76> close cupboard, <3.66,5.0> open drawer,<5.5,8.0> take knife, <5.55,6.36> take knife, <6.73,7.74> put down knife, <8.1,8.6> close drawer, <9.45,10.7> put down knife,<11.05,14.3> open tap,<13.86,16> wash courgette

Future 30 second: still washing courgette, wash carrot, still washing carrot, close tap, put down vegetables, open cupboard, take grater, take pan, put down pan, close cupboard, close cupboard, take courgette, cut courgette

Context type 2: Object Trajectory

'right hand': [[5.0, 0.295, 0.401], [6.0, 0.317, 0.419], [7.0, 0.294, 0.365], [8.0, 0.324, 0.406], [10.0, 0.303, 0.377], [12.0, 0.344, 0.366], [13.0, 0.336, 0.284]] 'left hand': [[5.0, 0.186, 0.407], [6.0, 0.207, 0.427], [7.0, 0.212, 0.402], [8.0, 0.214, 0.402], [11.0, 0.293, 0.368], [12.0, 0.113, 0.324], [13.0, 0.216, 0.358]] 'drawer': [[5.0, 0.258, 0.311], [6.0, 0.193, 0.319], [7.0, 0.227, 0.284], [8.0, 0.219, 0.31]] 'chopping board': [[9.0, 0.266, 0.108], [10.0, 0.203, 0.348], [11.0, 0.153, 0.373], [12.0, 0.045, 0.395], [13.0, 0.035, 0.404]] 'carrot': [[10.0, 0.218, 0.177], [11.0, 0.241, 0.233], [12.0, 0.096, 0.252], [13.0, 0.075, 0.264]]

Response type 1: Description

Question: Can you provide a concise summary of the main events occurring in the video?

Answer: Certainly. The video shows a sequence of kitchen activities, including opening and closing a cupboard and drawer, taking and putting down a knife, and washing a courgette.

Response type 2: Detailed Description

Question: Can you provide a detailed account of the sequence of events that take place between 3.66 and 8.6 seconds in the video, specifically focusing on the actions involving the kitchen drawer and knife? Please include the timing of these actions and describe how the person's hands interact with these objects, as indicated by the object trajectory data.

Answer: Between 3.66 and 8.6 seconds in the video, the individual engages in a sequence of actions involving a kitchen drawer and a knife. Initially, between [3.66, 5.0] seconds, the right-hand reaches for the drawer, with its trajectory indicating a movement slightly right and above the center, coinciding with opening the drawer. Then, the person takes a knife between [5.5, 6.36] seconds, as evidenced by the right hand moving downward and leftward towards the drawer. This action involves a brief re-adjustment, as the person takes the knife, adjusts their grip, and takes it again. There's a moment between [6.73, 7.74] seconds where the knife is put down, suggested by the right hand moving away from the drawer. Finally, the drawer is closed between [8.1, 8.6] seconds, with the right hand's trajectory returning to its original position, indicating the closing motion. Throughout this sequence, the left hand appears to provide support or stabilization, with minor adjustments in its position.

Response type 4: Event Localization

Question: At what timestamps does the individual open and close the drawer?

Answer: The individual opens the drawer between [3.66, 5.0] seconds and closes it between [8.1, 8.6] seconds in the video.

Response type 5: Temporal Reasoning

Question: How did the previous actions of taking vegetables and opening the fridge lead to the current situation in the video?

Answer: Previous actions like taking vegetables and opening the fridge led to the current situation by indicating food preparation activities, culminating in washing the courgette.

Response type 6: Event Localization

Question: Given the ongoing event of washing vegetables, what might be a likely future action sequence in the kitchen, and why?

Answer: Based on the ongoing event of washing vegetables, future actions may include chopping the courgette, possibly using the grater and pan (as indicated in the 'future' events list), suggesting a continuation of food preparation.

Response type 7: Cross-Referencing Events

Question: Can you explain the connection between opening the drawer (3.66-5.0 seconds) and the subsequent use of the chopping board (after 9.0 seconds)?

