

Action Without Interaction: Probing the Physical Foundations of Video LMMs via Contact-Release Detection

Daniel Harari¹ Michael Sidorov¹ Chen Shterental¹ Liel David¹
Abrham Kahsay Gebreselasie² Muhammad Haris Khan²

¹Weizmann Institute of Science

²Mohamed bin Zayed University of Artificial Intelligence

{hararid, michael.sidorov, chen.shterental}@weizmann.ac.il, olesya.liel@gmail.com

{abrham.gebreselasie, muhammad.haris}@mbzuai.ac.ae

Abstract

Large multi-modal models (LMMs) show increasing performance in realistic visual tasks for images and, more recently, for videos. For example, given a video sequence, such models are able to describe in detail objects, the surroundings and dynamic actions. In this study, we explored the extent to which these models ground their semantic understanding in the actual visual input. Specifically, given sequences of hands interacting with objects, we asked models when and where the interaction begins or ends. For this purpose, we introduce a first of its kind, large-scale dataset with more than 20K annotated interactions on videos from the Something-Something-V2 dataset. 250 AMTurk human annotators labeled core interaction events, particularly when and where objects and agents become attached ('contact') or detached ('release'). We asked SoTA LMMs, including GPT, Gemini and Qwen to locate these events in short videos, each with a single event. The results show that while models reliably name target objects and identify actions, they exhibit a form of 'shortcut learning' where semantic success masks a failure in physical grounding. Specifically, they consistently fail to identify the frame where the interaction begins or ends and poorly localize the physical event within the scene. This disconnect suggests that while LMMs excel at System 1 intuitive pattern recognition (naming the action and objects), they lack the System 2 cognitive foundations required to reason about physical primitives like 'contact' and 'release', hence truly ground dynamic scenes in physical reality.

1. Introduction

Discovering and understanding actions and interactions are fundamental cognitive capabilities of humans and other in-

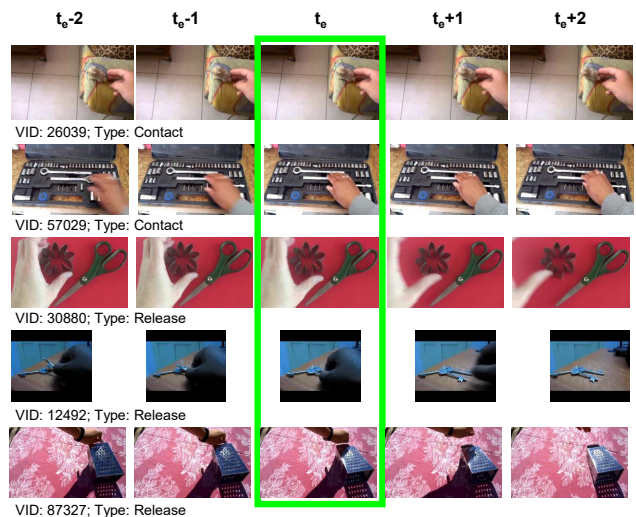


Figure 1. **Events of physical contact or release.** Each row presents 2 frames before and after a true physical contact or release event (marked with a green box) during a hand-object interaction. Humans are sensitive to the gap between the hand and the target object even one frame before a contact or after a release. Evidence of a physical contact is crucial to identify an interaction over a simple occlusion, or to predict the outcome of an interaction at its termination. These examples are from our annotated dataset, extracted from SSv2.

telligent beings, necessary in interpreting and planning dynamic events between objects and agents in the surrounding environment [22]. Infants develop early sensitivity to spatiotemporal continuity in simple events, for example, the physical contact between an agent and a target object (see Fig. 1). In developmental psychology, the perception of physical contact and release are considered 'primitives' of causal reasoning—foundational building blocks that allow

humans to move beyond surface correlations to an understanding of agency and physical laws. This sensitivity to the causality of perception guides infants at a very young age, in learning to detect and interpret interactions between objects and agents, including launching, entraining and expulsion events [2, 14, 18, 23].

Computationally, recent vision models showed increasing performance in recognizing actions and interactions in realistic video sequences [9, 19, 28]. Some models include special architectural designs to improve on the internal representation learning [38], while others use a common architecture, but train on very large unlabeled video datasets utilizing self-supervised learning (SSL) paradigms [5, 16, 30]. The introduction of large multi-modal models (LMMs), allowed the combination of semantic information from large language models (LLMs) and foundational visual representations, thus allowing to generalize to unseen videos and actions without explicit training [28].

Despite the increasing success of models in generalizing to more complex tasks without explicit training, recent studies revealed fundamental limitations in the models’ ability to reason about the performed tasks and develop human-like generalizable cognitive understanding [4, 13, 24]. In this study, we explore whether the enhanced performance of LMMs in video action recognition, reflects an improved cognitive understanding (System 2), or merely a superficial ‘story telling’ ability about detected objects in proximity of hands (System 1)?

For this purpose we introduce a large-scale dataset – *The Contact-Release Interaction Dataset* (CRID). This first of its kind dataset consists of more than 20K annotated interactions, based on 10,000 action videos from the “Something-Something v.2” (SSv2) dataset [10]. Using the Amazon Mechanical Turk (AMTurk) crowd sourcing platform, we conducted a survey, in which 250 human annotators labeled 24,222 core interaction events, including: (i) the type of agent acting upon the target object (e.g., a hand or another object), (ii) the type of core interaction event (‘contact’—attachment between a target object and an agent, or ‘release’—detachment of the object from the agent), (iii) the spatiotemporal location of the event (frame number and image coordinates). Fig. 2 depicts the annotation setting (see details in Sec. 3).

