

# While recognizing actions in videos, LMMs struggle to detect core interaction events

Daniel Harari<sup>1</sup> Michael Sidorov<sup>1</sup> Liel David<sup>1</sup> Chen Shterental<sup>1</sup>  
 Abrham Kahsay Gebreselasie<sup>2</sup> Muhammad Haris Khan<sup>2</sup>

<sup>1</sup>Weizmann Institute of Science

<sup>2</sup>Mohamed 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 two LMMs (Qwen-2.5VL and GPT-4o) to locate these events in short videos, each with a single event. The results show that although the models can reliably name the target objects, identify the action and provide coherent reasoning, they consistently fail to identify the frame where the interaction begins or ends and cannot localize the event within the scene. Our findings suggest that in struggling to pinpoint the moment and location of physical contact that defines the interaction, the models lack the perceptual grounding required for deeper understanding of dynamic scenes.*

## 1. Introduction

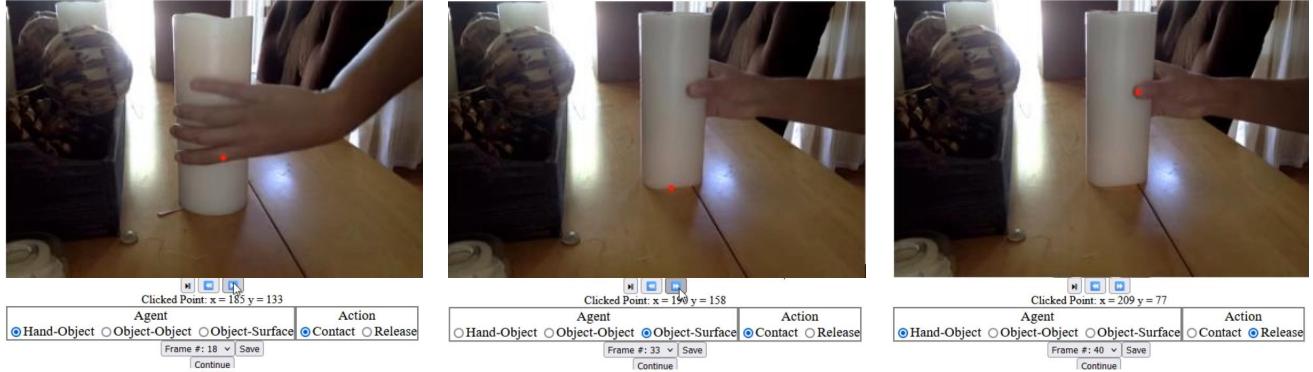
Discovering and understanding actions and interactions are fundamental cognitive capabilities of humans and other intelligent beings, necessary in interpreting and planning dynamic events between objects and agents in the surrounding environment [20]. Infants develop early sensitivity to spa-

tiotemporal continuity in simple events, for example, the physical contact between an agent and a target object. 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, 12, 16, 21].

Computationally, recent vision models showed increasing performance in recognizing actions and interactions in realistic video sequences [7, 17, 24]. Some models include special architectural designs to improve on the internal representation learning [31], while others use a common architecture, but train on very large unlabeled video datasets utilizing self-supervised learning paradigms [5, 14, 26]. 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 [24].

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 understanding [4, 11, 22]. We are interested to explore whether the enhanced generalization behavior in interpreting actions in LMMs, reflects a deeper understanding of dynamic events at the core of interactions, similar to intelligent beings, or does it merely reflect a seemingly convincing "story telling" ability of reliably detected objects in the scenes?

For this purpose we introduce a large-scale dataset – *The Contact-Release Interaction Dataset*. 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 [8]. Using the Amazon Mechanical Turk (AMTurk) crowd sourcing platform, we con-



**S-Figure 1. Collecting human annotations for interactions using Amazon Mechanical Turk platform.** Human subjects were asked to annotate core interaction events in videos from SSv2 dataset [8]. 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).

ducted 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, i.e., ‘contact’ if a target object becomes attached to an agent, or ‘release’ if an object becomes detached from an agent, (iii) the spatiotemporal location of the event, i.e., frame number and image coordinates (see Fig. 1 and Sec. 3).

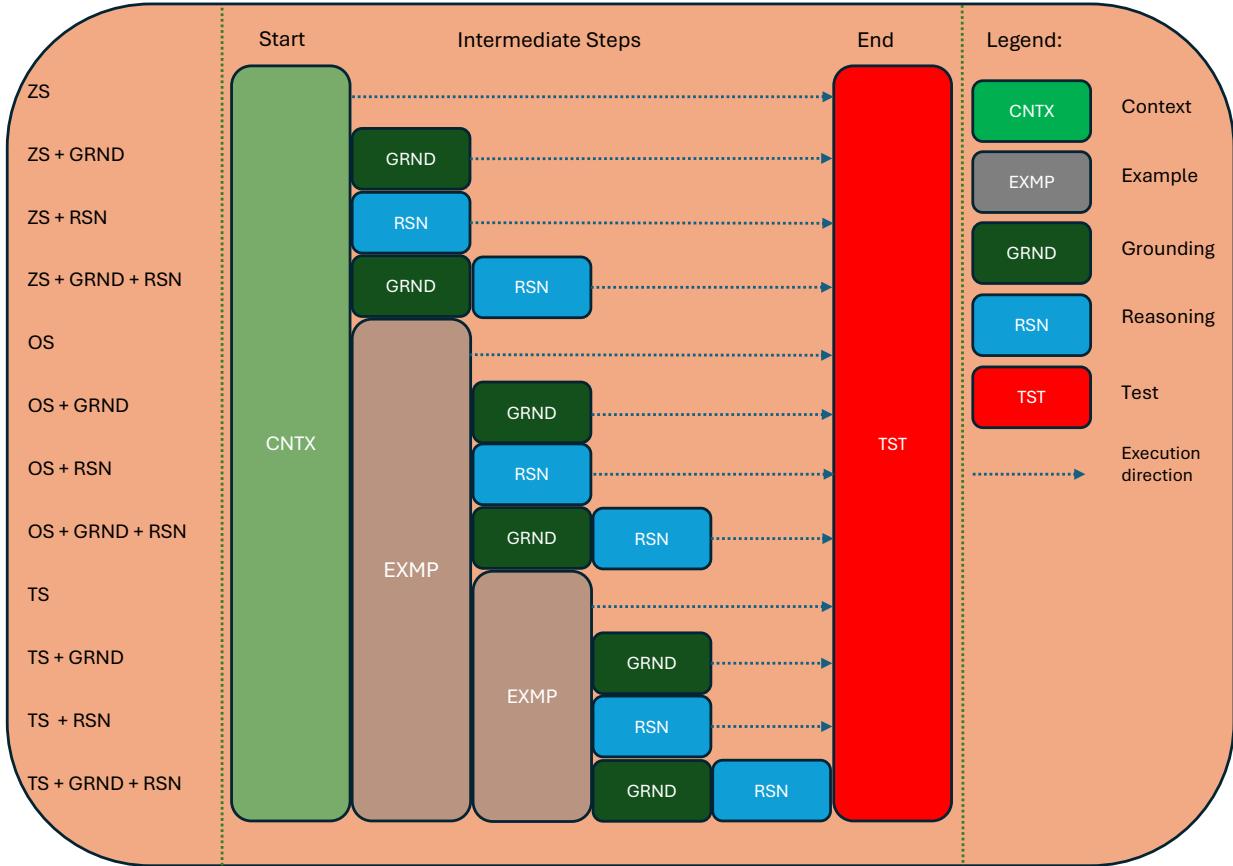
Using the new annotations, we conducted a series of experiments to evaluate the ability of current LMMs to detect the spatio-temporal location of core interaction events in real-world video sequences. The experiments were conducted under several In-Context-Learning regimes, while applying two modifying conditions on the models’ prompts – *Reasoning* and *Grounding* – inspired by earlier studies [10, 23, 27]. In the evaluation, we used an open source version of Alibaba’s Qwen-2.5VL-72B [1] and OpenAI’s online GPT-4o model [18].

