OSCaR: Object State Captioning and State Change Representation

Nguyen Nguyen[1], **Jing Bi**[1], **Ali Vosoughi**[1], **Yapeng Tian**[2], **Pooyan Fazli**[3], **Chenliang Xu**[1]
[1]University of Rochester, [2]University of Texas at Dallas, [3]Arizona State University
`{nguyen.nguyen, jing.bi, ali.vosoughi, chenliang.xu}@rochester.edu`,
`yapeng.tian@utdallas.edu`, `pooyan@asu.edu`

Abstract

The capability of intelligent models to extrapolate and comprehend changes in object states is a crucial yet demanding aspect of AI research, particularly through the lens of human interaction in real-world settings. This task involves describing complex visual environments, identifying active objects, and interpreting their changes as conveyed through language. Traditional methods, which isolate object captioning and state change detection, offer a limited view of dynamic environments. Moreover, relying on a small set of symbolic words to represent changes has restricted the expressiveness of language. To address these challenges, in this paper, we introduce the Object State Captioning and State Change Representation (OSCaR) dataset and benchmark. OSCaR consists of 14,084 annotated video segments with nearly 1,000 unique objects from various egocentric video collections. It sets a new testbed for evaluating Multimodal Large Language Models (MLLMs). Our experiments demonstrate that while MLLMs show some skill, they lack a full understanding of object state changes. The benchmark includes a fine-tuned model that, despite initial capabilities, requires significant improvements in accuracy and generalization ability for effective understanding of these changes. Our code and dataset are available at `https://github.com/nguyennm1024/OSCaR`.

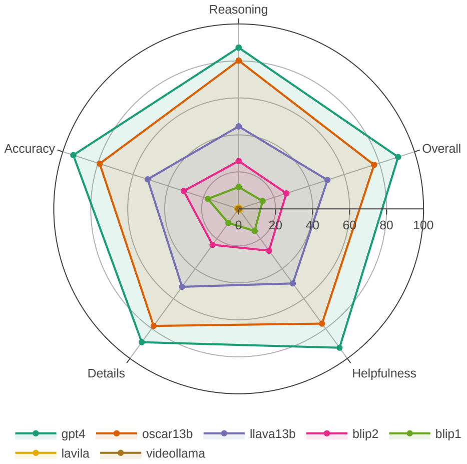


Figure 1: **Surpassing prior models in aligning with human judgements.** Our method achieves near parity with GPT-4V ratings across helpfulness, accuracy, reasoning, and other key metrics.

1 Introduction

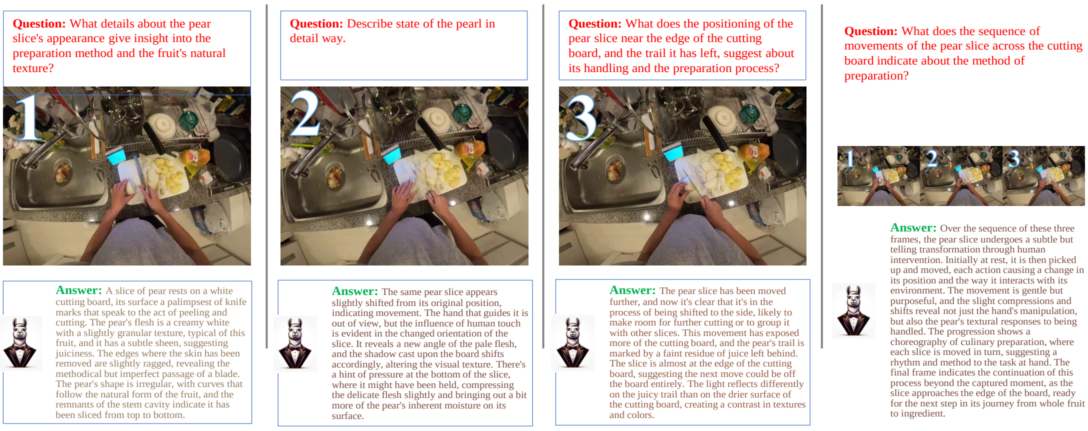
The field of Natural Language Processing (NLP) has evolved beyond mere text interpretation and generation, advancing into realms where understanding and interacting with the physical world becomes imperative. From studying causal reasoning (Gao et al., 2018) to building a world model for cause-effect prediction (Gao et al., 2016; Alayrac et al., 2017), researchers have been working on the problem of causation in the physical world.

In this paper, we investigate the very basic causal relations between a concrete action and the change of the object state caused by this action. For example, given an image as shown in Figure 2, we, as humans, would have no problem understanding which object is being actively interacted with. Furthermore, given the statement "cutting the bread", we would naturally imagine what state change may happen. However, Despite tremendous progress in knowledge representation, automated reasoning, and machine learning, artificial agents still lack the understanding of naive causal relations regarding the physical world (Gao et al., 2018).

Imagining a scenario where artificial agents collaborate with humans in the physical world, they will need to understand the physical action effect to reason, learn, and assist humans (Bi et al., 2023). To empower machines with such capabilities, this paper introduces a novel benchmark focusing on understanding object state changes from egocentric visual inputs, which has the advantage of the lens of human eyes.