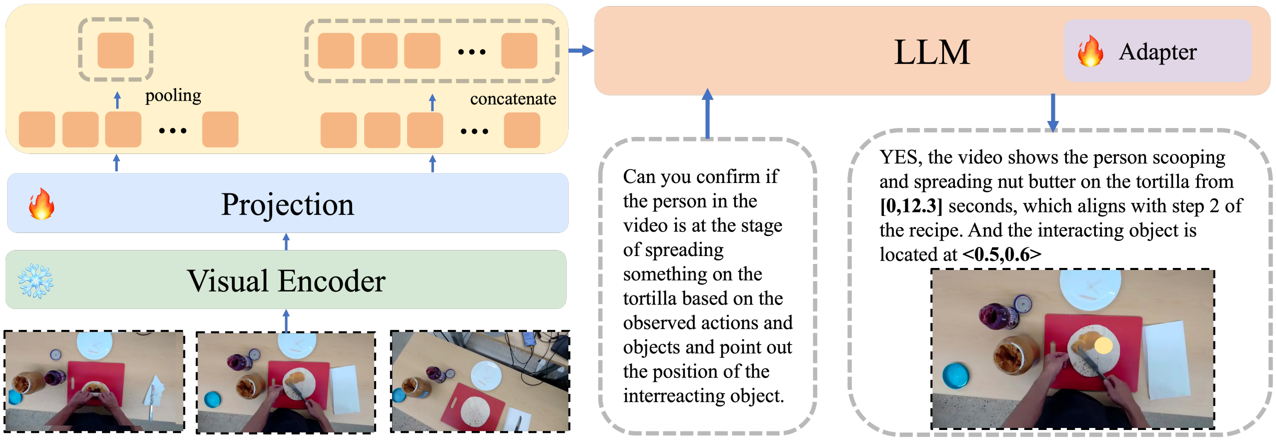
Answer: The opening of the drawer (3.66-5.0 seconds) and the use of the chopping board (after 9.0 seconds) are connected as both actions are part of setting up for the food preparation process; utensils are gathered first (from the drawer), followed by setting up the chopping board for cutting vegetables.

narration. This is based on our observation that longer durations reduce the relevance of actions. In this way, the context helps tasks like action anticipation and detection and encourages the development of new tasks by extrapolating relationships between labels. For instance, our framework enables advanced tasks such as Temporal Reasoning and Cross-Referencing Events, as shown in Table 3, enhancing the dataset's utility without additional annotation effort.

We use two types of symbolic representations to prompt GPT4: (i) Captions, which typically describe the visual scene from various perspectives. (ii) Objects trajectory in the scene, and each box encodes the object concept and its spatial location as shown in Figure 2. We collect 400K unique video instruction-following samples in total, including 350K for activity recognition as shown in Table 3 and 50K for procedure knowledge learning. We have undertaken multiple iterations to refine our method for creating accurate instruction data from task descriptions and object trajectories. We normalized object bounding boxes to a scale of 0-1 and used only the center points of objects, improving the spatial relationships in GPT-4's responses. Additionally, we added a post-processing step that uses interpolation to align GPT-4 output coordinates with actual object trajectories, ensuring high data accuracy. However, including complete trajectories in responses sometimes led to errors. To counter this, we selectively replaced faulty segments with ground truth data, enhancing the dataset's usability. As shown in Table 1, our approach provides longer question-to-clip correspondence than MIMIC-IT [42], focusing on video content comprehension. In contrast, MIMIC [23] often generates questions unrelated to the visual content. Compared to EgoSchema [57], our method emphasizes fine-grained understanding, while EgoSchema targets coarse-grained analysis with few multiple-choice questions for 3-minute video.

4 EAGLE MODEL

Existing image-based MLLMs such as Shikra [12] primarily focus on spatial information, while models like VTimeLLM [30] specifically target temporal dimensions. Given the unique aspects of our dataset, which encompasses both spatial and temporal attributes, our goal is to simplify the tuning process and construct a straightforward yet strong model by leveraging the existing MLLM model.



Can you confirm if the person in the video is at the stage of spreading something on the tortilla based on the observed actions and objects and point out the position of the interreacting object.