In summary our contributions include:

- Introduce a large scale dataset, based on 10K action 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 In-Context-Learning regimes and modifying conditions, including *Reasoning* and *Grounding*.
- A discussion around the question: Does the enhanced generalization behavior of LMMs in action recognition reflect an improved video understanding, or is it merely a convincing ‘story telling’ about detected objects in visual scenes?

## 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. [15], showed that an a ChatGPT agent can successfully answer complex questions when prompted with image data together with the verbal question. Wu et al. [28] 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. [23] 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 [25] 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] employed a two-step reinforcement learning (RL) technique, where in the fist step the LMM agent is prompted to extract a region of interest (ROI) from an image based on the query; then this image crop, together with the actual test image and the associated prompt, are fed into the LMM for the actual answer. They also used the correctness of the prediction of the model on the test task as a guiding signal for models’ improvement. Qi et al. conduct a comprehensive prompt-probing study revealing fundamental limitations in multimodal LLMs’ visual grounding capabilities [19]. Their analysis, which is focused exclusively on image-based task, shows that these models often rely more on textual priors than on actual vi-



S-Figure 2. 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).

sual evidence, and that their visual understanding remains shallow despite producing coherent textual explanations. In contrast, our study investigates grounding in video settings, where models must additionally reason over temporal dynamics, object–object interactions, and event boundaries. This highlights a complementary and more challenging dimension of multi-modal grounding.

### 3. Dataset

In this section we describe the dataset used in our experiments. As we explore the understanding of core interaction elements, for example the moment a hand picks up a target object including the spatial location of the hand and object in the scene, we introduce here a new large scale dataset – The Contact-Release Interaction Dataset. This first of

S-Table 1. **New annotated dataset of core interaction events.** Videos and corresponding labeled action templates were taken from the Something-Something v.2 dataset [8].

<b>Videos</b>	10130
<b>SSv2 action templates</b>	91
<b>Mean videos per template</b>	111
<b>'Contact' events</b>	13816
<b>'Release' events</b>	10406
<b>'Hand-object' interactions</b>	12550
<b>'Object-object' interactions</b>	5653
<b>'Object-surface' interactions</b>	6019

its kind dataset consists of more than 20K annotated interactions on 10,130 action videos from the "Something-Something-V2" (SSv2) dataset [8]. The SSV2 dataset is a collection of more than 200K labeled video clips of humans performing various everyday actions with objects in natural settings. The original dataset labels describe actions as generic templates, for example, "putting something into something", "turning something upside down" and "covering something with something". The labels also include the actual object names in the template placeholder ("something") for each video.

In this work we utilized the AMTurk platform to conduct 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, i.e., 'contact' if a target object becomes attached to an agent, or 'release' if an object becomes detached from an agent, (iii) the spatiotemporal location of the event, i.e., frame number and image coordinates (see Fig. 1). 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. The dataset is publicly available at [ssv2-contact-release-interaction-dataset](#).

## 4. Experimental Design

**Overview.** In this section we describe in more detail the experiments we conducted in the course of the presented study. We evaluated LMMs under three In-Context-Learning (ICL) [6] regimes: zero-shot, one-shot and two-shot.

We are interested in the detection of the spatiotemporal location of core interaction events in video sequences. In our experiments, we focused on the detection of the frame in the video, where a core interaction ('contact' or 'release') event occurs. Since the original videos in the SSV2 dataset may contain several core events, we extracted 3 isolated events for a subset of 33 videos, resulting with our experimental dataset, consisting of 99 short sequences, each with 10 frames consisting of a single core 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. In our experiments we evaluated their relative influence on the models performance. Fig. 2 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

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>"
```

(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.

**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 illustrated in Listing 2. We evaluate the models on each test video in a

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 imgaes ...
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>"
```

**Listing 3.** A grounding prompt 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 detailed reasoning.

```

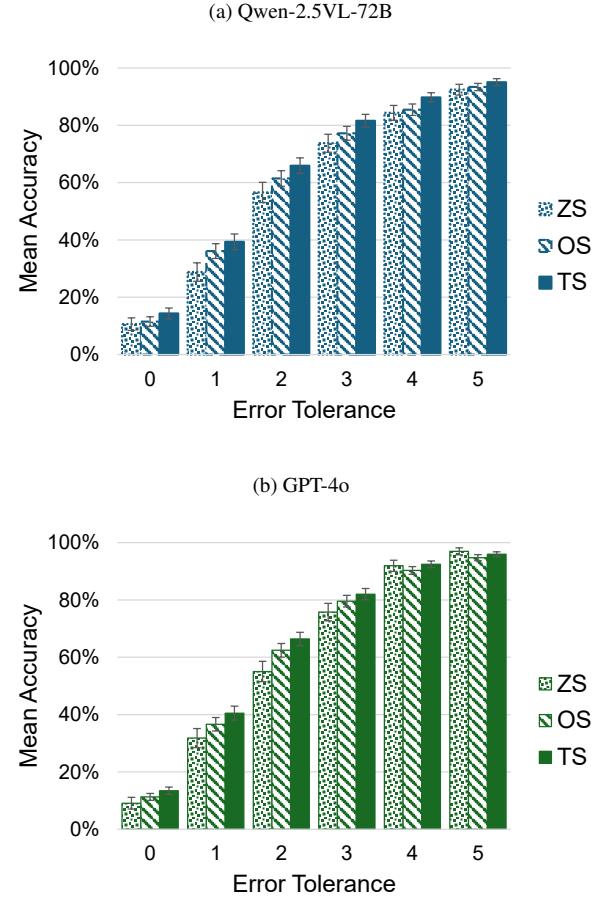
System:
You are ...
Images:
[First test video frame]
...
[Last test video frame]
User:
The uploaded images ...
Q: In which frame ... ?
Answer briefly with ...
Think step by step, and show all intermediate
    ↪ reasoning before giving the final answer.

```

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.

**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 model, 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.



**S-Figure 3. Mean accuracy vs. detection error tolerance.** A correct detection of the models represents a predicted frame within the allowed error tolerance, where an error tolerance of zero means the exact true frame was predicted. Results of Qwen-2.5VL-72B (a) and GPT-4o (b) are shown for the different ICL regimes under the “with reasoning” condition. Note, that the length of all videos in the experimental dataset is 10 frames.

**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. 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 [27].

## 5. Results

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

**S-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 two frames off the true frame). Results are reported as per 3 ICL regimes (ZS, OS, TS) for the models: (i) Qwen-2.5VL-72B and (ii) GPT-4o.

ICL	Reasoning	Mean Accuracy (%)					
		Exact		2-off			
		Qwen	GPT	Qwen	GPT		
ZS	W/O	7 ± 2	10 ± 2	52 ± 4	53 ± 4		
	W	<b>11 ± 2</b>	9 ± 2	<b>57 ± 4</b>	55 ± 4		
OS	W/O	6 ± 1	12 ± 2	53 ± 3	63 ± 3		
	W	<b>12 ± 2</b>	11 ± 1	<b>61 ± 3</b>	62 ± 2		
TS	W/O	10 ± 2	15 ± 2	60 ± 3	67 ± 3		
	W	<b>14 ± 2</b>	13 ± 1	<b>66 ± 3</b>	66 ± 2		

**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  $e_\tau$ . We considered an error tolerance in the range:  $e_\tau \in [0, 5]$ . For example,  $e_\tau = 0$  means the model predicted the true (human) labeled frame where the event occurs, and  $e_\tau = 1$  means that the predicted frame can be up to a frame off the true frame. Throughout this section, the models’ performance is measured as the mean detection accuracy (with standard error) across all test events (99) in the experimental dataset.

Fig. 3 depicts the models’ performance as a function of the allowed error tolerance for the models Qwen-2.5VL-72B [1] and GPT-4o [18] (other models yielded similar results, see Supplementary). The results shown are under the “WITH” *Reasoning* condition (see Sec. 4). The charts show how the very poor performance measured for zero tolerance (TS: 14% and 13% for Qwen and ChatGPT, respectively), increases with a more permitting tolerance, reaching more than 60% for both models. The charts also show a minor increase in performance among the ICL regimes, where TS performs mostly the best and ZS the lowest, as was demonstrated already in earlier studies [6].