Understanding object state change is not only a complex task but also practical and foundational

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3565–3576
June 16-21, 2024 ©2024 Association for Computational Linguistics



Question: What details about the pear slice's appearance give insight into the preparation method and the fruit's natural texture?

Answer: A slice of pear rests on a white cutting board, its surface a palimpsest of knife marks that speak to the act of peeling and cutting. The pear's flesh is a creamy white with a slightly granular texture, typical of this fruit, and it has a subtle sheen, suggesting juiciness. The edges where the skin has been removed are slightly ragged, revealing the methodical but imperfect passage of a blade. The pear's shape is irregular, with curves that follow the natural form of the fruit, and the remnants of the stem cavity indicate it has been sliced from top to bottom.

Question: Describe state of the pearl in detail way.

Answer: The same pear slice appears slightly shifted from its original position, indicating movement. The hand that guides it is out of view, but the influence of human touch is evident in the changed orientation of the slice. It reveals a new angle of the pale flesh, and the shadow cast upon the board shifts accordingly, altering the visual texture. There's a hint of pressure at the bottom of the slice, where it might have been held, compressing the delicate flesh slightly and bringing out a bit more of the pear's inherent moisture on its surface.

Question: What does the positioning of the pear slice near the edge of the cutting board, and the trail it has left, suggest about its handling and the preparation process?

Answer: The pear slice has been moved further, and now it's clear that it's in the process of being shifted to the side, likely to make room for further cutting or to group it with other slices. This movement has exposed more of the cutting board, and the pear's trail is marked by a faint residue of juice left behind. The slice is almost at the edge of the cutting board, suggesting the next move could be off the board entirely. The light reflects differently on the juicy trail than on the drier surface of the cutting board, creating a contrast in textures and colors.

Question: What does the sequence of movements of the pear slice across the cutting board indicate about the method of preparation?

Answer: Over the sequence of these three frames, the pear slice undergoes a subtle but telling transformation through human intervention. Initially at rest, it is then picked up and moved, each action causing a change in its position and the way it interacts with its environment. The movement is gentle but purposeful, and the slight compressions and shifts reveal not just the hand's manipulation, but also the pear's textural responses to being handled. The progression shows a choreography of culinary preparation, where each slice is moved in turn, suggesting a rhythm and method to the task at hand. The final frame indicates the continuation of this process beyond the captured moment, as the slice approaches the edge of the board, ready for the next step in its journey from whole fruit to ingredient.

Figure 2: **OSCaR's description of state, state change, and illustration of reasoning.** State description involves the characterization of a specific region of interest within the video and the associated activity. State change entails the description of the evolution of a system over a defined temporal sequence. Furthermore, the analysis of the state of an object is centered on comprehending and elucidating the mechanisms underlying the object's evolution.

for many other tasks, such as helping intelligent agents to understand the environment dynamics and complete task (Padmakumar et al., 2023; Sarch et al., 2023; Merullo et al., 2022), tracking the state of dialog(Le et al., 2022), creating causal graphs for knowledge representation for complex question and answering (Ates et al., 2020).

Modeling object state change requires two abilities: 1) scene understanding, which involves parsing the world through an object-centric lens, and 2) causal-effect understanding, which entails identifying likely actions and their effects by observing images before, during, and after an action.

Previous research efforts have concentrated on building symbolic representations to ground changes and states (Wu et al., 2023; Zellers et al., 2021; Nagarajan and Grauman, 2018). However, given the diversity and complexity of objects and their states, influenced by contextual and temporal factors, symbolic representation alone falls short. This paper proposes the use of natural language as a more expressive and intuitive medium for this task. This approach not only aligns the understanding of visual content between humans and AI systems but also enhances communication between them, providing a richer context than unimodal models.

Essentially, we form the scene understanding as an object-centric visual captioning problem. We can utilize natural language to describe the objects and any changes that may occur. On the other hand, the ability to understand the causal effect is formed as a visual question-answering problem based on 3 images: before, during, and after the action. Our dataset and experiments exhibit considerable potential for scalable application across various domains in future research. While conducting this study, another research was also conducted to understand object state change with a different approach (Xue et al., 2024). That shows the importance and significant potential of this research direction.

In summary, our contributions are threefold:

- We introduce a new problem to understand states and state changes of object through natural language.
- We present a method to generate good-quality visual instructions guided by simple annotations, applicable to both images and videos, advancing future research in visual instruction tuning. Our pipeline provides a good starting point for the data collection process.
- Our paper introduces OSCaR, a novel dataset and a benchmark leveraged by the power of GPT-4V that contains different tasks for object state understanding, including visual captioning, visual question answering visual dialog, and reasoning.

2 Related Works

Object state change: Localizing and recognizing changes of object states, play a key role in applications such as procedural planning (Bi et al., 2021), robotics, and video action understanding (Du et al., 2023; Zhong et al., 2023; Tang et al., 2023b; Wang et al., 2023; Song et al.). Recognizing object state changes necessitates the joint discovery of states and actions through an understanding of their causal relationship, as discussed in prior works (Alayrac et al., 2017; Liu et al., 2017; Souček et al., 2022; Naeem et al., 2021). Recently, a self-supervised method has been proposed to jointly localize action and state changes temporally from noisy untrimmed long videos (Souček et al., 2022). Moreover, (Saini et al., 2023) introduces a novel benchmark for the generation of object states, yet their focus is very limited to only the cutting action and a small dataset. However, previous studies often separate scene understanding from object state change recognition and tend to operate under a closed-world assumption, which limits their applicability in real-world scenarios. Our research aims to bridge the gap between human and machine perception by integrating egocentric views and language.