YES, the video shows the person scooping and spreading nut butter on the tortilla from **[0,12.3]** seconds, which aligns with step 2 of the recipe. And the interacting object is located at **<0.5,0.6>**

Figure 3: The architecture of the EAGLE model, highlighting the fine-tuned projection layer and adapter, enhances the LLM's ability to process time boundaries and object location tokens.EAGLE is capable of processing natural language queries to determine the temporal boundaries of events and pinpoint the location of the desired object, as denoted by the yellow dot.

Our model, in line with common MLLMs, integrates a vision encoder, an alignment layer, and a large language model (LLM), specifically employing the pre-trained ViT-L/14 from CLIP [70] as the frame encoder $\mathbb{E}$ and Vicuna-13B as LLM, as shown in Figure 3. Given a video sample $V_i \in \mathbb{R}^{T\times H\times W\times C}$ with T frames, the frame encoder $\mathbb{E}$ processes each frame independently, generating video embedding as $x_i \in \mathbb{R}^{T\times D}$.

After obtaining frame embeddings, selecting an optimal method for aggregating these features is critical. Video-LLaMA [103] employs temporal position embedding and a q-former, which typically demands a large amount of paired video-text data (rare in video datasets). Compared with image-language datasets such as CC3M [80] utilized by LLaVA [49], video-language datasets like WebVid [4] contain shorter and less detailed language descriptions. Consequently, when models are pretrained on these video datasets, their expressiveness is often limited, which can result in a less effective image-language alignment layer. Instead, we choose to leverage existing alignment layers from LLAVA to obtain language token from visual feature. We have two streteges, (i) adapt recent advancements [73] and employ an average-pooling strategy to aggregate a video-level representation $v_i \in \mathbb{R}^D$, where D is 5,120 for Vicuna-13B. We donate this model as EAGLE-pool. (ii) Instead of using pooling, we employ alignment layers to extract language tokens directly from each visual frame and concatenate these tokens into a long sequence. This method does not require the explicit embedding of position tokens. Instead, it implicitly incorporates temporal learning, thus leveraging the strengths of the LLAVA alignment layer, which ensures more reliable alignment compared to Q-former aggregation methods.

To enhance the LLM's ability to capture both temporal and spatial information, we integrated Adapters [28] into various self-attention layers of Vicuna-13B, allowing the model to effectively incorporate coordinates from both time boundaries and object trajectories. During training, the visual embedding can be inserted anywhere in the input sequence. Regarding the frame encoder $\mathbb{E}$, we decided to keep the visual encoder frozen throughout all training phases, as fine-tuning the visual encoder even with a small-scale dataset can affect its image representation capabilities and yield performance drop, as discussed in [86].

Followed by [50, 103], the model training is done in two phases. In the first phase, we only focus on fine-tuning the projection layer with a subset of `(VIDEO, INSTRUCTION, RESPONSE)` pairs that do not include time boundary and object trajectory. During the second phase, both the newly integrated Adapters and the projection layer are trained with the entire dataset with 8 NVIDIA A100 GPUs. our model establishes a strong baseline and lays the groundwork for future research into more accurate temporal-spatial grounding abilities and context modeling.

5 EXPERIMENTS

5.1 Evaluation Metrics.

Following the evaluation methods [53, 106] for recent LLMs, we use GPT-4 to assess the quality of responses generated by models. Due to the time-consuming nature of evaluating all 7,700 samples across nine models with GPT-4, we adopt a square root sampling strategy, selecting approximately ($\sqrt{7700} \approx 88$) 100 samples as a representative subset. To deepen our analysis, we further sampled 200 additional responses to evaluate the top four performing models and donate result as EAGLE-pool$_2$ Shikra$_2$ BLIP-2$_2$ and EAGLE$_2$ as shown in Table 5. The results from this extended dataset are presented in the subsequent table and are consistent with the findings from our initial sample of 100 responses.

Given the nature of the egocentric dataset, which offers only action labels, recipe steps, and corresponding timestamps, we need to develop ground truth sentences for evaluation purposes. Our empirical findings indicate that compared to using polished sentences of ground truth labels, template-based construction reduces the occurrence of hallucination errors. The evaluation prompt was refined iteratively through trial and error, aiming to improve the accuracy in identifying event boundaries and objects, and to enhance clarity. The evaluation prompt will be included in the supplementary materials.

Table 4: Video sources and the corresponding number of videos and average actions for training and validation sets.