Tab. 2 reports the mean accuracy for  $e_\tau = 0$  and  $e_\tau = 2$  (equivalent to Top-1 and Top-5 results) as per each of the three ICL regimes (ZS, OS, TS) and under the *Reasoning* condition (w or w/o reasoning, see Sec. 4). Similarly, Tab. 3 lists the mean accuracy under the *Grounding* condition (see Supplementary for more detailed results).

The results in Tab. 2 indicate that instructing the models to provide reasoning for their answers usually increases their performance. For the Qwen-2.5VL model the perfor-

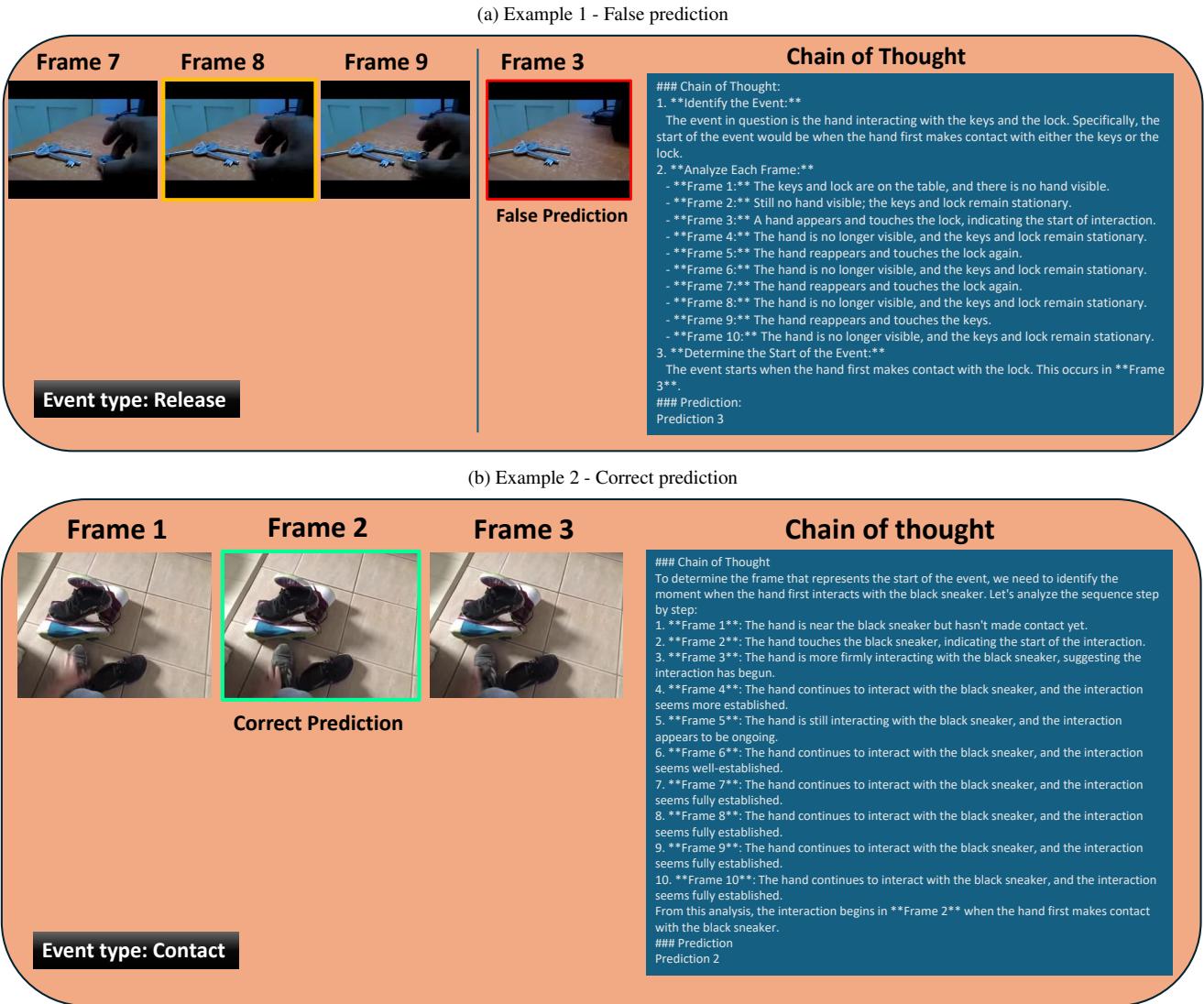
**S-Table 3. Grounding effect on LMMs’ performance.** Prior to the main interaction detection task, models are instructed to name the target object and specify the length of the video sequence, to improve their perceptual grounding. Mean accuracy is measured for the detection of the frame where the test event occurs, within an allowed error tolerance (here the exact true frame or up to 2 frames off). Results are reported as per 3 ICL regimes (ZS, OS, TS) for the models: (i) Qwen-2.5VL-72B and (ii) GPT-4o.

ICL	Grounding	Mean Accuracy (%)					
		Exact		2-off			
		Qwen	GPT	Qwen	GPT		
ZS	W/O	10 ± 2	<b>11 ± 2</b>	<b>56 ± 4</b>	<b>58 ± 4</b>		
	W	8 ± 2	8 ± 2	52 ± 4	51 ± 4		
OS	W/O	9 ± 3	12 ± 1	<b>59 ± 3</b>	63 ± 3		
	W	8 ± 1	11 ± 1	56 ± 3	62 ± 3		
TS	W/O	13 ± 2	14 ± 1	63 ± 3	67 ± 3		
	W	12 ± 2	14 ± 1	63 ± 3	66 ± 3		

mance increase is significant, while for GPT-4o it is not (with respect to the standard error). We performed additional ablation experiment, inspired by recent work on boosting LLMs reasoning with verification feedback and similar to reinforcement learning approach [10, 27, 29], but this did not improve the models’ performance on the detection task (see Supplementary). Surprisingly, the results in Tab. 3 indicate that in contrast to the *Reasoning* condition, instructing the model to provide grounding information from the contents of the instruction prompt and video, yields mostly a decrease in the models’ performance. We further discuss this in Sec. 6.

**Action recognition.** As a comparison with models ability to detect core interaction events, here we tested the models’ ability to recognize the action in the test videos. The SSv2 dataset, includes action template labels, for example, “Putting something, something and something on the table”, each correspond to many videos in the dataset. In addition, the dataset provides a unique label for each video, where the “something” placeholder is replaced with the corresponding object names in that video, for example, “Putting a scale, an eraser and a sd card on the table”.

In this experiment, we evaluated the models on the experimental dataset, which is a subset of the SSv2 dataset, consisting of 33 videos spanning 15 action templates. For template recognition, the prompt included a verbal description of four representative visual examples for a few templates, used solely to illustrate the interaction style. No actual video frames were included. The models were first instructed to classify the general actions depicted in the videos using the predefined list of SSv2 action templates. Impor-



S-Figure 4. **Example predictions of the model Qwen-2.5VL-72B.** The model provides the presented chain-of-thought under the "WITH Reasoning condition. (a) A false prediction. (b) A Correct prediction. The examples show that the reasoning seems logical and realistic, but the relation to the actual video frames is often very loose. Orange and green boxes mark the true frame. Red box marks a false prediction.

tantly, the models were required to output only template-IDs rather than the textual template labels. This allowed us to evaluate predictions by comparing numerical IDs instead of free-form strings, thereby avoiding mismatches caused by formatting, spacing, or minor textual variations.

Using this ID-based evaluation, the Top-1 classification accuracy for the Qwen model was **61 ± 9%**, and **48 ± 9%** for the ChatGPT model. The Top-5 accuracy was **85 ± 6%** and **82 ± 7%** for Qwen and ChatGPT, respectively.