Multimodal Large Language Models: Recent advancements in Large Language Models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023; Chiang et al., 2023; Chung et al., 2022) have led to significant achievements in language understanding and generation. This progress has sparked an interest in the creation of MLLMs that blend the advanced linguistic processing of LLMs with capabilities for multi-modal perception (Zhang et al., 2023a; Ye et al., 2023; Li et al., 2023a; Gao et al., 2023; Peng et al., 2023; Tang et al., 2023a). The core of this research is the fusion of pre-training visual encoder representations with the input embedding space of LLMs, achieved by pretraining with datasets that interleave images and text. (Li et al., 2023c; Zhu et al., 2023a; Liu et al., 2023a). In this paper, we aim to provide a comprehensive evaluation of these models, particularly focusing on their performance in object state change recognition.

3 The OSCaR Dataset

This section outlines our pipeline for creating visual instructions on object states. We begin with the process of collecting diverse visual data from public sources, detailed in section 3.1. Following this, section 3.2 describes our approach to enhancing data quality using simple human annotations across various tasks, facilitating a deeper understanding of object states. Our method enables the generation of detailed captions, visual question answering, and visual dialogue.

3.1 Video Collections

OSCaR is a curated compilation of videos sourced from two distinct datasets: EPIC-KITCHENS (Damen et al., 2018) and Ego4D (Grauman et al., 2021). Acknowledging that changes in object states occur progressively over time rather than abruptly within a single frame, we have selectively included video clips that effectively illustrate these state transitions. Our selection process ensures that these videos depict the dynamic changes in object states and capture moments where the objects remain stationary for short enough durations. This approach enabled us to compile a comprehensive visual dataset encompassing the object's static and transitional states.

We initially analyzed the verbs from the original videos of the EPIC-KITCHENS dataset to ensure that the videos highlighted objects undergoing state changes. We categorized these verbs into three groups: *change*, *not sure*, and *not change*. The *change* group consists of verbs likely to alter the state of objects, including actions like Open, Close, Wash, Cut, and Mix. Conversely, the *not change* group encompasses verbs with a minimal likelihood of inducing state changes, such as Take, Put, Move, Check, etc. Lastly, the *not sure* group includes verbs with ambiguous potential for state change, covering actions like Shake, Flip, Use, Pull, and others. After filtering the EPIC-KITCHENS dataset, we were able to identify 69 verb classes that consisted of a total of 650 verbs. Using this verb list, we retrieved all video segments containing those actions.

Upon analyzing the videos, we discovered that some objects only appeared in a few times. As a result, we split the videos into two groups. The first group comprises videos that focus on objects that occurred more than ten times, and it will be used to construct our training and testing set. The second group includes videos with objects that occurred less than ten times. These objects are rare in EPIC-KITCHENS and can be used for open-world evaluation, which will be discussed in section 4.2. In the first group, we randomly selected 10 to 50 video segments per object, resulting in 7442 with 306 different objects from EPIC-KITCHENS.

We leveraged Ego4D, the largest egocentric video dataset, selecting video segments tagged with "$object\_of\_change$" to enhance our data's diversity. This tag highlighted videos showcasing object state changes. By gathering these specific videos, along with details of the objects and their narrations, we informed our data generation and compiled relevant statistics. From this dataset, we extracted 5942 segments featuring 296 unique objects for our OSCaR project.

3.2 GPT-assisted Data Generation

Caption Generation: Captioning plays an important role in visual understanding. Understanding object states requires detailed and informative captions to capture the exact state of objects. To achieve this goal, we generated captions for all collected videos by leveraging GPT-4V and human's weak annotations. This problem requires two types of annotations, including 1) Start and end frame ID in videos during the event to make state changes and 2) A short description of what happens in the video. The short description can be a verb representing the action and a noun representing the object humans interact with (e.g., washing tray). We designed adaptive prompts to inject this annotation as context to guide GPT-4V to generate high-quality captions. We found that GPT-4V often suffers from ambiguity without this guidance, and the quality of generated captions is degraded. With simple human guidance, GPT-4V can reduce ambiguity and produce better-quality captions.

Multiple-choice QA Generation: The multiple-choice question is a method of presenting a set of answers, including incorrect options, to teach machine learning models how to distinguish between correct and incorrect answers. This type of question can also be used as a form of instruction, where the question serves as the prompt, and the answer serves as the response for the models. We created multiple-choice question and answer sets based on generated captions.

Conversation Generation: Visual dialog is a complex task requiring understanding of visual content and conversation context, and faces challenges in data collection due to its need for natural dialogues between two people viewing the same content. This process is time-consuming and resource-intensive, especially when involving reasoning and explanations. With the growth of machine learning models, generating visual dialog data is increasingly vital. We've developed a method that uses captions to create visual conversation data, enhanced by GPT-4V's ability to provide explanations, offering flexible and diverse data. This approach, labeling input data for images and videos, is cost-effective and faster than manual methods, generating vast amounts of training data for future models.