Based on these annotations, we conducted a series of experiments to evaluate the ability of current LMMs to detect the spatiotemporal location of core interaction events in real-world video sequences. The experiments were conducted under several In-Context-Learning (ICL) regimes, while applying two modifying conditions on the models’ prompts – *Reasoning* and *Grounding* – inspired by earlier studies [12, 25, 32]. We applied the evaluation scheme (see Fig. 3) to five SoTA LMMs including, OpenAI’s GPT-5.2 and GPT-4o models [20, 26], Google’s Gemini-Pro-3 and

Gemini-2.5-Flash models [6, 7] and an open source version of Alibaba’s Qwen-2.5VL-72B [1].

In summary our contributions include:

- Introduce CRID - a large scale dataset, based on 10K interaction videos from SSv2 dataset, with more than 20K first of their kind human annotations of core interaction events (‘Contact’ and ‘Release’). The annotations include details on the event and agent types, as well as the spatiotemporal locations of the events.
- A set of prompting experiments under several ICL regimes and modifying *Reasoning* and *Grounding* conditions.
- A discussion around the striking ‘grounding gap’ revealed in our results: models can identify a high-level action label through statistical association, but remain ‘blind’ to the underlying physical events that define that action.

2. Related Work

Video understanding was thoroughly researched in the past due to its high value to the advancement in the domain of AI. Unlike action recognition in video, such as identifying people jumping, playing tennis etc., video understanding is a more complex task, often requiring a high level of generalization. Recent studies in this area of research, such as Maaz et al. [17], showed that a ChatGPT agent can successfully answer complex questions when prompted with image data together with the verbal question. Wu et al. [33] used segmentation masks, which the model provided in a grounding step, to further improve the model’s understanding of the input images. Shao et al. [25] showed that by asking the model to produce a Chain-of-Thought of a general task related to the image of the main task, improves significantly the final answer of the LMM. Tian and Wu [29] also used a prompt tuning technique to improve action recognition performance of an LMM agent, but their method requires training of additional adapter models that embed the prompt, actions and the image to be tokenized and sent to the LMM. Chen et al. [3] utilize reinforcement learning to improve LMM grounding via iterative ROI extraction and prediction-based feedback.

While Qi et al. [21] reveal that image-based LMMs often prioritize textual priors over visual evidence, their analysis remains limited to static scenes. Our work extends this critique to the temporal domain, demonstrating that even when models produce coherent action narratives, they fail to ground them in the physical primitives—such as contact and release—that define dynamic interactions.

Finally, several studies focused on changes that a manipulated object undergoes during long interaction videos

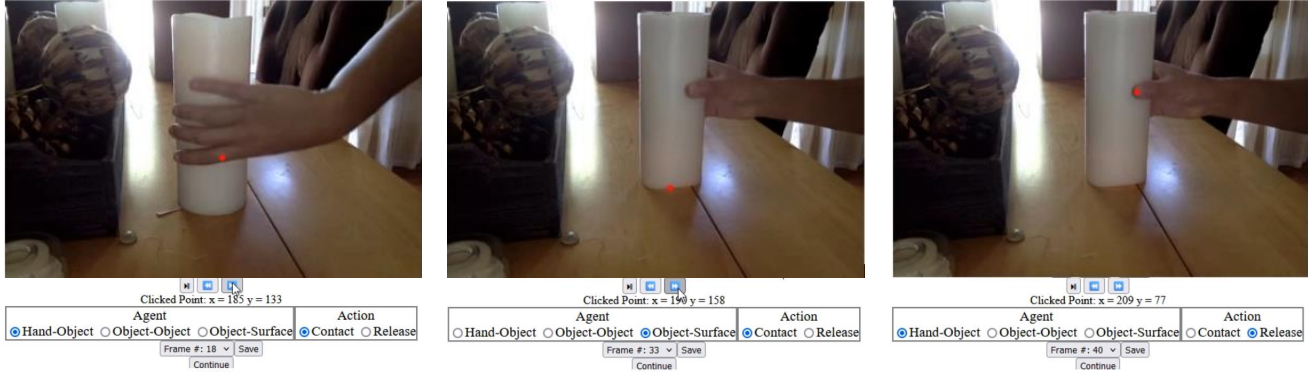


Figure 2. **Collecting human annotations for interactions using Amazon Mechanical Turk platform.** Human subjects were asked to annotate core interaction events in videos from SSv2 dataset [10]. Shown here are example annotations for ‘contact’ and ‘release’ events, where the target object (white candle) comes in contact with a hand (left) and a surface (middle), or is detached from the hand (right). The annotations include the event type, the kind of agent-object pair and the spatiotemporal location of the event (frame and image coordinates).

(e.g., paper folding, fruit peeling, etc.), utilizing existing benchmarks for object affordance and object manipulation [27, 31, 34, 36]. The videos are often at low resolution and include coarse categories of background and actions, rather than the begin or end of a physical contact. In contrast, our evaluation is focused on the initial moment of contact or release in an interaction, regardless of the kind of target object or the kind of manipulation, which are the core of the object affordance studies. We believe that the generic ability to detect physical contact is fundamental for learning about object affordance and manipulations in a self-supervised manner.