In addition, we evaluated whether the models could correctly identify the objects involved in each interaction. For this task, the model was given the correct action template as input and was asked to replace every occurrence of the

placeholder "something" with the corresponding object(s) appearing in the video. The number of required object names matched the number of placeholder occurrences in the template. To compute accuracy, the predicted placeholders were compared manually against the ground-truth object labels, allowing for differences in ordering and for reasonable synonyms referring to the same object.

The accuracy of correctly naming all objects in a template was **70 ± 8%** for Qwen-2.5VL-72B and **82 ± 7%** for GPT-4o.

**Event bounding-box detection.** We evaluated the models ability to locate the interaction region within the pre-

dicted 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, a square bounding box centered around the annotation point. We tested several box sizes in the range  $20 \times 20$  to  $200 \times 200$  pixels. We computed the Intersection-over-Union (IoU) between the models' predicted bounding boxes and the boxes around the true event position. Using a  $120 \times 120$  pixel size boxes around the true location point yielded mean IoU of only **6.2%** – the best performance in the range of boxes used around the true location. None of the predicted bounding boxes yielded IoU above the common 50% detection threshold. Despite the low IoU scores, the true event location points were included inside the models' predicted bounding boxes in more than 90% of the cases. However, further inspection revealed that the models often predicted over-sized bounding boxes, which in many cases even exceeded the actual image dimensions, indicating that the models internally rescaled the input frames or ignored the original spatial scale. It should be noted, that we explicitly instructed the model to use the original image scale and respond with the smallest possible bounding box. Overall, these findings indicate that the model is unable to reliably localize interaction regions and tends to output bounding boxes whose spatial scale does not correspond to the input frame.

## 6. Discussion

Recent advancement of current computational vision models improve significantly LMMs ability to recognize actions in real-world video sequences with high level of generalization to unseen actions and scenes [22, 24, 26]. Our experiments on action recognition verify this increased performance for the models Qwen-2.5VL-72B and GPT-4o (as well as others, see Supplementary). However, previous studies have indicated a possible limitation of current vision models in understanding the interactions and core events underlying the general action [9]. Do LLMs overcome this limitation by leveraging their vast common knowledge and visual semantics?

In this paper we conducted a series of experiments, under several in-context-learning (ICL) regimes, to test LMMs' (Qwen-2.5VL-72B and GPT-4o) 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 (see Sec. 4). We introduce a new large scale dataset with more than 20K human annotated events in videos from the SSV2 dataset [8] (see Sec. 3).

Our experimental results indicate that despite the ability 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 core in-

teraction events and ground them visually by associating a particular frame and image location, where they occur ( $\leq 11\%$ ). Introducing similar examples using few-shot ICL paradigm slightly improves the performance, which still remains very low ( $\leq 12\%$  in OS;  $\leq 15\%$  with TS).

As demonstrated in earlier studies [27], applying Chain-of-Thought prompting seems to increase the models' reasoning ability ("WITH" *Reasoning* condition in our experiments), but yielded a minor performance increase only for Qwen (+4%, see Tab. 2). Interestingly, explicitly instructing the models to attend and describe information in the instructing prompt and the input video (e.g., name the target objects in the interaction scene), did not improve the models' performance and even yielded lower accuracy in the majority of the experiments (under "WITH" *Grounding* condition, see Tab. 3). These results seem to be somewhat in contrast with previous studies such as by Shao et al. [23]. They suggest that this level of visual grounding is not enough to enhance the models' understanding 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 [30]. 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 [13] and complex dynamic events, despite their huge semantic knowledge. In a sense, the models are merely able to provide a convincing "story" about the interactions based on the reliably detected objects in the scenes and the models' familiarity with verbal action descriptions. 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 [16, 22], 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. Despite the ability to describe the action and name

the participating objects, the models struggle to point back to the spatiotemporal location of core interaction events in videos. We introduce an extension to the SSv2 dataset with more than  $20K$  detailed annotations of such core events in more than  $10K$  videos. These annotations may be used in future efforts to develop new foundation models with better understanding of visual dynamic interactions.

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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, 6
- [2] Renee Baillargeon. Physical reasoning in infancy. *The cognitive neurosciences*, pages 181–204, 1995. 1
- [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. 1
- [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. 1
- [6] 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, 6
- [7] Ping Gong and Xudong Luo. A survey of video action recognition based on deep learning. *Knowledge-Based Systems*, 320:113594, 2025. 1
- [8] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, 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. 1, 2, 3, 4, 8
- [9] 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. 8
- [10] 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, Jiniao 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, Qiushi 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, Xiaosha 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, Yuduan 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, Zhigang 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, 6, 1
- [11] 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. 1
- [12] Alan M Leslie. The perception of causality in infants. *Perception*, 11(2):173–186, 1982. 1
- [13] Jingping Liu, Ziyan 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
- [14] Runze Liu, Yaqun 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. 1
- [15] 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

- [16] Albert Michotte. *The perception of causality*. Basic Books, 1963. [1](#), [8](#)
- [17] 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. [1](#)
- [18] OpenAI. Gpt-4o system card. *arXiv preprint arxiv:2410.21276*, 2024. [2](#), [6](#)
- [19] 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](#)
- [20] Gabriel A Radvansky and Jeffrey M Zacks. *Event cognition*. Oxford University Press, 2014. [1](#)
- [21] Rebecca Saxe and Susan Carey. The perception of causality in infancy. *Acta psychologica*, 123(1-2):144–165, 2006. [1](#)
- [22] 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. [1](#), [8](#)
- [23] 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](#)
- [24] 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. [1](#), [8](#)
- [25] 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](#)
- [26] 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. [1](#), [8](#)
- [27] 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](#), [6](#), [8](#), [1](#)
- [28] Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24710–24721, 2025. [2](#)
- [29] 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. [6](#), [1](#)
- [30] 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](#)
- [31] 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. [1](#)

# While recognizing actions in videos, LMMs struggle to detect core interaction events

## Supplementary Material

### 8. Ablation Experiments

#### 8.1. Action and object recognition

This section provides extended descriptions of the experimental setup of the evaluation on the tasks related to action and object recognition, mentioned in Section 4 of the main manuscript. All experiments were conducted using the extracted video frames of the full length videos from the original SSv2 dataset [8]. The models were given, as input, the sequence of video frames and prompted with the task-specific textual information detailed below. Both **Qwen-2.5-VL-72B** and **GPT-4o** were evaluated under the zero-shot regime (without any examples), and without any of the modifying conditions (*Grounding* and *Reasoning*).

Task-specific textual information was provided only when required:

- **Action-template recognition:** The prompts included the list of candidate template labels and their corresponding template IDs.
- **Object-placeholder extraction:** The prompts included the template ID, the template sentence, and the number of placeholder slots required by the template.
- **Event bounding-box detection:** No additional textual information was provided beyond the sequence of video frames.

**Action and object recognition.** Listing 5 and Listing 8 presents the prompts used in the experiments for testing the models performance in the tasks of object and action recognition, respectively. In the listed prompts, the strings `template_sentence`, `template_id`, and `n_slots` denote variables that were automatically replaced during execution for each experiment, according to the specific interaction template. The prompt structure itself was fixed, while these fields changed to reflect the corresponding template text, its numeric ID, and the number of placeholder slots.

**Event bounding-box detection.** Listing 6 present the prompt used in the experiment for testing the models performance in the task of detecting the spatial location where the event occurs in the predicted frame.