Video sources	Training set		Validation set		Total
	# videos	# actions (avg)	# videos	# actions (avg)	
EPIC-KITCHENS [16]	16,570 (57%)	4.78	2,901 (38%)	3.98	19,471 (53%)
Ego4D [23]	9,050 (31%)	2.30	3,669 (47%)	2.80	12,719 (35%)
PTA	3,355 (12%)	1.55	1,167 (15%)	1.53	4,522 (12%)
Total	28,975		7,737		36,712

Table 5: We evaluated existing models and our EAGLE model. The scores reflect the models' performance in key aspects, with EAGLE achieving the highest scores in Accuracy and Helpfulness, and competitive scores in other areas.Higher scores indicate better performance.

Model	Accuracy	Helpfulness	Detail	Conciseness	Consistency	Average
Video-LLaMA [102]	1.00	1.00	1.60	1.85	1.43	1.38
LaViLa [105]	1.17	1.15	1.95	4.63	2.73	2.33
BLIP-1 [44]	1.56	1.48	1.85	4.50	3.75	2.63
LLaVA [50]	2.81	2.9	4.56	4.12	3.38	3.55
ImageBind-LLM [25]	2.96	2.97	5.45	4.64	3.71	3.95
InstructBLIP [15]	3.81	3.68	5.29	5.46	4.81	4.61
Shikra [12]	4.21	4.52	6.80	4.78	5.15	5.09
Shikra$_2$	4.31	4.55	6.85	4.20	5.20	5.02
BLIP-2 [43]	4.62	4.78	6.14	5.51	5.53	5.32
BLIP-2$_2$	4.43	4.80	6.20	5.45	5.38	5.25
EAGLE-pool	7.13	7.32	6.52	6.45	6.10	6.70
EAGLE-pool$_2$	7.21	7.40	6.72	6.42	6.30	6.81
EAGLE	**7.32**	**7.51**	**6.90**	**6.75**	6.65	**7.03**
EAGLE$_2$	7.28	7.48	6.83	6.67	**6.77**	7.01

These selected responses will be scored by GPT-4 based on five key metrics, each rated on a scale from 1 to 10, with higher scores indicating superior performance. The evaluation metrics are

1. *Accuracy*: This metric involves assessing if the response reflects the video's content, focusing on activity recognition for EPIC-KITCHENS and Ego4d samples, and the match between predicted and ground truth procedure steps for PTA samples.
2. *Helpfulness*: evaluating how much the response aids in comprehending the video's content and its broader context. It involves assessing whether the model's output provides actionable insights or clarifies complex elements within the video.
3. *Level of Detail*: This involves assessing the comprehensiveness and specificity with which the video is described. A high score in this area indicates that the model captures essential objects and events of the video.
4. *Conciseness*: This metric measures the succinctness and clarity of the response, focusing on delivering essential information without superfluous content. Effective conciseness involves distilling complex information into a clear and brief explanation, which is critical for provide esstial information of the video.
5. *Consistency*: This assesses the uniformity and reliability of the narrative or description provided by the model across multiple instances or parts of the video.

Details of the responses from different models will be included in the supplementary material.

5.2 Baseline Models.

For our baseline models, we use both image-based and video-based approaches. Image-based models include:

1. *BLIP-2* [43] trained a lightweight Q-Former for multimodal representation alignment and vision-to-language generation, capable of following instructions without multimodal instruction tuning.
2. *BLIP-1* [44], pre-trained with web data, using a captioner and filter for synthetic captions, excelling in zero-shot video language tasks.
3. *InstructBLIP* [15], built on BLIP-2, reformats 26 public datasets for instruction tuning, updating only the Q-Former during training.
4. *LaViLa* [105] is a video narration method that pairs a video encoder with a GPT-2 [71] as language decoder and a T-5 [72] to reduce overfitting and enhance natural language data.
5. *LLaVA* [50] introduces visual instruction tuning, using GPT-generated data and instructions for conversation, detailed description, and complex reasoning.
6. *ImageBind-LLM* [25] is an open-source MLLM, with its algorithm details pending publication.
7. *Shikra* [12] encodes regions in natural language as numerical coordinates to specify regions in user queries.
8. *Video-LLaMA* [102] trains adapters for aligning video and audio modalities with LLMs, sampling only eight frames from arbitrarily long videos.