3. Dataset

Overview. In this section we describe the dataset used in our experiments. We explore models’ understanding of core interaction elements, for example the moment a hand picks up a target object and the spatial location where the physical contact occurs in the scene. We introduce a new large scale dataset – *The Contact-Release Interaction Dataset (CRID)*. This first of its kind dataset consists of more than 20K annotated interactions on more than 10K action videos from the “Something-Something-v2” (SSv2) dataset, which is a collection of smartphones’ videos depicting humans performing various everyday interactions in natural settings [10]. The SSv2 labels include generic action templates, for example, “putting something into something”, and the object names in the template placeholders (“something”) for each video.

Human annotations. In this work we used the AMTurk online platform to conduct a survey, in which 250 human annotators labeled more than 20K core interaction events, including: (i) the type of agent acting upon the target object

(e.g., a hand or another object), (ii) the type of core interaction event, i.e., ‘contact’ or ‘release’, (iii) the spatiotemporal location of the event, i.e., the frame number and image coordinates where the physical event occurs (see Fig. 1 and Fig. 2). The survey and the annotations collection procedure were approved by the institutional review board of the Weizmann Institute of Science, Rehovot, Israel. All human subjects gave informed consent before participating in the survey. Details about the new annotated dataset, including the videos and the labels are summarized in Tab. 1. Due to time and budget constraints, we could not hire multiple annotators for each of the 10K videos. However, we evaluated the inter-rater agreement between randomly selected annotators, by measuring the Intraclass Correlation Coefficient (ICC) between 3 random annotators on a randomly selected subset of videos. The results indicate a high agreement level on the frame annotation (ICC of 0.95), and a medium agreement level on the spatial location where the events occur (ICC of 0.73 and 0.39 for the exact x and y coordinates, respectively, and an IoU of 0.57 for a 120×120

Table 1. *The Contact-Release Interaction Dataset (CRID)*. New annotations of core interaction events for videos extracted from the SSv2 dataset [10].

Videos	10130
SSv2 action templates	91
Mean videos per template	111
‘Contact’ events	13816
‘Release’ events	10406
‘Hand-object’ interactions	12550
‘Object-object’ interactions	5653
‘Object-surface’ interactions	6019

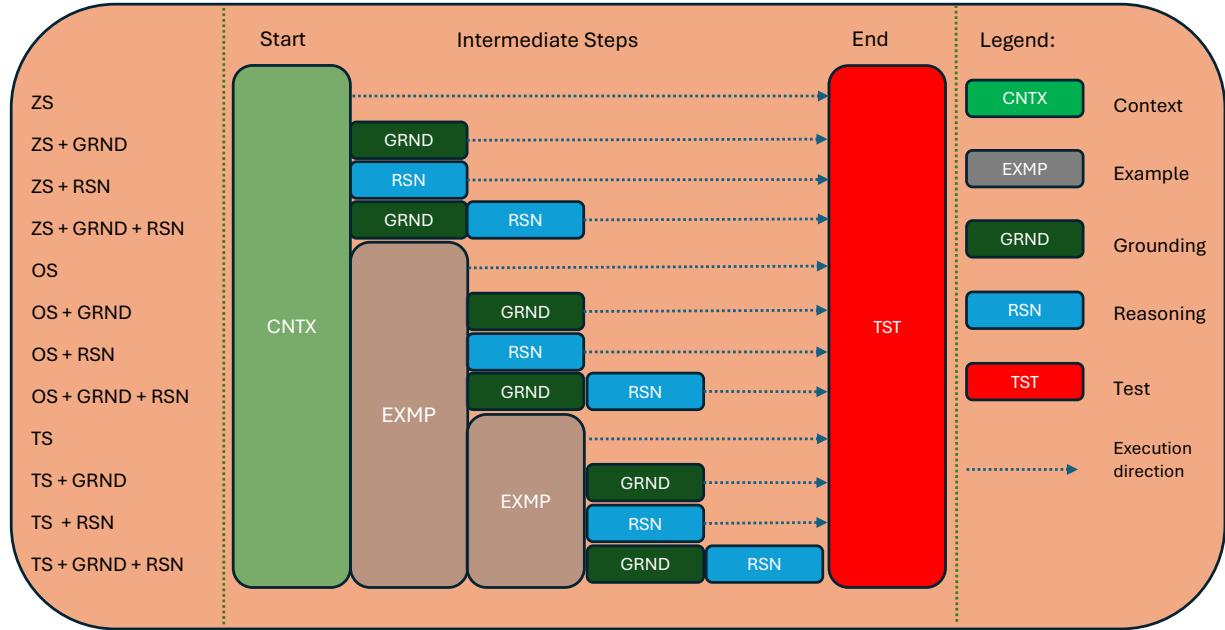


Figure 3. A schematic flow chart of the experiments under the different In-Context-Learning (ICL) regimes (i.e., ZS, OS, TS) and modulating conditions. The blocks represent different components of intermediate procedures. Each row represents an experiment using a particular ICL regime and condition (the experiment flow is directed left to right). The CNTX block indicates an introductory prompt about the agent. The EXMP block represents a prompt of an example, including the task instruction, an input video and the correct response for this example. The RSN block indicates a prompt instructing the model to include in the response a step-by-step description of the reasoning behind the predicted answer. The GRND block represents a prompt instructing the model to describe the content of the input video and the instructing prompt. In this block, the model provides an intermediate response, prior to the main task. The TST block indicates the prompt of the main test task, including the instruction and test video (see Sec. 4 for more details).

bounding box centered around the labeled location).