#### 8.2. Tow-Shot with Feedback

Inspired by recent work on boosting LLMs reasoning with verification feedback [10, 27, 29], we designed a variant to the common TS regime, in which the label of the second example was presented to the model indirectly, through an iterative feedback session. In this session, the model performed the detection task on the second example video and then prompted a numerical verification feedback by the user side. The feedback indicate a metric on the gap between the predicted frame ( $f_p$ ) and the true frame ( $f_t$ ) of the event. We defined the error in prediction,  $\epsilon(f_t, f_p)$ , via a sigmoid function, as shown in Eq. (1). The error function is shifted to the middle of the frame range ( $\frac{N}{2}$ ) (as we constrained the error to be in the range of  $[0, 1]$ , and required it to be  $\epsilon(\frac{N}{2}) = 0.5$ ). The score was defined as shown in Eq. (2). The iterative feedback session ended when the model predicted the correct frame, or after it exceeded a limit of allowed trials  $T_{th}$ , which in our experiments was set to 10 (equivalent to the maximal number of iterations in the naive case where the model simply scans all the frames in turn until it gets to the right frame). After the iterative session ended, the model was instructed to perform the main detection task on the test video. In our experiment, we included examples only of other events from the same full video from which the test event was cropped, thus providing the model context with familiar context from the test video. The protocol followed the algorithm in Algorithm 1. The instructing prompt is presented in Listing 7.

$$\epsilon(f_t, f_p) = \sigma(|f_t - f_p| + \frac{N}{2}) = \frac{1}{1 + e^{-|f_t - f_p| + \frac{N}{2}}} \quad (1)$$

$$s(f_t, f_p) = 1 - \epsilon(f_t, f_p) \quad (2)$$

---

**Algorithm 1** Iterative feedback algorithm
 

---

```

Ensure: fbScr = 0.5
Ensure: prvPred0 = 0
Ensure: prvPred1 = 0
while 0.98 > |fbScr| do
    if (sgn(fbScr) > 0) & (prvPred1 ≥ prvPred0) then
        curPred ← [prvPred1, 10]
    else if (sgn(fbScr) > 0) & (prvPred1 ≤ prvPred0) then
        curPred ← [1, prvPred1]
    else if (sgn(fbScr) < 0) & (prvPred1 ≥ prvPred0) then
        curPred ← [prvPred0, prvPred1]
    else if (sgn(fbScr) < 0) & (prvPred1 ≤ prvPred0) then
        curPred ← [prvPred1, prvPred0]
    end if
    prvPred0 ← prvPred1
    prvPred1 ← curPred
    fbScr ← MeasureScore(prvPred0, prvPred1, trueFrame)
end while
  
```

---

**Listing 6.** Instruction prompt for event bounding-box detection.

**System:**  
You are a useful assistant and an expert in video  
    ↳ understanding.  
**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\_0.jpg  
        ↳ is the first frame, followed by frame\_1.jpg  
        ↳ which is the second frame, etc.  
Q: In which frame does the interaction end?  
Answer briefly with: 'Predicted frame: <frame number  
    ↳ only>'  
Q: Where in the predicted frame does the interaction  
    ↳ occur?  
Answer with bounding box coordinates in the original  
    ↳ frame's pixel scale:  
'Predicted BBox: [<left\_x>, <top\_y>, <width>, <height>]'  
Bounding box requirements:  
- Use the SAME coordinate scale as the original video  
    ↳ frame (pixels, not normalized).  
- The box should tightly enclose the MAIN CONTACT REGION  
    ↳ between the interacting objects (e.g., where a  
        ↳ hand touches an object), not the entire objects.  
- Make the bounding box as SMALL as possible while still  
    ↳ fully containing this contact region.  
- The bounding box MUST stay within the frame boundaries  
    ↳ ; never extend beyond the image edges.  
- If uncertain, err on slightly smaller rather than  
    ↳ larger, as long as the contact area is included.

**Listing 5.** Instruction prompt for object detection.

**System:**  
You are an expert video-interaction classifier.  
**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\_0.jpg  
        ↳ is the first frame, followed by frame\_1.jpg  
        ↳ which is the second frame, etc.  
You will see frames from a short video and one template  
    ↳ sentence.  
Template (with placeholders):  
{template\_sentence}  
Return JSON ONLY (no extra text, no markdown fences):  
{  
 "template\_id": {template\_id},  
 "placeholders": [  
 "<slot1>",  
 "<slot2>",  
 ...  
 ]  
}  
Rules:  
- Provide exactly {n\_slots} placeholders, ordered left-to-right as they appear in the template.  
- Use short, concrete noun phrases for visible objects (e.g., "potato", "vicks vaporub bottle").  
- Avoid generic words such as "object", "thing", or "item".  
- Do not include explanations, labels, confidence scores , or any additional fields.

**Listing 7.** Instruction prompt for TS with feedback experiment.

**User:**  
In the next procedure follow these rules:  
1) Your output should ALWAYS be the word "Prediction"  
    ↳ followed by the frame number.  
2) You will be provided a score with absolute values in  
    ↳ the range [0, 1], representing the correctness  
        of your prediction.  
3) Score of 0 means that your prediction is incorrect,  
    ↳ while an absolute value of 1 means you have  
        ↳ found the correct frame.  
4) The absolute values in the range of [0,1] reflect the  
    ↳ proximity of the predicted frame to the true  
        ↳ target frame, where the higher value is better.  
5) The sign of the score signifies the direction for  
    ↳ your next prediction. If the sign is positive,  
        ↳ your next prediction should be in the same  
        ↳ direction as the previous prediction. A negative  
        ↳ sign means you should change the direction of  
        ↳ your next prediction. For example, if your last  
        ↳ prediction was frame 5, your current prediction  
        ↳ is frame 6 and the score is negative - your next  
        ↳ prediction should be smaller than 5. On the  
        ↳ other hand, if the last prediction was frame 7,  
        ↳ the current is frame 4 and your score is  
        ↳ negative - you should predict values greater  
        ↳ than 4.  
6) You should never predict the same frame twice.  
7) Stop your predictions only when the score is above  
    ↳ 0.98.  
Now we will perform an iterative session, during which  
    ↳ you will need to find the frame in the provided  
        ↳ image sequence. Follow the instructions above  
        ↳ when prompted with the feedback score.

**Listing 8. Instruction prompt for action recognition.**

```
System:
You are an expert video-interaction classifier.

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_0.jpg is the first frame, followed by frame_1.jpg which is the second
    ↪ frame, etc.
You will see frames from a short video.
Choose the FIVE best-matching interaction templates (ranked by confidence).
Your goal:
Pick the template ID corresponding to the template label that best describes what happens in the sequence of
    ↪ frames. Focus on the physical interaction between visible objects.
Illustrative Examples (for clarity):
- If the frames show a human hand placing several books one after another on a shelf, the correct template is "
    ↪ Putting number of something onto something" - because the action repeats multiple times and involves a
    ↪ series of objects being placed on another object (the shelf).
- If the frames show a person putting three distinct objects on a table, the correct template is "Putting
    ↪ something, something and something on the table"
- because exactly three objects are placed on the surface at once.
- If the frames show one object being placed next to another, the correct template is "Putting something next to
    ↪ something".
- If the frames show an object being placed inside another, the correct template is "Putting something into
    ↪ something".
Return JSON ONLY (no text or markdown fences):
[
    { "template_id": <int> },
    { "template_id": <int> },
    { "template_id": <int> },
    { "template_id": <int> },
    { "template_id": <int> }
]
Rules:
- Return exactly 5 objects, ranked most->least confident.
- Each object MUST have only one field: "template_id" (integer from the list below).
- Do NOT include any text, explanations, or reasoning.
- Choose IDs based purely on what the video depicts.
- Prefer the main **physical interaction** over camera motion.
Disambiguation:
- Return exactly 5 objects, ranked most->least confident.
- "into" -> containment / inside relation.
- "onto" -> on top of.
- "next to" -> lateral adjacency without contact stacking.
- "slanted surface" -> object accelerates along a plane.
- "on a flat surface w/o rolling" -> stable placement without motion.
Scoring & Specificity Rules (very important):
- Prefer the template whose action AND number of involved objects best match the scene.
- If THREE distinct objects interact, prefer a 3-object template over any 2-object option, if the action fits.
- Tie-breakers: object count > verb precision > physical outcome > surface/slant qualifiers.
- Don't pick a broader template if a more specific one fits.
Templates (id: label):
    1: Attaching something to something
    33: Moving part of something
    48: Piling something up
    54: Poking a stack of something without the stack collapsing
    57: Poking something so that it falls over
    58: Poking something so that it spins around
    88: Pulling something onto something
    97: Pushing something so it spins
    98: Pushing something so that it almost falls off but doesn't
    99: Pushing something so that it falls off the table
    102: Putting number of something onto something
    120: Putting something, something and something on the table
    122: Rolling something on a flat surface
    144: Stacking number of something
    148: Taking something out of something
```

## 9. Detailed results

**Action and object recognition.** For completeness, we also evaluated GPT-5 on the action-recognition task under the same zero-shot protocol described in Section 4 of the main manuscript. GPT-5 obtained a Top-1 accuracy of **63.6%** and a perfect Top-5 accuracy of **100%**, outperforming both GPT-4o and Qwen-2.5VL on this subset.