4 OSCaR Benchmarks

4.1 Evaluation with Text Generation Metrics

The dataset we are providing consists of 500 videos from the Ego4D and EPIC-KITCHENS datasets, which are specifically designed for benchmarking purposes. Each video is annotated by four detailed captions, all of which have undergone rigorous human verification to ensure the quality and reliability of this evaluation set. To ensure a comprehensive and accurate assessment of performance, text generation metrics such as BLEU, Rouge, LSA, among others, can be used for evaluation purposes.

4.2 Open-world Object State Understanding

Collecting data for all objects worldwide and then training models is not feasible. However, humans can describe new or unfamiliar objects, which can be challenging for AI, especially when they are in a new domain or serve a different purpose. Fortunately, recent achievements in MLLMs have opened up the potential for AI to have this ability. During pre-training with large amounts of data, MLLMs can learn general knowledge about the world. Besides, models will learn how to perform tasks during the visual instruction tuning process. In both processes, the models may or may not have been exposed to objects not in the object state understanding training set. The question is whether models can generalize to objects of this type. To answer this question, we provide two evaluation sets to test the generalizability of the models.

Cooking domain objects have not occurred in the training set for object state understanding: For this evaluation, we want to investigate the model's ability to understand objects that have not appeared in the training set in a similar scenario with the training domain. We provided a set of 2,485 videos with 1,024 objects that have not occurred in the object state training set. This testing set will evaluate how in-domain knowledge can help models understand object states and state changes. We used GPT-4V to annotate 344 videos for evaluation purposes.

Out-of-domain objects state understanding: This evaluation focuses on judging the ability of models to understand objects beyond the training domains. Our training set contains only the cooking domain data, while this testing set has diverse domains, such as baker, household management, cleaning/laundry, bike mechanic, etc. This set was extracted from the Ego4D dataset and contains 43,367 videos with more than 500 objects. This testing not only can be used for evaluation but also has the potential to scale up using our pipeline for object state understanding in other specific domains. For this evaluation set, we selected 10 videos from each of the 51 different domains, totaling 356 videos. Domains with fewer than 10 videos have all their videos included. This set is also annotated by GPT-4V.

4.3 Data Quality Verification

We evaluated the quality of descriptions for object states and activities across video frames using Amazon MTurk for human feedback. Our assessment framework included five guidelines for spotting inaccuracies, focusing on frame-specific description accuracy, two for assessing state change accuracy, two for identifying hallucinations, and three for recognizing incomplete descriptions. Annotators were asked to categorize each description under one of four labels: 1) Fully Detailed and Comprehensive, 2) Generally Complete with Minor Omissions, 3) Lacks Important Details or Contains Errors, or 4) Incomplete, Misleading, or Hallucinating, and provide reasoning to discourage random responses. This study utilized 500 samples from the EPIC-KITCHENS and Ego4D datasets, leading to the validation of 2000 natural language descriptions.

5 Data Statistics

In order to help models generate concise and informative answers, we have defined short answers as those with less than ten words and long answers as those with more than ten words. Short answers provide brevity, while long answers offer detailed and informative information. The distribution of these two types of answers can be seen in Figure 3. The average answer length in the dataset is 47.06 words. Long answers make up about 75% of the data, with an average length of 63 words, while short answers account for about 25% of the data, with an average length of 3.32 words. By splitting the data accordingly, future models can provide short, direct, and informative answers with explanations. To showcase the uniqueness of our OSCaR dataset, we have presented a comparison between OSCaR and other related datasets in Table 1. The OSCaR dataset comprises a vast number of instructions, along with images and videos. Additionally, it also provides data for object state captioning and object state change captioning.

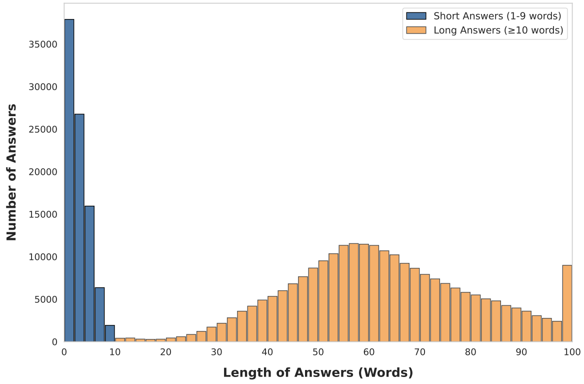


Figure 3: **Distribution of answer lengths.** The figure shows how answers are distributed by length in the dataset. It separates short answers (1-9 words) from long answers ($\geq$ 10 words). The histogram displays the number of answers on the y-axis based on increasing answer lengths on the x-axis. There is a category at 100 words for answers with lengths greater than or equal to 100 words. This breakdown emphasizes the balance between brief, direct answers and more detailed, explanatory responses.