Open access availability. The dataset is publicly available at [ssv2-contact-release-interaction-dataset](https://github.com/robotics-robotics/SSv2-Contact-Release-Interaction-Dataset). Despite concerns about benchmark data leaking into LMM pre-training, recent research (e.g., [11]) indicates that even top-tier SSL visual encoders struggle with physical interactions and contact, even after fine-tuning. Consequently, we provide open access to our annotations. This allows future models to leverage these scarce examples alongside their inherent reasoning and semantic knowledge (similar to humans) to bridge current limitations.

4. Experimental Design

Overview. In this section we describe in more detail the evaluation experiments conducted in the course of this study. We considered three ICL regimes: zero-shot, one-shot and two-shot [8].

As we are interested in models’ ability to understand core interaction events in videos, we focused on the detection of the frame in the video, where a core interaction event (‘contact’ or ‘release’) occurs. Since the original SSv2 videos often contain several simultaneous inter-

action events, we manually extracted 3 temporally isolated events from a subset of 33 videos, resulting with 99 short sequences, 10 frames each, consisting of a single core event (see supplementary for details). As a control, we extracted from each video, an additional short sequence (a ‘non-event’), which depicts objects and hands in a dynamic scene without a physical contact/release event.

To better understand the models behavior, we employed two explainability methods, that were used in previous studies with LMMs, as modulating conditions: (i) *Grounding*, (ii) *Reasoning*. We elaborate about these conditions later in this section. In our experiments we evaluated their relative influence on the models performance. These prompting conditions do not modify the models’ internal reasoning parameters. Fig. 3 presents the different experimental settings, including the ICL regimes and modulating conditions.

Zero-Shot (ZS) regime. In the baseline ZS regime, models are instructed to perform the main task on a test video (i.e., detect the frame where the a core event occurs), without any examples. The prompt includes an introduction, the image frames of the test video sequence and the instructions for the main task, as shown in Listing 1.

Listing 1. The prompt used in the baseline ZS experiment.

```

System:
You are a useful assistant and an expert in video
    ↪ understanding.
Images:
[First test video frame]
...
[Last test video frame]
User:
The uploaded images are consecutive frames from a video.
    ↪ The numbers in the file name indicate the order
    ↪ of the frames in the sequence, so frame_1.jpg
    ↪ is the first frame, followed by frame_2.jpg
    ↪ which is the second frame, etc. The sequence
    ↪ shows an interaction between a hand and an
    ↪ object. An interaction usually begins when an
    ↪ object starts to move with the hand. An
    ↪ interaction usually ends when the hand starts to
    ↪ move without the object.
Q: In which frame does the interaction end?
Answer briefly with: "Prediction: <frame number>"

```

Listing 2. A prompt used in the OS experiment.

```

System:
You are ...
Images:
[First frame of example video]
...
[Last frame of example video]
User:
The uploaded images ...
Q: In which frame does the interaction end?
Answer briefly with: "Prediction: <frame number>"
A: Prediction: 3
Images:
[First test video frame]
...
[Last test video frame]
User:
The uploaded images ...
Q: In which frame does the interaction end?
Answer briefly with: "Prediction: <frame number>"

```

One-Shot (OS) regime. In the OS regime, models are provided with a single example of the main task prior to performing the task on the test video sequence. The example, includes a video with a different event from the experimental dataset and the annotated true frame for this video, where the core event occurs. The prompt is provided in Listing 2. We evaluate the models on each test video in a leave-one-out approach, and average the accuracy for that video across all trials.

Two-Shot (TS) regime. We further extended the experiment of the OS regime, to the TS regime, by presenting to the model a second example of another event with the corresponding labeled frame, prior to the test. As in the OS regime, we average the accuracy across many trials for each test video (same total trials as in the OS regime), where in each trial the two examples are drawn randomly from the set of experimental dataset, excluding the test video.

Listing 3. A grounding prompt in the OS experiment to verify model’s understanding of both verbal and visual contents of a provided example.

```

System:
You are ...
Images:
[First example video frame]
...
[Last example video frame]
User:
The uploaded images ...
To check your understanding, please repeat the correct
    ↪ answer to the example question, and specify
    ↪ which object was in contact with the hand in
    ↪ these frames? Answer with the object’s name.

```

Listing 4. An instructing prompt for step-by-step reasoning.

```

System:
You are ...
Images:
[First test video frame]
...
[Last test video frame]
User:
The uploaded images ...
Q: In which frame ... ?
Answer briefly with ...
Think step by step, and show all intermediate reasoning
    ↪ before giving the final answer.

```

Grounding condition. The grounding procedure precedes the main test task. The model is instructed to describe the contents of the provided image sequence, and to repeat a particular piece of information from the instructing prompt. The motivation behind this procedure came from a preliminary experiment that was conducted with the online User-Interface of ChatGPT-4o, in which the model had better understanding of the instructing prompt and more attention to the image contents, when explicitly was asked about them. The prompt used for the grounding condition is shown in Listing 3.