**Event bounding-box detection.** We first examine a basic property: whether the model-predicted bounding box contains the true event location point. We additionally evaluated GPT-5 using the same setup, and include its results alongside GPT-4o and Qwen-2.5VL. S-Table 4 reports, for each model, the fraction of ground-truth event points that fall inside the predicted bounding box.

Model	Total label points in bbox	Percentage
GPT-4o	56 / 99	56.57%
GPT-5	86 / 99	86.87%
Qwen-2.5VL	99 / 99	100%

S-Table 4. Ground-truth event point locations contained inside the model-predicted bounding boxes.

Despite this relatively high containment rate (especially for Qwen), models often predicted over-sized bounding boxes, which in many cases even exceeded the actual image dimensions. S-Table 5 presents a per-model breakdown of the Intersection-over-Union (IoU) between the predicted box and a  $120 \times 120$  pixel ground-truth box centered on the annotated event location.

Model	Mean IoU (%)	#IoU $\geq 50\% / 99$
GPT-4o	1.48%	0 / 99
GPT-5	11.23%	0 / 99
Qwen-2.5VL	9.55%	1 / 99

S-Table 5. Per-model IoU statistics for the event localization task.

Although the true event point often lies inside the predicted box (S-Table 4), the IoU values are extremely low (S-Table 5). This discrepancy indicates that the models tend to predict overly large bounding boxes, possibly ignoring detections of the participating objects and hands involved in the interactions. In many cases the predicted event bounding box extended beyond the frame boundaries despite explicit instructions to preserve the original spatial scale (see Listing 6). Overall, all those models struggle to localize the interaction event regions reliably.

**Reasoning and Grounding conditions.** S-Table 7 lists the detailed results from our experiments with (W) and

without (W/O) the two modifying conditions: *Reasoning* and *Grounding*. The table includes the results under the three ICL regimes (ZS, OS, TS) for the models Qwen-2.5VL-72B and GPT-4o. The mean accuracy is evaluated for the detection of the exact true frame, where a core interaction event occurs, and also for a detection within an error tolerance of two frames off.

**Example predictions and associated CoT.** S-Figure 5 and S-Figure 6 show additional examples of false and correct frame predictions. The examples include the models’ Chain-of-Thought, which seemingly presents a logical reasoning text for detecting an interaction event, but the grounding to the video frames is often very loose.

**Two-shot with feedback.** The results of this experiment showed no improvement in the model’s performance on the test task of predicting the frame where an event occurs in the test video. The mean accuracy is reported in Tab. 6. The results suggest that the feedback session may even interfere with the main prediction task, by shifting away the model from the visual input to the number of the frame, while trying to maximize the feedback score, which is a metric on the prediction error of the frame number.

Nevertheless, an analysis of the model’s weighted success rate and its test error shown in Fig. 7, indicate that when the model is required to provide reasoning which lead to its predictions, the test error remains low up until the 6th feedback iterations, suggesting that some learning may occur. However, as discussed in the main text, the loose grounding to the visual input, becomes even worst with this feedback approach since the focus of the model is drawn away from the image contents, trying to satisfy the feedback score around the frame number, rather than grounded visual cues.

The weighted success rate of a test trial (see Fig. 7a,b)

S-Table 6. **Effect of feedback on the model’s performance.** The true label of the second example is provided through an iterative session with feedback, indicating the gap between the predicted and the true frame. Mean accuracy is measured for the detection of the frame where the test event occurs, within an allowed error tolerance (here the exact true frame or up to 2 frames off). The evaluation is performed under the two modifying conditions: *Reasoning* and *Grounding*. The evaluated model is Qwen-2.5VL-72B.

Reasoning	Grounding	Mean Accuracy (%)	
		Exact	2-off
W/O	W/O	$4 \pm 2$	$29 \pm 4$
W/O	W	$4 \pm 2$	$29 \pm 4$
W	W/O	$9 \pm 2$	$53 \pm 4$
W	W	$9 \pm 2$	$50 \pm 4$

was calculated with the conditioned probability as follows

$$p(\text{success} = n | \text{iterations} = k) = \frac{n}{N} \frac{n}{N_k} = \frac{n^2}{NN_k} \quad (3)$$

where  $N$  is the total number of successful test trials (i.e., where the agent predicted the correct frame in the test task), and  $N_k$  is the number of successful test predictions conditioned on training session having  $k$  iterations.

## 10. Experimental Dataset

S-Table 8 lists all 33 videos included in the experimental dataset, which was used in our experiments. For each video, we cropped short 10-frame video clips around three temporally separated core interaction events. The table includes the video ID, action template and object placeholders from the original SSv2 dataset. In addition, for each short video clip, the table includes the crop start frame, the frame where the event occurs and the type of the event, i.e., 'contact' or 'release'. The full annotations are available online at: [ssv2-contact-release-interaction-dataset](#).

Event type: Release	Frame 4	Frame 5	Frame 6	Frame 3	Chain of Thought
					<p>To determine the frame number that represents the start of an event, let's analyze the sequence of actions depicted in the frames:</p> <ol style="list-style-type: none"> <li>**Frame 1**: The hand is positioned above the white circular object, indicating an intention to interact with it.</li> <li>**Frame 2**: The hand moves closer to the white circular object.</li> <li>**Frame 3**: The hand makes contact with the white circular object.</li> <li>**Frame 4**: The hand lifts the white circular object.</li> <li>**Frame 5**: The hand places the white circular object back on the surface.</li> <li>**Frame 6**: The hand moves away from the white circular object.</li> <li>**Frame 7**: The hand returns to the white circular object.</li> <li>**Frame 8**: The hand lifts the white circular object again.</li> <li>**Frame 9**: The hand places the white circular object back on the surface.</li> <li>**Frame 10**: The hand moves away from the white circular object.</li> </ol> <p>### Chain of Thought</p> <ul style="list-style-type: none"> <li>- The event of interest here is the interaction between the hand and the white circular object.</li> <li>- The interaction begins when the hand first makes contact with the white circular object.</li> <li>- This contact occurs in Frame 3.</li> </ul> <p>### Prediction</p> <p>Prediction 3</p>
	Frame 5	Frame 6	Frame 7	Frame 1	Chain of Thought
					<p>To determine the frame number that represents the start of an event, let's analyze the sequence of images step by step:</p> <p>### Chain of Thought:</p> <ol style="list-style-type: none"> <li>**Frame 1**: A hand is holding a small black object near a white object on a blue circular surface. The hand is positioned as if it is about to place the black object down.</li> <li>**Frame 2**: The hand is still holding the black object, and it appears to be in the process of placing it on the blue surface.</li> <li>**Frame 3**: The hand is still holding the black object, and it appears to be in the process of placing it on the blue surface.</li> <li>**Frame 4**: The hand is still holding the black object, and it appears to be in the process of placing it on the blue surface.</li> <li>**Frame 5**: The hand is no longer visible, and the black object is now resting on the blue surface. The white object remains in place.</li> <li>**Frame 6**: The hand reappears, holding the black object again, suggesting a new interaction.</li> <li>**Frame 7**: The hand is no longer visible, and the black object is still resting on the blue surface. The white object remains in place.</li> <li>**Frame 8**: The hand is no longer visible, and the black object is still resting on the blue surface. The white object remains in place.</li> <li>**Frame 9**: The hand is no longer visible, and the black object is still resting on the blue surface. The white object remains in place.</li> <li>**Frame 10**: The hand reappears, holding the black object again, suggesting a new interaction.</li> </ol> <p>From the analysis, the start of an event occurs when the hand first appears holding the black object and moves it towards the blue surface. This happens in Frame 1, Frame 6, and Frame 10. Since we need to find the earliest start of an event, we focus on Frame 1.</p> <p>### Prediction:</p> <p>Prediction 1</p>
	Frame 8	Frame 9	Frame 10	Frame 1	Chain of Thought
					<p>### Chain of Thought:</p> <ol style="list-style-type: none"> <li>**Identify the Event**: <ul style="list-style-type: none"> <li>- The event in question is the hand interacting with the container. Specifically, the hand is pointing at or touching the container.</li> </ul> </li> <li>**Temporal Analysis**: <ul style="list-style-type: none"> <li>- In Frame 1, the hand is already in contact with the container.</li> <li>- From Frame 2 onwards, the hand continues to interact with the container in a similar manner.</li> </ul> </li> <li>**Determine the Start of the Event**: <ul style="list-style-type: none"> <li>- Since the hand is already in contact with the container in Frame 1, this indicates that the interaction has already started.</li> <li>- There is no clear indication of a prior state where the hand was not interacting with the container.</li> </ul> </li> <li>**Conclusion**: <ul style="list-style-type: none"> <li>- The start of the event, which is the hand beginning to interact with the container, is in Frame 1.</li> </ul> </li> </ol> <p>### Prediction:</p> <p>Prediction 1</p>