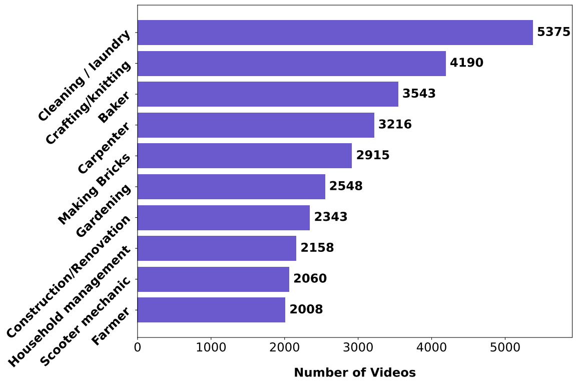


Figure 4: **Top 10 open-world domains (excluding cooking)**. The figure shows non-cooking domains present in the open-world test set used to assess model generalization. By evaluating performance on household and occupational activities unseen during training, we benchmark the trained models' capacity to understand new objects and actions beyond cooking tasks.

In section 4.2, we discussed two types of open-world datasets for object state understanding: in-domain cooking and open domains. Although we trained on videos with object state changes, in open-world evaluation, we tested the models on both types of videos, with and without object state changes, to ensure their generalizability. The in-domain evaluation set consists of 2,485 videos with 1,024 novel objects extracted from EPIC-KITCHENS.

We have extracted an open-domain evaluation set from the Ego4D dataset. The top 10 most frequent domains in the open-world testing set are shown in Figure 4. This evaluation set from 51 different domains, contains annotations for domain, action, object name, and action narrations extracted from annotations of Ego4D. The set includes 43,367 open-world videos for which we know their domains and 56,231 videos of unknown domains, but we still have information about their object names and action narrations. Thus, this set can be utilized not only for open-world evaluation but also for the advancement of general domain object-state understanding in the future when applying our method to generate labels. This set of data hasn't been annotated, but the data we extracted from Ego4D are ready to use our pipeline to scale up the data generation.

6 Experiments

In this section, we will discuss the experimental design we used and how we trained our model. Our fine-tuning process will be described in Section 6.1. Additionally, we included other vision language models such as BLIP (Li et al., 2023b) , BLIP2 (Li et al., 2023c), LaViLa (Zhao et al., 2022), and Video-LLaMA (Zhang et al., 2023b) for comparison purposes. Firstly, we will evaluate our model's performance in the cooking domain in Section 6.3. After that, we will also evaluate its performance in an open-world setting in Section 6.4.

6.1 Model Training

We conducted extensive experiments to showcase the effectiveness of our data generation pipeline in solving object-state understanding problems. A straightforward approach to solving these types of problems is using a model with a text encoder to encode prompts and a visual encoder to encode visual content. After that, both of these inputs will be used as conditions to generate text answers with a text decoder. Ideally, this text decoder will be an LLM.

We fine-tuned LLaVA, an open-source MLLM featuring capabilities like visual dialogue, question-answering (Agrawal et al., 2015), and OCR (Nguyen et al., 2021, 2024), to achieve our goals. Notably, the generated data can enhance any future vision-language models beyond LLaVA.

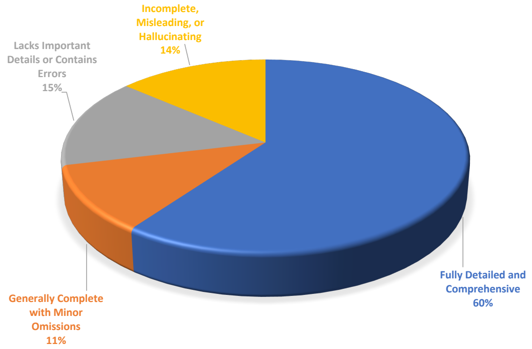


Figure 5: **GPT-4V zero-shot caption quality human evaluation.** The figure shows the distribution of quality ratings assigned by human annotators evaluating frame descriptions automatically generated by the GPT-4V model under zero-shot conditions. Descriptions for 500 video frames were rated.

We experimented with LLaVA using Vicuna 7B and 13B models under two conditions: with and without its original visual instruction tuning data, referring to the former as OSCaR.

For training, we employed Lora fine-tuning with a configuration of rank 128 and alpha 256, using Vicuna 13B and 7B models alongside the OpenAI/CLIP-ViT-Large-Patch14-336 vision encoder. A projector transformed visual features into tokens. Our fine-tuning parameters included a single epoch, a learning rate of 2e-4, a batch size of 16 per device, and a maximum model length of 2048.

6.2 Evaluating GPT-4V

Because our pipeline uses GPT-4V as the knowledge model to annotate our data, evaluating GPT-4V's ability is crucial. Evaluating GPT-4V's performance has two purposes: 1) Understanding the performance of GPT-4V on this task and 2) Producing a clean benchmark beyond the ability of GPT-4V for future research. As discussed in section 4.3, we ask humans to check data quality and classify quality into four levels with text explanation. Figure 5 shows the distribution of data quality from 500 videos sampled from the dataset for benchmarking.

6.3 Evaluation on Cooking Domain Objects

Text Generation Metrics Evaluation: The table 2 in this document displays the results of two text generation metrics, BLEU and ROUGE. As per the table, LaViLa and BLIP1 models have scored very low, whereas BLIP2, Video-LLaMA, and LLaVA models, which are currently the most advanced models, have achieved significant improvements.

Table 1: Comparison of OSCaR dataset versus other related datasets. OSC and OSCC represented for Object State Captioning and Object State Change Captioning, respectively.