Reasoning condition. The reasoning procedure is combined with the main task, by instructing the model to describe in detail the reasoning behind its answer to the main frame detection task (see prompt in Listing 4). This prompt does not modify the internal reasoning parameters of the model. The motivation for this experimental condition comes from recent studies showing that models’ performance increases when they are explicitly instructed to provide a step-by-step description, and thus may extract and combine more relevant information in the final answer [32]. Fig. 4 shows an example of a false prediction together with the model’s step-by-step chain of thought.

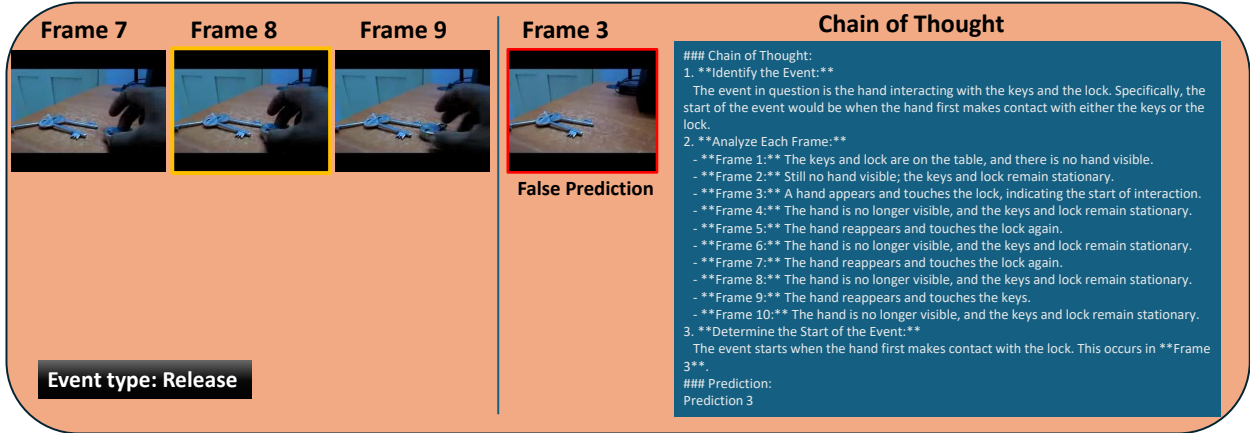


Figure 4. **Example false prediction.** The model provides the chain-of-thought under the “WITH” *Reasoning* condition. The example shows that the reasoning seems logical and realistic, but the grounding of the physical release event in the video frames is very loose. The orange box marks the true frame. The red box marks the false prediction.

5. Results

In this section we describe in detail the results from the experiments conducted in this study.

Event frame detection. The main task for the models was to detect the frame where a test event (‘contact’ or ‘release’) occurs. A correct detection is considered within an allowed error tolerance e_τ . For example, $e_\tau = 0$ means the model predicted the exact true (human) labeled frame where the event occurs, and $e_\tau = 1$ means that the predicted frame can be up to one-frame off (before or after) the true frame. As demonstrated in Fig. 1, it is often evident that there is no physical contact between the hand and target object already one frame after a release or before a contact event. Throughout this section, the models’ performance is mea-

Table 2. **Reasoning effect on LMMs’ performance.** Models are instructed to describe step-by-step the reasoning behind their prediction of the frame where the interaction occurs. Mean accuracy is measured for the detection of the frame where the test event occurs, within an allowed error tolerance (here the exact or one frame off the true frame). Results are reported as per 3 ICL regimes (ZS, OS, TS) for the models: (i) Gemini-2.5-Flash and (ii) GPT-5.2.

ICL	RSN	Mean Accuracy Percentage (SD)			
		Exact		1-off	
		Gemini	GPT	Gemini	GPT
ZS	W/O	8.08 (1.94)	9.60 (2.10)	30.81 (3.29)	35.35 (3.41)
	W	9.09 (2.05)	12.12 (2.33)	29.80 (3.26)	33.84 (3.37)
OS	W/O	19.87 (2.84)	18.15 (2.29)	47.15 (3.55)	42.96 (3.19)
	W	16.74 (2.65)	15.93 (2.02)	43.71 (3.53)	39.43 (3.05)
TS	W/O	20.71 (20.71)	18.03 (2.73)	49.07 (3.55)	44.31 (3.53)
	W	17.98 (2.73)	17.12 (2.68)	47.53 (3.55)	40.82 (3.49)

sured as the mean detection accuracy (with standard error) across all test events (99) in the experimental dataset.

Fig. 5 presents performance charts for the evaluated models. The results in Fig. 5a indicate that at the ZS regime, the models’ accuracy is about chance level (10%) for detecting the exact frame where the event occurs. The accuracy for an error tolerance of one-frame off peaks at 42.9% with Gemini-3-Pro. However, as indicated by Fig. 5b, the models’ failure often tends to frames with no visual grounding of a physical contact. In addition, when evaluated on ‘non-event’ control videos, which contained hands and objects but did not present a physical contact, most of the models completely failed. Except for Qwen, which responded correctly at 18.2% of the trials that there is no event in these video, the rest of the models always predicted a frame and provided an explanation for their choice. The chart in

Table 3. **Grounding effect on LMMs’ performance.** Prior to the frame detection task, models are instructed to name the target object, to improve their perceptual grounding. Mean accuracy is measured for the detection of the frame where the test event occurs, within an allowed error tolerance (here the exact or one frame off the true frame). Results are reported as per 3 ICL regimes (ZS, OS, TS) for the models: (i) Gemini-2.5-Flash and (ii) GPT-5.2.