Among baseline models, LaViLa is specifically trained on egocentric videos (Ego4d, EPIC-KITCHEN) to generate narrations. Despite this targeted training, our research reveals that in zero-shot learning scenarios, MLLM, in zero-shot setting, outperformed LaViLa for handling egocentric data.

Additionally, to ensure a fair comparison, we chose not to finetune the vision encoder in our EAGLE model for egocentric vision adaptation. Instead, we focused on refining the model to improve its spatial-temporal video analysis capabilities. Our findings indicate that our dataset significantly contributes to enhancing the performance of current MLLMs in understanding and interpreting video content.

5.3 Results and Analysis

. To validate the performance of EAGLE, we compare it with recent MLLMs [12, 50, 102], on the EAGLE-400K dataset. As Table 5 shows, Shikra and BLIP-2 demonstrate remarkable proficiency, scoring highest in most categories, indicating their reliability, helpfulness, and detailed response capability. Although Video-LLaMA is targeted at video analysis, it exhibits the lowest performance when compared to image-based multimodal large language models (MLLMs), with outputs often arbitrary and failing to capture essential visual information from videos. LLaVA and InstructBLIP demonstrate balanced and above-average performances across all metrics, showcasing their versatility in handling diverse tasks.

Interestingly, while LaViLa is specifically trained on egocentric data, its performance is hindered by its relatively weaker language backbone (GPT-2), resulting in it being outperformed by more advanced MLLMs in a zero-shot setting. This highlights the significant impact that a robust language model can have on performance.

Moreover, ImageBind-LLM excels in providing detailed and consistent responses. This suggests that superior language modeling capabilities, coupled with a more generalized visual encoder, can enhance overall performance significantly.

Comparing the two variants of EAGLE, which utilize different methods for processing video content: Using concatenation of frame features preserves the temporal order of each frame, allowing the model to capture more detailed temporal dynamics and intricate interactions within the video content.EAGLE-pool, on the other hand, employs temporal pooling to aggregate features over time. This approach helps reduce the impact of less relevant information and noise but may also gloss over finer temporal details that are crucial for understanding complex dynamics. Despite these tradeoffs, EAGLE-pool still benefits from the extensive EAGLE-400k dataset and performs better than spatial grounding models like Shikra, which focuses more on spatial rather than temporal data.

These scores provide valuable insights into each model's strengths and weaknesses, allowing for informed decisions on their optimal application areas based on specific needs and criteria.

Ablation Study. Ablation studies were conducted on the EAGLE-400k dataset using varied training data splits to investigate the impact of spatial and temporal information on egocentric video understanding. The ablation included: removing time boundaries (*w/o time*), excluding object trajectories (*w/o obj*), and eliminating both (*only desc*). As shown in Table 6, Performance tends to decrease when either time or object information is excluded, with the least effective results observed when relying solely on descriptions. Surprisingly, PTA exhibits the most significant decline in performance when detailed information is removed, indicating procedure learning relies more on temporal and object details.

Table 6: Ablation study with the different split of dataset

Dataset	EPIC-KITCHEN	Ego4D	PTA
w/o time	5.9	6.1	5.9
w/o object	6.2	6.2	5.8
only desc	5.5	5.8	5.5
all	6.8	6.4	6.5

6 CONCLUSION

In this work, we present the EAGLE-400K dataset and the EAGLE model for holistic egocentric video understanding. The EAGLE-400K dataset consists of 40K question-answer pairs from 36K diverse video clips and EAGLE offers a unified framework for diverse visual computational tasks. We also provide an evaluation method for egocentric vision tasks and demonstrate EAGLE's superior performance. The introduction of a new evaluation metric enhances the understanding of video-based MLLMs. We hope our work can pave the way for augmented reality assistants that aid in complex physical tasks with multimodal perception.

The EAGLE system exhibits a remarkable proficiency in interpreting temporal information from egocentric videos. Despite its impressive capabilities, the system's reliance on human annotation for defining time boundaries and the necessity of a teacher model to generate high-quality reasoning pairs are areas that warrant further exploration. Additionally, the system's limited capacity to identify and track infrequently appearing objects in the dataset is a challenge that needs to be addressed.