Dataset	**Video**	**#Clip**	**#Instruction**	**OSC**	**OSCC**
MiniGPT-4 (Zhu et al., 2023b)	✗	✗	5K	✗	✗
Shikra-RD (Chen et al., 2023)	✗	✗	5.9K	✗	✗
LLaVA (Liu et al., 2023b)	✗	✗	345K	✗	✗
VideoChat (Li et al., 2023d)	✓	11K	20.8K	✗	✗
OSCaR	✓	**18K**	**400K**	✓	✓

Table 2: **Performance comparison based on BLEU and ROUGE scores.** OSCaR is LLaVA fine-tuned with OSCaR data, mixed data is a combination of LLaVA data and OSCaR data.

Model	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
LaViLa (Zhao et al., 2022)	0.006	3.3	0.26	3.27
BLIP1 (Li et al., 2022)	0.008	1.38	0.08	1.35
BLIP2 (Li et al., 2023c)	0.1	11.53	2.12	10.51
Video-LLaMA (Zhang et al., 2023b)	1.0	17.75	2.69	16.02
LLaVA v1.5 13B (Liu et al., 2023b)	3.72	27.09	6.59	24.01
LLaVA v1.5 7B (Liu et al., 2023b)	3.23	25.37	6.22	22.60
OSCaR 13B (OSCaR data only) (Ours)	5.28	27.93	7.67	24.45
OSCaR 7B (OSCaR data only) (Ours)	5.1	28.27	7.42	24.77
OSCaR 13B (Mixed data) (Ours)	5.76	29.26	8.24	25.78
OSCaR 7B (Mixed data) (Ours)	**5.79**	**29.94**	**8.34**	**26.24**

Our proposal has surpassed every previous state-of-the-art model by a large margin on these metrics.

GPT4 Evaluation: The experimental results of evaluating LLaVA, OSCaR, and GPT-4V captions on five criteria using GPT-4V are shown in Table 4. According to the metric used, OSCaR performs significantly better than LLaVA. Additionally, OSCaR achieved 88.19%, 87.01%, 90.81%, 89.21%, and 97.94% in accuracy, helpfulness, detail level, reasoning, and overall, respectively, compared to GPT-4V. On average, OSCaR is **90%** as good as GPT-4V. The visualization can be seen at Figure 1.

Human Study: In our study to assess caption quality from various models, seven evaluators reviewed five videos with four captions each (three for frames, one for state changes), provided by seven models. Each caption had seven different options generated by seven different models. Evaluators could select up to two options per caption that they think are the best. Figure 6 shows the results of this experiment. We calculated the percentage of times each model was selected and found that OSCaR achieved 73.93%, which was only 8.57% lower than GPT-4V. OSCaR significantly outperformed LLaVA by more than two times. These results demonstrate that OSCaR is a promising model for generating high-quality captions.

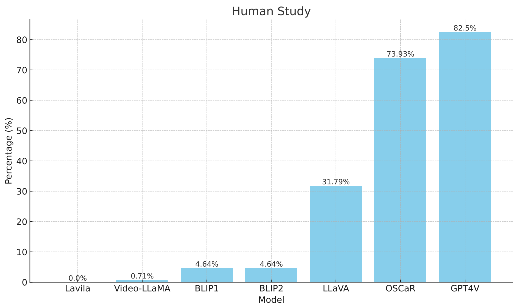


Figure 6: **Human study results.** The figure shows the percentage that each model was selected by participants as producing favorable descriptions in a human rating study.

6.4 Open-world Objects Evaluation

Evaluating the performance of machine learning models solely based on objects seen during training isn't enough. To more thoroughly test their effectiveness, we also evaluated them on objects not included in the training set, representing the open world. In this part of our study, we compare the quality of text produced by our model and GPT-4V for these open-world objects, using BLEU and ROUGE scores as our metrics.

In-domain Objects Evaluation: The evaluation results on objects in the cooking domain that were not included in the instruction fine-tuning data are presented in Table 3. When compared with the results in Table 2, the overall performance is better when testing with in-domain open-world objects. One of the reasons for this is that the evaluation set in Table 2 was corrected by humans, while the data used in Table 3 was generated

Table 3: **Open-world performance comparison based on BLEU and ROUGE scores.** OSCaR is LLaVA fine-tuned with OSCaR data, mixed data is a combination of LLaVA data and OSCaR data.

Open World	**Model**	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
In Domain	OSCaR 13B (OSCaR data only)	5.86	28.64	8.43	24.91
	OSCaR 7B (OSCaR data only)	5.73	29.10	8.38	25.47
	OSCaR 13B (Mixed data)	**6.19**	29.36	8.74	25.69
	OSCaR 7B (Mixed data)	6.13	**30.00**	**8.95**	**26.25**
Out of Domain	OSCaR 13B (OSCaR data only)	5.32	27.20	7.62	23.67
	OSCaR 7B (OSCaR data only)	5.18	27.07	7.50	23.65
	OSCaR 13B (Mixed data)	5.24	26.18	7.36	23.09
	OSCaR 7B (Mixed data)	**5.69**	**28.99**	**8.29**	**25.38**

Table 4: Evaluation scores using GPT-4V under different criterion are listed in the table.