S-Figure 5. **Example false predictions of the model Qwen-2.5VL-72B.** The model provides the presented chain-of-thought under the "WITH" Reasoning condition. The examples show that the reasoning text seems logical and realistic, but the relation to the actual video frames is often very loose. A red box marks a false prediction, while the orange box marks the true frame.

Frame 5	Frame 6	Frame 7	Chain of Thought
			<p>### Chain of Thought</p> <p>To determine the frame number that represents the start of an event, we need to identify the first frame where a noticeable change or action occurs compared to the previous frames. Let's analyze the sequence step by step:</p> <ol style="list-style-type: none"> <li>**Initial State (Frames 1-4)**: - The hand is consistently touching the scissors' blade.</li> <li>**Transition (Frame 5)**: - The hand moves away from the scissors' blade.</li> <li>**Subsequent Frames (Frames 6-10)**: - The hand returns to touch the scissors' blade in Frame 6.</li> <li>**Conclusion** The start of the event is marked by the hand moving away from the scissors' blade for the first time. This occurs in **Frame 5**.</li> </ol> <p>### Prediction</p> <p>Prediction 5</p>
False Prediction	True Frame		

Event type: Release

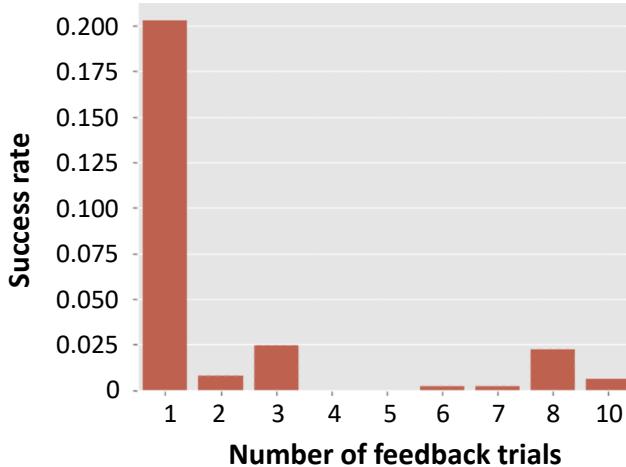
Frame 1	Frame 2	Frame 3	Chain of Thought
			<p>### Chain of Thought</p> <p>To determine the frame that represents the start of the event, we need to identify the first frame where the hand begins interacting with the pen. Here is a detailed analysis of the sequence:</p> <ol style="list-style-type: none"> <li>**Frame 1**: The hand is approaching the pen, but there is no clear contact yet.</li> <li>**Frame 2**: The hand is closer to the pen, but still no contact.</li> <li>**Frame 3**: The hand is making contact with the pen, indicating the start of the interaction.</li> <li>**Frames 4-10**: The hand continues to hold and manipulate the pen, showing ongoing interaction.</li> </ol> <p>Given this analysis, the interaction starts in Frame 3 when the hand makes contact with the pen.</p> <p>### Prediction</p> <p>Prediction 3</p>
True Frame		False Prediction	

Event type: Contact

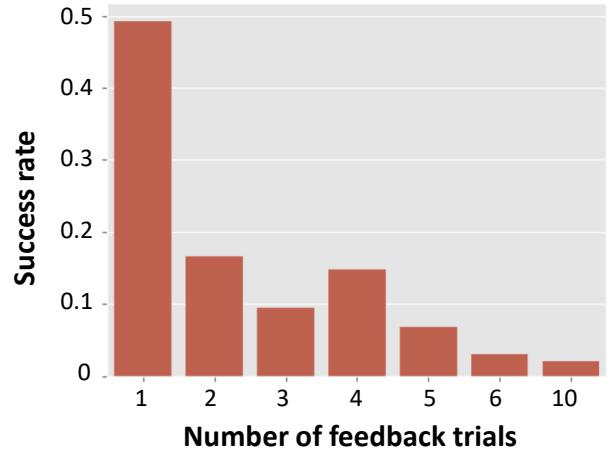
Frame 7	Frame 8	Frame 9	Chain of Thought
			<p>### Chain of Thought</p> <p>To identify the frame number that represents the start of an event in the given sequence, let's analyze the changes between consecutive frames:</p> <ol style="list-style-type: none"> <li>**Frames 1 to 7**: The hand is consistently interacting with the fork, holding it in place. There is no significant change in the position or state of the objects.</li> <li>**Frame 8**: The hand begins to move away from the fork, indicating the start of a new event where the hand is no longer interacting with the fork.</li> <li>**Frames 9 and 10**: The hand continues to move away, confirming the continuation of the event started in Frame 8.</li> </ol> <p>Based on this analysis, the start of the event where the hand begins to move away from the fork is in **Frame 8**.</p> <p>### Prediction</p> <p>Prediction 8</p>
Correct Prediction			

Event type: Release

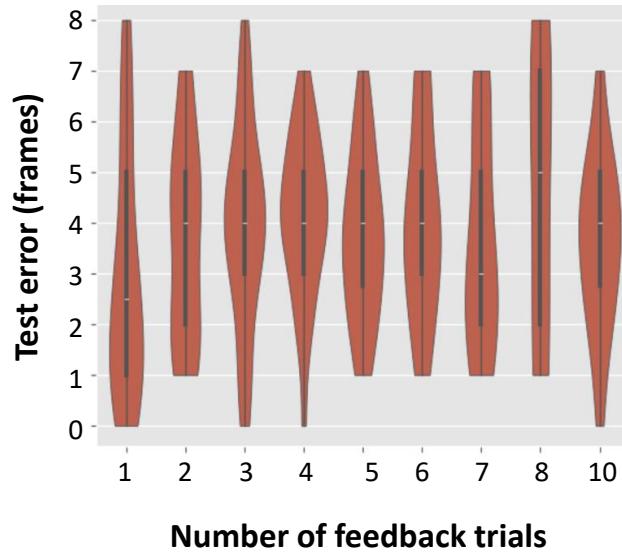
S-Figure 6. Additional example predictions of the model Qwen-2.5VL-72B. The top two examples show false predictions of the model, in which the predicted frame was one frame before or after the true frame, where the event occurs. Humans can see clearly the moment of release or contact in the true frames, but the visual cues are too subtle for the models to detect. The Chain-of-Thought text lists the correct reasoning flow, but is not well grounded in the video frames. Orange and green boxes mark the true frame. Red boxes mark false predictions.



