HanDyVQA: A Video QA Benchmark for Fine-Grained Hand-Object Interaction Dynamics

Masatoshi Tateno[1,2] **Gido Kato**[3,2] **Kensho Hara**[2]
Hirokatsu Kataoka[2,4] **Yoichi Sato**[1] **Takuma Yagi**[2]
[1]Institute of Industrial Science, The University of Tokyo
[2]National Institute of Advanced Industrial Science and Technology (AIST)
[3]Waseda University [4]Visual Geometry Group, University of Oxford
{masatate,ysato}@iis.u-tokyo.ac.jp
{takuma.yagi,katou.1999,kensho.hara,hirokatsu.kataoka}@aist.go.jp

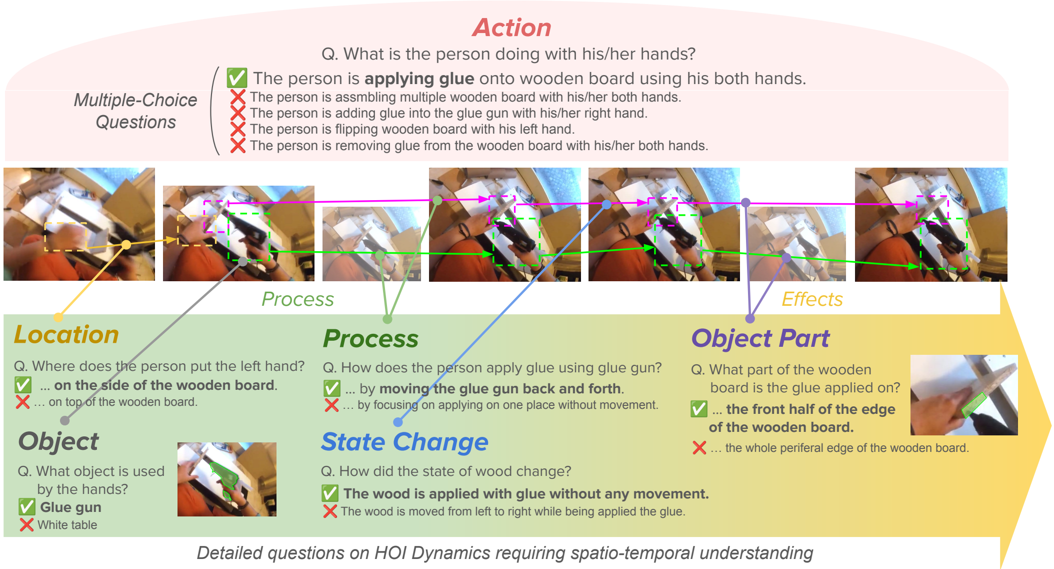

Action
Q. What is the person doing with his/her hands?
Multiple-Choice Questions
✅ The person is **applying glue** onto wooden board using his both hands.
❌ The person is assmbling multiple wooden board with his/her both hands.
❌ The person is adding glue into the glue gun with his/her right hand.
❌ The person is flipping wooden board with his left hand.
❌ The person is removing glue from the wooden board with his/her both hands.
Process
Effects
Location
Q. Where does the person put the left hand?
✅ ... **on the side of the wooden board**.
❌ ... on top of the wooden board.
Object
Q. What object is used by the hands?
✅ **Glue gun**
❌ White table
Process
Q. How does the person apply glue using glue gun?
✅ ... by **moving the glue gun back and forth**.
❌ ... by focusing on applying on one place without movement.
State Change
Q. How did the state of wood change?
✅ **The wood is applied with glue without any movement.**
❌ The wood is moved from left to right while being applied the glue.
Object Part
Q. What part of the wooden board is the glue applied on?
✅ ... **the front half of the edge of the wooden board.**
❌ ... the whole periferal edge of the wooden board.
Detailed questions on HOI Dynamics requiring spatio-temporal understanding

Figure 1: Overview of HandyVQA Dataset.