Criteria	LLaVA	OSCaR	GPT-4V
Accuracy	53.60	82.93	94.04
Helpfulness	51.63	80.78	92.83
Reasoning	53.64	79.20	87.22
Detail	40.56	87.30	89.14
Overall	51.96	80.92	90.72

from GPT-4V. Nevertheless, the outcomes of this experiment indicate the generalizability of models when dealing with new objects.

Objects Beyond Cooking Domain: Table 3 presents the open-world evaluation for various domains. The dataset employed in this experiment is discussed in section 4.2, which comprises 356 videos from 51 distinct domains. Compared to the experiment in table 2, the outcomes of this experiment are generally lower. Specifically, for LLaVA 7B with mixed data, this experiment shows a decline of 0.1, 0.95, 0.05, and 0.85 on BLEU, ROUGE-1, ROUGE-2, and ROUGE-L, respectively. This decline indicates two things: 1) the open domain is challenging and may require domain-specific data for fine-tuning to achieve better performance, and 2) even in the absence of new domain data, the decrease in performance is not too significant, and showing the generalizability of our model.

6.5 Ablation Study

Our research also examined the accuracy of video frame annotations in the EPIC-KITCHENS and Ego4D datasets. We used Amazon Mechanical Turk annotators to evaluate 500 video data points for the precision and completeness of descriptions, categorizing them into four classes. In addition, we analyzed 100 samples from each setting of zero-shot and two-shot to determine the best strategy for scaling up data annotation. Our findings indicate that zero-shot is the more effective approach for annotating our task's data.

Our findings, detailed in Table 5, compare zero-shot and two-shot performance in aligning descriptions with human standards of accuracy and relevance, as derived from video frame analyses. This table illustrates how well the GPT-4V model's natural language descriptions, evaluated by Amazon Mechanical Turk annotators in zero and two-shot scenarios, match human judgment. The percentages indicate the extent to which these descriptions accurately and relevantly depict the video content, based on a frame-by-frame review. Each description was judged for its thoroughness and relevance in detailing the object and its activities. Annotators followed established guidelines to determine the quality of data in their assessments.

The results reveal a notable disparity in description quality between the zero-shot and two-shot methods. The zero-shot approach yielded a higher proportion of Fully Detailed and Comprehensive descriptions, while the two-shots method indicated a greater occurrence of descriptions with errors or misleading content. This variation highlights the differences in data quality and annotator perceptions under varying evaluation conditions, underscoring the importance of method selection in annotation studies.

Table 5: The table lists the distribution of Amazon Mechanical Turk annotators' choices of descriptions of objects and object state changes in 0 and two-shot tests by the GPT-4V model in %.

Satisfaction Class	Zero-shot	Two-shots
Fully Detailed	56.25	33.25
Minor Mistakes	16.75	28.25
Lacks Important Details	13.25	23.00
Hallucinating	13.75	15.50

In Table 6, we present the results of our experiment where we evaluate various models in open-

Table 6: **Performance comparison based on BLEU and ROUGE scores in different domains.** The table compares various models with open-world benchmarks.

Domain	Method	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Cooking Domain	LLaVA (Liu et al., 2023b)	2.56	23.77	5.64	21.08
	BLIP1 (Li et al., 2022)	4.33×10^{-5}	0.75	0.026	0.73
	BLIP2 (Li et al., 2023c)	0.043	8.2	1	7.4
	LaViLa (Zhao et al., 2022)	4.34×10^{-5}	3.09	0.27	3.07
Other Domains	LLaVA (Liu et al., 2023b)	2.88	23.96	5.85	21.26
	BLIP1 (Li et al., 2022)	7.39×10^{-5}	1.15	0.077	1.13
	BLIP2 (Li et al., 2023c)	0.028	9.04	1.05	8.2
	LaViLa (Zhao et al., 2022)	6.95×10^{-5}	3.1	0.29	3.07

world benchmarks, including the cooking domain and other domains. We have observed that the performance of other baselines has generally decreased in open-world benchmarks. These results demonstrate the importance of building models that can be generalized in the world. However, capturing the state of objects while dealing with diverse objects and domains is still a major challenge.

7 Conclusion

This paper presents a new task for comprehending the state of objects and their changes using natural language. We also propose a data generation pipeline that utilizes the capabilities of GPT-4V to tackle this task. Furthermore, we introduce OSCaR, a dataset that includes training data and a benchmark with various protocols. Our comprehensive experiments not only demonstrate the superiority of our methods in comparison to previous state-of-the-art open-source solutions but also examine the limitations of GPT-4V in addressing this challenge.

8 Limitations

This study explores a new research problem that focuses on understanding the states of objects. Although it has provided valuable insights, some limitations and areas still require further investigation, as outlined below.

Lack of audio integration: A limitation of this work is the lack of audio data, which could be useful in scenarios where sound is essential for indicating changes or properties of objects.

Challenges in long-term state transition tracking: Tracking changes in object state over extended periods is challenging because many current models, especially foundation models and models based on LLMs, do not yet have the ability to capture long-term information. This limitation highlights the difficulty in understanding complex, long-term transitions in object states, which is critical to comprehending object dynamics in various environments.