Moreover, there's an increased potential for misinformation from model hallucination, where MLLMs might generate plausible but entirely fictitious responses. This can be particularly concerning when models are used to provide feedback or guidance as an AR assistant. The risk is magnified by the model's ability to produce highly realistic and convincing outputs, blurring the line between reality and fiction for users.

7 ETHICS STATEMENT

We must admit that the data collection process in our study may inherently carry a certain degree of bias. In an attempt to mitigate this, we have implemented several measures in our pipeline. Initially, we sourced visual data from EPIC-KITCHEN, Ego4D, and PTA datasets, which are collected from a diverse range of sources and are extensively utilized in various research fields. However, we must also consider that the data annotation phase could potentially introduce additional bias, given its dependence on the prior annotations of the source datasets and GPT-4V. To counteract this, we manually sampled and scrutinized the data quality during the generation process from GPT-4V. In the event of identifying any potential issues, we immediately halt the process for a more in-depth investigation.

REFERENCES

[1] Peri Akiva, Jing Huang, Kevin J Liang, Rama Kovvuri, Xingyu Chen, Matt Feiszli, Kristin Dana, and Tal Hassner. 2023. Self-Supervised Object Detection from Egocentric Videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 5225–5237.

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023).

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023).

[4] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE. https://doi.org/10.1109/iccv48922.2021.00175

[5] Siddhant Bansal, Chetan Arora, and CV Jawahar. 2022. My view is the best view: Procedure learning from egocentric videos. In European Conference on Computer Vision. Springer, 657–675.

[6] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. 2023. Uncertainty-aware State Space Transformer for Egocentric 3D Hand Trajectory Forecasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 13702–13711.

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding?. In ICML, Vol. 2. 4.

[8] Jing Bi, Jiebo Luo, and Chenliang Xu. 2021. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 15611–15620.

[9] Jing Bi, Nguyen Manh Nguyen, Ali Vosoughi, and Chenliang Xu. 2023. MISAR: A Multimodal Instructional System with Augmented Reality. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 1–5.

[10] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, 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 (CVPR).

[11] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. 2023. Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models. arXiv preprint arXiv:2308.13437 (2023).

[12] 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 (2023).

[13] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. 2023. Tracking Anything with Decoupled Video Segmentation. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE. https://doi.org/10.1109/iccv51070.2023.00127

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022).

[15] 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 (2023).

[16] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. International Journal of Computer Vision (IJCV) 130 (2022), 33–55. https://doi.org/10.1007/s11263-021-01531-2

[17] Dima Damen, Oliver Haines, Andrew Calway, Anthony G Cohn, Steven J Vine, and Antonio Criminisi. 2014. You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 3639–3646.

[18] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. 2022. EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. In Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks.

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast networks for video recognition. In Proceedings of the International Conference on Computer Vision (ICCV).

[20] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. 2023. AssistGPT: A General Multi-modal Assistant that can Plan, Execute, Inspect, and Learn. arXiv preprint arXiv:2306.08640 (2023).

[21] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. arXiv preprint arXiv:2304.15010 (2023).

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition. 6904–6913.

[23] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 18995–19012.

[24] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 18995–19012.

[25] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. 2023. Imagebind-llm: Multi-modality instruction tuning. arXiv preprint arXiv:2309.03905 (2023).

[26] Rishi Hazra, Brian Chen, Akshara Rai, Nitin Kamra, and Ruta Desai. 2023. EgoTV: Egocentric Task Verification from Natural Language Task Descriptions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 15417–15429.

[27] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 961–970. https://doi.org/10.1109/CVPR.2015.7298698

[28] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning. PMLR, 2790–2799.

[29] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. arXiv preprint arXiv:2211.09699 (2022).

[30] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2023. VTimeLLM: Empower LLM to Grasp Video Moments. arXiv:2311.18445 [cs.CV]

[31] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. 2023. Egocentric Audio-Visual Object Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 22910–22921.

[32] Mingzhen Huang, Xiaoxing Li, Jun Hu, Honghong Peng, and Siwei Lyu. 2023. Tracking Multiple Deformable Objects in Egocentric Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 1461–1471.