ICL	GRND	Mean Accuracy Percentage (SD)			
		Exact		1-off	
		Gemini	GPT	Gemini	GPT
ZS	W/O	7.58 (1.89)	11.62 (2.28)	29.29 (3.24)	34.85 (3.39)
	W	9.60 (2.10)	10.10 (2.15)	31.31 (3.30)	34.34 (3.38)
OS	W/O	17.78 (3.84)	17.31 (2.18)	44.77 (3.53)	41.94 (3.16)
	W	18.84 (2.78)	16.77 (2.14)	46.09 (3.54)	40.45 (3.09)
TS	W/O	19.32 (2.81)	19.31 (2.81)	49.49 (3.55)	45.22 (3.54)
	W	19.37 (2.81)	15.84 (2.60)	47.10 (3.55)	39.92 (3.48)

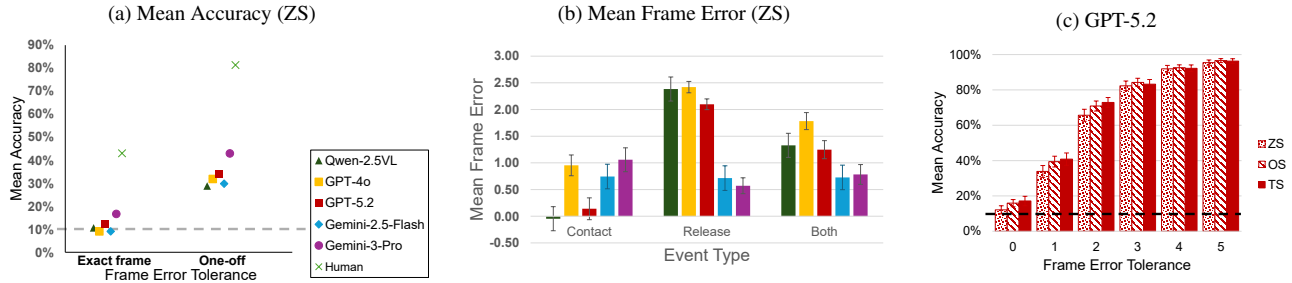


Figure 5. **Models’ performance.** (a) A comparison between models’ performance at Zero-Shot ICL. Exact detection of the physical event is around chance level for most models. At one-frame off error tolerance Gemini-pro-3 peeks at 42.9%. Humans outperform the models by far. (b) The models’ mean frame error. A positive error indicates a prediction which precedes a physical contact event, or follows a physical release event. In these frames there is often a clear gap between the hand and the target object. (c) Detail mean accuracy vs. frame error tolerance for GPT-5.2. The chart depicts a slight performance increase across the different ICL regimes (ZS, OS, TS). Results are under the “WITH” Reasoning condition. Chance level is 10% (shown as a dashed horizontal line) as the length of all videos in the experimental dataset is 10 frames.

Fig. 5c depicts the GPT-5.2 model’s performance as a function of the allowed error tolerance (other models yielded similar results, see Supplementary). The chart also shows a minor increase in performance between the different ICL regimes as was demonstrated already in earlier studies [8].

Tab. 2 reports the mean accuracy for $e_\tau = 0$ and $e_\tau = 1$ as per each of the three ICL regimes (ZS, OS, TS) under the Reasoning condition (see Sec. 4). Similarly, Tab. 3 lists the mean accuracy under the Grounding condition (see supplementary for the results on other models).

These results indicate that explicitly instructing the models to provide visual grounding evidence and reasoning for their answers does not necessarily increase their performance. Surprisingly, the results indicate that these instructions, may sometimes yield a decrease in the models’ performance. This reflects the foundational gap between System 1 and System 2 abilities of the models. We further discuss this in Sec. 6.

Finally, we performed an ablation experiment, inspired by recent work on boosting LLMs reasoning with a verification feedback procedure – similar to reinforcement learning approach [12, 32, 35] – but this did not improve the models’ performance on the detection task (see supplementary).

Action and object recognition. To contextualize the failure in fine-grained grounding, we benchmarked the models’ high-level action and object recognition capabilities. Using our experimental dataset consisting of a subset of 33 videos across 15 action templates from SSv2, we evaluated two distinct tasks:

(i) Action Classification: Models were prompted to identify the correct SSv2 action template for each video. To ensure evaluation precision and avoid linguistic ambiguity, we required models to output numerical template IDs rather than text strings. Accuracy was measured using Top-1 and Top-5 metrics. For few-shot guidance, the prompt included

verbal descriptions of four representative interaction styles, excluding actual video frames.

(ii) Object Naming: Given the correct action template, models were tasked with replacing “something” placeholders with the specific objects identified in the video. Accuracy was determined by manually verifying predicted objects against ground-truth labels, accounting for ordering and synonyms.