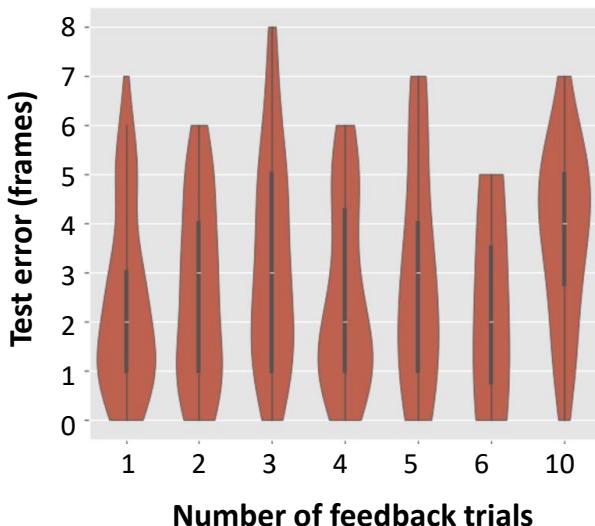
(a) Test prediction success rate - Without *Reasoning*



(b) Test prediction success rate - With *Reasoning*



(c) Distribution of test error (frames) - Without *Reasoning*



(d) Distribution of test error (frames) - With *Reasoning*

**S-Figure 7. Effect of feedback on test success rate and error distribution.** In the TS with feedback experiment, the model goes through an iterative session, when it is provided with a feedback score until it predicts correctly the event frame of the second example. (a-b) The model's weighted success rate (see Sec. 9) on the test video vs. the number of trials in the feedback session, without and with *Reasoning*, respectively. (c-d) The distribution of the test error (in frames) without and with *Reasoning*, respectively. The results show that when the model is not required to provide reasoning for its prediction, the feedback session does not improve the model's performance in the test task and the test error increases with the number of feedback trials (a, c). On the other hand, when the model is required to provide reasoning, the test success rate decreases monotonically with the number of feedback trials up till the 6th iteration.

S-Table 7. **Reasoning and Grounding conditions' effect on LMMs' performance.** In the "WITH" (W) *Reasoning* condition, models are instructed to describe in step-by-step the reasoning behind their prediction of the frame where the interaction occurs. In the "WITH" *Grounding* condition, prior to the main interaction detection task, models are instructed to name the target object and specify the length of the video sequence, 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 two frames off the true frame). Results are reported as per 3 ICL regimes (ZS, OS, TS) for the models: (i) Qwen-2.5VL-72B and (ii) GPT-4o.

ICL Regime	Reasoning	Grounding	Mean Accuracy (%)			
			Exact		2-off	
			Qwen-2.5VL	GPT-4o	Qwen-2.5VL	GPT-4o
<b>Zero Shot</b>	W/O	W/O	8.1 ± 2.8	11.1 ± 3.2	57.6 ± 5.0	56.6 ± 5.0
	W/O	W	6.1 ± 2.4	8.1 ± 2.8	46.5 ± 5.0	49.5 ± 5.1
	W	W/O	11.1 ± 3.2	10.1 ± 3.0	54.5 ± 5.0	58.6 ± 5.0
	W	W	10.1 ± 3.0	8.1 ± 2.8	58.6 ± 5.0	51.5 ± 5.0
<b>One Shot</b>	W/O	W/O	7.2 ± 2.1	13.0 ± 2.2	56.0 ± 4.4	64.3 ± 3.8
	W/O	W	5.2 ± 1.8	11.7 ± 2.2	51.2 ± 4.4	61.9 ± 4.0
	W	W/O	11.6 ± 2.3	11.3 ± 1.7	62.1 ± 3.8	62.4 ± 3.4
	W	W	11.5 ± 2.4	11.3 ± 1.7	60.8 ± 3.9	62.4 ± 3.4
<b>Two Shot</b>	W/O	W/O	12.2 ± 2.6	14.7 ± 2.2	61.8 ± 4.6	66.9 ± 3.8
	W/O	W	8.6 ± 2.2	14.8 ± 2.4	58.8 ± 4.6	66.3 ± 3.8
	W	W/O	14.0 ± 2.6	13.4 ± 1.8	64.9 ± 3.9	66.6 ± 3.3
	W	W	14.7 ± 2.8	13.5 ± 1.8	66.9 ± 3.8	66.2 ± 3.3

S-Table 8. Experimental dataset information.

Video ID	Action template	Object placeholders	Event clip (10 frames)		
			Start frame	Event frame	Event type
1979	Putting [something], [something] and [something] on the table	scale, eraser, sd card	12 40 60	14 39 55	release release release
2648	Attaching [something] to [something]	dummy peach, peach tree	11 30 46	18 36 52	contact contact release
3996	Putting number of [something] onto [something]	books, shelf	8 29 47	13 33 50	release release release
4042	Pushing [something] so it spins	green candy	7 18 27	10 21 34	release contact release
4144	Poking [something] so that it falls over	pen	12 19 32	20 22 34	contact release release
9257	Piling [something] up	kool-aid packs	8 34 52	14 38 57	release release release
12492	Putting [something], [something] and [something] on the table	keys, lock, bulb	1 10 28	7 17 34	release release release
14990	Putting [something], [something] and [something] on the table	perfume bottle, naphthalene ball, silver ring	22 32 48	25 37 52	release release release
17127	Putting [something], [something] and [something] on the table	prescribers guide book, medicine bottle, vape pen	13 36 56	17 39 62	release release release
26039	Pushing [something] so that it falls off the table	toy	1 13 24	9 20 30	contact contact contact
30880	Putting [something], [something] and [something] on the table	scissors, cookie cutter, grater	12 35 56	17 39 62	release release release
41434	Stacking [number of] [something]	3, coins	0 17 37	4 19 43	contact contact contact
57029	Taking [something] out of [something]	tools, toolbox	0 14 40	1 17 45	contact contact contact
66464	Moving [part] of [something]	tuner, electric guitar	12 26 36	14 32 40	contact release contact
67618	Putting [something], [something] and [something] on the table	bottle, tube, purse	0 13 40	1 21 47	release release release
73232	Taking [something] out of [something]	cd, book	4 11 20	8 18 23	contact contact release
74722	Taking [something] out of [something]	phone, drawer	0 24 52	8 29 56	contact contact release

Continued on the next page

S-Table 8. . (cont.)

Video ID	Action template	Object placeholders	Event clip (10 frames)		
			Start frame	Event frame	Event type
84410	Attaching [something] to [something]	pen's cover, pen	6 18 27	10 25 28	contact release contact
87327	Putting [something], [something] and [something] on the table	grater, whisk, corkscrew	4 28 42	11 31 50	release release release
92626	Poking [something] so that it spins around	flashlight	2 29 41	7 31 44	contact release contact
95238	Attaching [something] to [something]	toy train engine, its coach	3 13 40	7 18 43	contact contact contact
96903	Rolling [something] on a flat surface	perfume	2 15 25	7 18 29	contact contact contact
153413	Putting [something], [something] and [something] on the table	fork, spoon, dish	10 27 54	17 33 61	release release release
158080	Putting [something], [something] and [something] on the table	toothpick container, showpiece, padlock	2 21 44	6 27 47	release release release
158915	Putting [something], [something] and [something] on the table	mug, spoon, gum	1 20 31	5 25 34	release release contact
163090	Putting [something], [something] and [something] on the table	popcorn, vicks vaporub bottle, purple water bottle	5 26 42	10 29 45	release release release
164784	Pushing [something] so that it almost falls off but doesn't	roll	3 30 43	6 35 48	contact release contact
166894	Poking a stack of [something] without the stack collapsing	lincoln logs	15 38 46	19 41 51	contact release contact
175159	Stacking [number of] [something]	5, hot pads	6 20 39	9 26 41	contact contact contact
175167	Piling [something] up	water color containers	4 25 41	8 31 44	contact release contact
181367	Piling [something] up	shoes	0 19 29	2 26 30	contact release contact
186500	Pulling [something] onto [something]	nail clipper, envelope	11 21 29	17 26 31	contact release contact
217743	Putting [something], [something] and [something] on the table	glass vase, child's shoe, coffee mug	21 41 58	24 46 65	release release release