[33] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. 2020. Mutual Context Network for Jointly Estimating Egocentric Gaze and Action. IEEE Transactions on Image Processing 29 (2020), 7795–7806. https://doi.org/10.1109/tip.2020.3007841

[34] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR.

[35] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In International Conference on Learning Representations.

[36] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. 2019. Multitask Learning to Improve Egocentric Action Recognition. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE. https://doi.org/10.1109/iccvw.2019.00540

[37] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017).

[38] Evangelos Kazakos, Silvio Giancola, Long Mai, Dima Damen, Bernhard Rinner, Devis Tuia, Assumpta Parres-Padros, Davide Moltisanti, and Dima Damen.

2019. Epic-Kitchens: A Dataset for Object Recognition in Egocentric Video. In Proceedings of the IEEE International Conference on Computer Vision (ICCV). 3752–3761.

[39] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 787–798.

[40] Hilde Kuehne, Hueihan Jhuang, Estefania Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: A large video database for human motion recognition. In Proceedings of the International Conference on Computer Vision (ICCV).

[41] 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 (2023).

[42] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023. Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425 (2023).

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023).

[44] 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. PMLR, 12888–12900.

[45] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023).

[46] Yin Li, Alireza Fathi, and James M Rehg. 2015. Delving into Egocentric Actions. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 4077–4085.

[47] Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. VideoXum: Cross-modal Visual and Textural Summarization of Videos. arXiv preprint arXiv:2303.12060 (2023).

[48] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. UniVTG: Towards Unified Video-Language Temporal Grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2794–2804.

[49] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023).

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023).

[51] Pinxin Liu, Luchuan Song, Daoan Zhang, Hang Hua, Yunlong Tang, Huaijin Tu, Jiebo Luo, and Chenliang Xu. 2024. Emo-Avatar: Efficient Monocular Video Style Avatar through Texture Rendering. arXiv preprint arXiv:2402.00827 (2024).

[52] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv:2303.05499 [cs.CV]

[53] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634 (2023).

[54] Diogo Luvizon, David Picard, and Hedi Tabia. 2020. Multi-task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020), 1–1. https://doi.org/10.1109/tpami.2020.2976014

[55] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv preprint arXiv:2306.05424 (2023).

[56] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. 2023. EgoLoc: Revisiting 3D Object Localization from Egocentric Videos with Visual Queries. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 45–57.

[57] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. arXiv preprint arXiv:2308.09126 (2023).

[58] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In Conference on Computer Vision and Pattern Recognition (CVPR).

[59] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In 2019 international conference on document analysis and recognition (ICDAR). IEEE, 947–952.

[60] Tirumala Nagarajan, Reza Lehrmann, Anurag Xie, and Naren Ramakrishnan. 2019. Grounded Human-Object Interaction Tasks with Real-World Object Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3628–3638.

[61] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. 2022. Egocentric vision-based action recognition: A survey. Neurocomputing 472 (2022), 175–197.

[62] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. 2023. AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 12999–13008.

[63] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[64] Yulin Pan, Xiangteng He, Biao Gong, Yiliang Lv, Yujun Shen, Yuxin Peng, and Deli Zhao. 2023. Scanning Only Once: An End-to-end Framework for Fast Temporal Grounding in Long Videos. arXiv preprint arXiv:2303.08345 (2023).

[65] 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 (2023).

[66] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2847–2854.

[67] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. 2023. An Outlook into the Future of Egocentric Vision. arXiv:2308.07123 [cs.CV]

[68] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022. Egocentric Video-Language Pretraining. arXiv e-prints (2022), arXiv–2206.

[69] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. 2023. Multimodal Distillation for Egocentric Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 5213–5224.

[70] 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. PMLR, 8748–8763.

[71] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. https://api.semanticscholar.org/CorpusID:160025533

[72] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21, 1 (2020), 5485–5551.

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

[74] Nicholas Rhinehart and Kris M Kitani. 2017. First-person activity forecasting with online inverse reinforcement learning. In Proceedings of the IEEE International Conference on Computer Vision. 3696–3705.

[75] Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. arXiv:1706.05098 [cs.LG]

[76] Fiona Ryan, Hao Jiang, Abhinav Shukla, James M. Rehg, and Vamsi Krishna Ithapu. 2023. Egocentric Auditory Attention Localization in Conversations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 14663–14674.