As shown in Tab. 4, the models demonstrate robust global understanding, with most achieving over 85% Top-5 action accuracy and high object naming precision. This confirms that their struggle to pinpoint interaction events is not due to a lack of semantic or object-level recognition, but rather a specific deficit in spatiotemporal grounding.

Event bounding-box (BBox) detection. We evaluated the models ability to locate the interaction region within the predicted video frame, where the event occurs. Since the annotations contain a point image location per event (point of contact or release), we considered for the evaluation several formulations for the prediction of the event location, including exact point location and different BBox formats. We report here the best performing setting, which was predicting a BBox, and measuring the Intersection-over-Union (IoU) between the prediction and a square BBox centered

Model	Action Classification Acc. (%)		Obj. Naming Acc. (%)
	Top-1	Top-5	
GPT-4o	48.48 ± 8.69	81.82 ± 6.71	81.82 ± 6.71
GPT-5.2	45.45 ± 8.67	96.97 ± 2.98	93.94 ± 4.15
Gemini-2.5	66.67 ± 8.33	90.91 ± 5.08	93.94 ± 4.15
Gemini-3-Pro	63.64 ± 8.37	87.88 ± 5.68	96.97 ± 2.98
Qwen-2.5VL	60.61 ± 8.51	84.85 ± 6.24	69.70 ± 8.00

Table 4. Per-model accuracy statistics for global action classification and naming the correct target objects in each test video.

around the label point location. We tested several box sizes in the range 20 to 200 pixels. Using a 120×120 pixel size boxes around the annotation point yielded the best performance: mean IoU of only 0.147 for Gemini-3-Pro and 0.086 for GPT-5.2 (see supplementary). Overall, these findings indicate that the model is unable to reliably localize the regions around the physical interaction.

Human performance. As a comparison to the models performance, we asked two naive humans to perform the same tasks on the experimental dataset. For humans, the mean accuracy on the event frame detection task was $43.00\% \pm 5.04\%$ for the exact frame, and $81.36\% \pm 4.18\%$ for 1-frame off - high above the best performing model as shown in Fig. 5a. The ICC between the human subjects was 0.87. The accuracy on the 'non-event' control set was $70.26\% \pm 6.93\%$. The mean IoU for the event location detection was 0.71 ± 0.02 .

6. Discussion

Recent advancement of current vision models improve significantly LMMs generalization ability in recognizing unseen actions and scenes in real-world videos [24, 28, 30]. Our action recognition evaluation verify this increased performance for five models: Qwen-2.5VL-72B, GPT-4o, GPT-5.2, Gemini-2.5-Flash and Gemini-3-Pro (see Tab. 4). However, previous studies have indicated a possible limitation of current vision models in understanding the core interaction events underlying the general action [11]. Do LMMs overcome this limitation by leveraging their vast common knowledge and visual semantics?

In this study we conducted a series of experiments, under several ICL regimes, to test LMMs' ability in detecting where and when in the video, core interaction events occur. Specifically, we focused on 'contact' and 'release' events, where a target object becomes attached to an agent (e.g., a hand) or detached from the agent (Fig. 1; Sec. 4). We introduced the *Contact-Release Interaction Dataset (CRID)* – a new large scale dataset with more than 20K human annotated events in videos from the SSv2 dataset [10] (Sec. 3).

Our experimental results indicate that despite the System 1 ability of the models to classify correctly the action in the videos ($> 80\%$ in Top-5) and even name the correct target objects in the scenes ($> 70\%$), the models struggle with detecting the physical core events and ground them visually in the videos ($< 17\%$). Introducing similar examples using few-shot ICL paradigm slightly improves the performance, which still remains slightly above chance level (Fig. 5c).

In contrast with earlier studies [25, 32], applying Chain-of-Thought prompting does not necessarily increase the models' performance (see Tab. 2). Similarly, explicitly instructing the models to attend and describe the input video (e.g., name the target objects in the interaction scene), does

not improve the models' grounding ability and hence the performance (see Tab. 3). These results suggest that the models lack System 2 understanding abilities of core interaction events.

We find that models struggle with the perceptual grounding of the core events underlying actions and interactions in the visual input, despite their general ability to describe the action and participating objects and agents in the interactions. This limitation is partly related to the challenge of complex question decomposition as was already shown in previous studies [37]. However, it seems that there is more to this limitation. We hypothesize that the main limitation is rooted in a loose integration between the visual representation (often of pretrained visual transformers) and the language representation, which are mostly trained separately. This limitation projects also to the models' inability to overcome current challenges of visual models in interpreting spatial relations between objects [15] and complex dynamic events, despite their huge semantic knowledge. In a sense, the models exhibit a "shortcut learning" behavior and merely able to tell a "good story" about possible interactions when hands appear in proximity of objects in scenes. In struggling to pinpoint the moment and location of physical contact that defines the interactions, the models lack the perceptual grounding required for deeper understanding of dynamic scenes

The implication of this limitation may be that current LMMs lack the capacity to develop full visual understanding of dynamic interactions, similar to intelligent beings [18, 24], and thus can have only limited ability in interpreting unfamiliar and complex interactions, as well as in planning interactions on their own for artificial systems.