Reliance on GPT-4V's imperfect outputs: Although GPT-4V has shown strength in generating data for this research problem, its outputs are imperfect. This limitation highlights the need for strategies to efficiently learn from and improve upon the imperfect data provided by GPT-4V.

9 Ethics Statement

We acknowledge that bias could be present in the process of collecting data for our paper. To minimize this issue, we have taken several measures. Firstly, we have collected videos from two highly diverse data sources: EPIC-KITCHENS and Ego4D. Secondly, when labeling the data using GPT-4V, we are aware that bias could occur from the behavior of GPT-4V. To address this, we regularly take test samples during the data generation process. If we detect any significant issues, we are prepared to stop the process and conduct an inspection. On the human side, we use the Amazon Mechanical Turk platform to hire people to label data for both the GPT4 zero-shot and few-shot quality assessment steps and the user study. Our data collection was classified as an approved exempt protocol by the IRB.

Acknowledgement

This work was supported by NSF 2202124, NNSA NA0004078, NIH R01EY034562, and DARPA HR00112220003. The content of the information does not necessarily reflect the position of the Government, and no official endorsement should be inferred.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31.

Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. 2017. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2127–2136.

Tayfun Ates, M Samil Atesoglu, Cagatay Yigit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. 2020. Craft: A benchmark for causal reasoning about forces and interactions. *arXiv preprint arXiv:2012.04293*.

Jing Bi, Jiebo Luo, and Chenliang Xu. 2021. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Jing Bi, Nguyen Nguyen, Ali Vosoughi, and Chenliang Xu. 2023. MISAR: A multimodal instructional system with augmented reality. *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshop on AV4D: Visual Learning of Sounds in Spaces*.

Ke Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *ArXiv*, abs/2306.15195.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 30 March 2024)*.

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

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The epic-kitchens dataset. *ArXiv*, abs/1804.02748.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. 2023. Learning universal policies via text-guided video generation. *arXiv preprint arXiv:2302.00111*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Chai. 2016. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1814–1824.

Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. 2018. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–945.

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 K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, 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 Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. 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. 2021. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990.

Hung Le, Nancy F Chen, and Steven CH Hoi. 2022. Multimodal dialogue state tracking. *arXiv preprint arXiv:2206.07898*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023d. Videochat: Chat-centric video understanding. *ArXiv*, abs/2305.06355.

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

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

Yang Liu, Ping Wei, and Song-Chun Zhu. 2017. Jointly recognizing object fluents and tasks in egocentric videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2924–2932.

Jack Merullo, Dylan Ebert, Carsten Eickhoff, and Ellie Pavlick. 2022. Pretraining on interactions for learning grounded affordance representations. *arXiv preprint arXiv:2207.02272*.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. 2021. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962.

Tushar Nagarajan and Kristen Grauman. 2018. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185.

Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, and Minh Hoai. 2021. Dictionary-guided scene text recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7379–7388.

Nguyen Nguyen, Yapeng Tian, and Chenliang Xu. 2024. Efficiently leveraging linguistic priors for scene text spotting. *ArXiv*, abs/2402.17134.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.

Aishwarya Padmakumar, Mert Inan, Spandana Gella, Patrick L Lange, and Dilek Hakkani-Tur. 2023. Multimodal embodied plan prediction augmented with synthetic embodied dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6114–6131.

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*.

Nirat Saini, Hanyu Wang, Archana Swaminathan, Vinoj Jayasundara, Bo He, Kamal Gupta, and Abhinav Shrivastava. 2023. Chop & learn: Recognizing and generating object-state compositions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20247–20258.

Gabriel Sarch, Yue Wu, Michael J Tarr, and Katerina Fragkiadaki. 2023. Open-ended instructable embodied agents with memory-augmented large language models. *arXiv preprint arXiv:2310.15127*.

Luchuan Song, Jing Bi Chao Huang, and Chenliang Xu. Audio-visual action prediction with soft-boundary in egocentric videos.

Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. 2022. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13956–13966.

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. 2023a. Video understanding with large language models: A survey.

Yunlong Tang, Jinrui Zhang, Xiangchen Wang, Teng Wang, and Feng Zheng. 2023b. Llmva-gebc: Large language model with video adapter for generic event boundary captioning. *arXiv preprint arXiv:2306.10354*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jianren Wang, Sudeep Dasari, Mohan Kumar Srirama, Shubham Tulsiani, and Abhinav Gupta. 2023. Manipulate by seeing: Creating manipulation controllers from pre-trained representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3859–3868.

Te-Lin Wu, Yu Zhou, and Nanyun Peng. 2023. Localizing active objects from egocentric vision with symbolic world knowledge. *arXiv preprint arXiv:2310.15066*.

Zihui Xue, Kumar Ashutosh, and Kristen Grauman. 2024. Learning object state changes in videos: An open-world perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. Piglet: Language grounding through neuro-symbolic interaction in a 3d world. *arXiv preprint arXiv:2106.00188*.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2023a. Transfer visual prompt generator across llms. abs/23045.01278.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858.

Yue Zhao, Ishan Misra, Philipp Krahenbuhl, and Rohit Girdhar. 2022. Learning video representations from large language models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6586–6597.

Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. 2023. Learning procedure-aware video representation from instructional videos and their narrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14825–14835.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592.