[77] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In European Conference on Computer Vision. Springer, 146–162.

[78] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. https://doi.org/10.1109/cvpr52688.2022.02042

[79] Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal video summarization via time-aware transformers. In Proceedings of the 29th ACM International Conference on Multimedia. 1756–1765.

[80] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2556–2565. https://doi.org/10.18653/v1/P18-1238

[81] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, 742–758.

[82] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[83] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. 2023. Video Understanding with Large Language Models: A Survey. arXiv:2312.17432 [cs.CV]

[84] Yunlong Tang, Siting Xu, Teng Wang, Qin Lin, Qinglin Lu, and Feng Zheng. 2022. Multi-modal Segment Assemblage Network for Ad Video Editing with Importance-Coherence Reward. In Proceedings of the Asian Conference on Computer Vision (ACCV). 3519–3535.

[85] Yunlong Tang, Jinrui Zhang, Xiangchen Wang, Teng Wang, and Feng Zheng. 2023. LLMVA-GEBC: Large Language Model with Video Adapter for Generic Event Boundary Captioning. arXiv preprint arXiv:2306.10354 (2023).

[86] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. 2023. What Makes for Good Visual Tokenizers for Large Language Models? arXiv preprint arXiv:2305.12223 (2023).

[87] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. 2023. Ego-Only: Egocentric Action Detection without Exocentric Transferring. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 5250–5261.

[88] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023. All in one: Exploring unified video-language pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6598–6608.

[89] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. 2023. Scene-Aware Egocentric 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 13031–13040.

[90] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. 2023. Caption anything: Interactive image description with diverse multimodal controls. arXiv preprint arXiv:2305.02677 (2023).

[91] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-End Dense Video Captioning With Parallel Decoding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 6847–6857.

[92] Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. 2022. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In European Conference on Computer Vision. Springer, 709–725.

[93] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. 2023. Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition From Egocentric RGB Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 21243–21253.

[94] Jinheng Xie, Kai Ye, Yudong Li, Yuexiang Li, Kevin Qinghong Lin, Yefeng Zheng, Linlin Shen, and Mike Zheng Shou. 2023. VisorGPT: Learning Visual Prior via Generative Pre-Training. arXiv preprint arXiv:2305.13777 (2023).

[95] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 (2021).

[96] Mengmeng Xu, Yanghao Li, Cheng-Yang Fu, Bernard Ghanem, Tao Xiang, and Juan-Manuel Pérez-Rúa. 2023. Where Is My Wallet? Modeling Object Proposal Sets for Egocentric Visual Query Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2593–2603.

[97] Yue Xu, Yong-Lu Li, Zhemin Huang, Michael Xu Liu, Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. 2023. EgoPCA: A New Framework for Egocentric Hand-Object Interaction Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 5273–5284.

[98] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. 2023. Pink: Unveiling the Power of Referential Comprehension for Multi-modal LLMs. arXiv preprint arXiv:2310.00582 (2023).

[99] Zihui Xue, Yale Song, Kristen Grauman, and Lorenzo Torresani. 2023. Egocentric Video Task Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2310–2320.

[100] Wenpeng Yin, Qinyuan Ye, Pengfei Liu, Xiang Ren, and Hinrich Schütze. 2023. LLM-driven Instruction Following: Progresses and Concerns. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, Qi Zhang and Hassan Sajjad (Eds.). Association for Computational Linguistics, Singapore, 19–25. https://doi.org/10.18653/v1/2023.emnlp-tutorial.4

[101] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. CoCoT: Contrastive Chain-of-Thought Prompting for Large Multimodal Models with Multiple Image Inputs. arXiv preprint arXiv:2401.02582 (2024).

[102] 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 (2023).

[103] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv:2306.02858 [cs.CL]

[104] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. 2023. Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023).

[105] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning Video Representations from Large Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6586–6597.

[106] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]

[107] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2015. Temporal action localization in untrimmed videos via multi-stage CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1049–1058.

[108] Chenchen Zhu, Fanyi Xiao, Andres Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. 2023. EgoObjects: A Large-Scale Egocentric Dataset for Fine-Grained Object Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 20110–20120.

[109] 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 (2023).