7. Conclusions

In this paper we demonstrate a major limitation of current large multi-modal models in understanding dynamic interactions. Our analysis suggests that current models are operating as sophisticated System 1 engines. They recognize a 'picking up' action by the presence of a hand and a cup, but they do not perform the System 2 'mental simulation' required to pinpoint the exact moment of physical attachment. To move toward genuine multi-modal intelligence, future architectures must incorporate structured priors or causal world models that treat interaction events not just as pixels, but as discrete physical state changes (e.g., attend to motion and motion boundaries around the hand and the object). We introduce CRID - an extension to the SSv2 dataset with more than 20K detailed annotations of core physical events in more than 10K videos. These annotations may be used in future efforts to develop new architectures or foundation models with cognitive understanding of visual dynamic interactions.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2
- [2] Renee Baillargeon. Physical reasoning in infancy. *The cognitive neurosciences*, pages 181–204, 1995. 2
- [3] Zhangquan Chen, Xufang Luo, and Dongsheng Li. Visrl: Intention-driven visual perception via reinforced reasoning. *arXiv preprint arXiv:2503.07523*, 2025. 2
- [4] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A. Smith, and Joshua B. Tenenbaum. Are deep neural networks smarter than second graders? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10834–10844, 2023. 2
- [5] Ishan Dave, Malitha Gunawardhana, Limalka Sadith, Honglu Zhou, Liel David, Daniel Harari, Mubarak Shah, and Muhammad Khan. Unifying video self-supervised learning across families of tasks: A survey. *Preprints*, 2024. 2
- [6] Google DeepMind. Gemini 2.5. <https://deepmind.google>, 2025. 2
- [7] Google DeepMind. Gemini 3 pro. <https://deepmind.google>, 2026. Accessed: 2026-02-24. Model version: 3.0 Pro. 2
- [8] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, 2024. Association for Computational Linguistics. 4, 7
- [9] Ping Gong and Xudong Luo. A survey of video action recognition based on deep learning. *Knowledge-Based Systems*, 320:113594, 2025. 2
- [10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yanilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2, 3, 8
- [11] M. Gunawardhana, L. Sadith, L. David, D. Harari, and M. H. Khan. How effective are self-supervised models for contact identification in videos. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 117–131, Singapore, 2024. Springer Nature Singapore. 4, 8
- [12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiusi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaoshua Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhiqiang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning, 2025. 2, 7
- [13] Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. Do vision-language models really understand visual language? In *Forty-second International Conference on Machine Learning*, 2025. 2
- [14] Alan M Leslie. The perception of causality in infants. *Perception*, 11(2):173–186, 1982. 2
- [15] Jingping Liu, Ziyang Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–632, Vienna, Austria, 2025. Association for Computational Linguistics. 8
- [16] Runze Liu, Yaquang Fang, Fan Yu, Ruiqi Tian, Tongwei Ren, and Gangshan Wu. Deep video understanding with video-language model. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 9551–9555, New York, NY, USA, 2023. Association for Computing Machinery. 2
- [17] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and

- Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 2
- [18] Albert Michotte. *The perception of causality*. Basic Books, 1963. 2, 8
- [19] Md Golam Morshed, Tangina Sultana, Aftab Alam, and Young-Koo Lee. Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors*, 23(4), 2023. 2
- [20] OpenAI. Gpt-4o system card. *arXiv preprint arxiv:2410.21276*, 2024. 2
- [21] Shuhan Qi, Zhengying Cao, Jun Rao, Lei Wang, Jing Xiao, and Xuan Wang. What is the limitation of multimodal llms? a deeper look into multimodal llms through prompt probing. *Information Processing and Management*, 60(6): 103510, 2023. 2
- [22] Gabriel A Radvansky and Jeffrey M Zacks. *Event cognition*. Oxford University Press, 2014. 1
- [23] Rebecca Saxe and Susan Carey. The perception of causality in infancy. *Acta psychologica*, 123(1-2):144–165, 2006. 2
- [24] Luca M. Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96, 2025. 2, 8
- [25] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2, 8
- [26] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 2
- [27] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
- [28] 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, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2, 8
- [29] Mengxiao Tian, Xinxiao Wu, and Shuo Yang. Llm-enhanced action-aware multi-modal prompt tuning for image-text matching. *arXiv preprint arXiv:2506.23502*, 2025. 2
- [30] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 2, 8
- [31] Mahiro Ukai, Shuhei Kurita, and Nakamasa Inoue. Status bench: A rigorous benchmark for evaluating object state understanding in vision-language models. page 4718–4727, New York, NY, USA, 2025. Association for Computing Machinery. 3
- [32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 2, 5, 7, 8
- [33] Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-Imm: Grounding frozen large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24710–24721, 2025. 2
- [34] Zihui Xue, Kumar Ashutosh, and Kristen Grauman. Learning object state changes in videos: An open-world perspective. *CVPR*, 2024. 3
- [35] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 7
- [36] Parnian Zameni, Yuhan Shen, and Ehsan Elhamifar. Moscato: Predicting multiple object state change through actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11600–11611, 2025. 3
- [37] Haowei Zhang, Jianzhe Liu, Zhen Han, Shuo Chen, Bailan He, Volker Tresp, Zhiqiang Xu, and Jindong Gu. Visual question decomposition on multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1926–1949, Miami, Florida, USA, 2024. Association for Computational Linguistics. 8
- [38] Long Zhao, Nitesh B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming-Hsuan Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. Videoprism: a foundational visual encoder for video understanding. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 2