Abstract

Hand-Object Interaction (HOI) is inherently a dynamic process, involving nuanced spatial coordination, diverse manipulation styles, and influences on interacting objects. However, existing HOI benchmarks tend to emphasize high-level action recognition and hand/object localization while neglecting the fine-grained aspects of hand-object dynamics. We introduce HanDyVQA, a video question-answering benchmark for understanding the fine-grained spatiotemporal dynamics in hand-object interactions. HanDyVQA consists of six types of questions (Action, Process, Objects, Location, State Change, and Object Parts), totaling 11.7k multiple-choice question-answer pairs and 11k instance segmentations that require discerning fine-grained action contexts, hand-object movements, and state changes caused by manipulation. We evaluated several video foundation models on our benchmark and found that even the powerful Qwen2.5-VL-72B reached only 68.8% average accuracy, uncovering new challenges in component-level geometric and semantic understanding through extensive analyses.

Preprint. Under review.

	Objective	Source	View	Question Scope					Answer Type	#Questions	Avg. Duration
				HOI	Spatial	Temporal	Process	Effect			
Next-QA [45]	Causal / Temporal / Descriptive	YFCC-100M	TPV	✗	✗	✓	✓	✓	MC + OP	52K	44 s
EgoTaskQA [16]	Spatial / Temporal / Causal	LEMMA	FPV	✗	✓	✓	✗	✓	OP	40K	25 s
EgoSchema [22]	Long-Term Reasoning	Ego4D	FPV	✗	✗	✓	✓	✗	MC	5K	180 s
MVBench [18]	Spatial / Temporal	Mixed	TPV	✗	✓	✓	✗	✓	MC	4K	5–8 s
EgoThink [4]	Reasoning / Forecasting / Planning	Ego4D	FPV	✗	✓	✓	✗	✗	OP	700	Single frame
HOI-QA [2]	Hand and Object Location Referral	EK/Ego4D	FPV	✓	✓	✗	✗	✗	OP + BBox	3.9M	Single frame
EgoHOIBench [46]	Action / Objects	Ego4D	FPV	✓	✓	✓	✗	✗	MC	30K	1 s
AMB [12]	Long-Term Object Interactions	EK	FPV	✓	✓	✓	✗	✗	MC	21K	20 m
HD-EPIC [27]	Fine-grained Video / 3D Understanding	HD-EPIC	FPV	✓	✓	✓	✓	✗	MC	26K	Variable
HanDyVQA(Ours)	Dynamics / Processes / Effects	Ego4D	FPV	✓	✓	✓	✓	✓	MC + Seg	12K	5 s

Table 1: Comparison agaist related QA datasets: TPV/FPV refers to third-person-view and first-person-view videos, respectively. MC stands for multiple-choice question-answering and OP represents open-ended question-answering. BBox indicates bounding box, and Seg refers to segmentation.

1 Introduction

Hand-Object Interaction (HOI) is inherently a dynamic process [11]. To perform tasks with precision, people choose appropriate tools, carefully coordinate their hands, tools, and objects, and modify the environment to accomplish their goals. Accurately recognizing the spatiotemporal dynamics of hand-object interactions opens up various applications, such as worker assistance [10], dexterous manipulation in robots [36], and motor function analysis [40].

While there has been a surge in hand-object interaction recognition methods and benchmarks in recent years, they tend to focus on either (i) high-level action recognition such as action recognition [7, 17, 13, 46], long-form actions [22], and procedural steps [34, 37, 50] or (ii) low-level localization such as hand-object localization [35, 5, 2] and hand pose estimation [25, 9, 55] while neglecting the semantically rich aspects of hand-object dynamics.

We propose HanDyVQA (**Hand Dy**namics **V**ideo **QA**), a video question-answering benchmark designed to evaluate spatiotemporal reasoning in dynamics of HOI (see Figure 1). HanDyVQA requires an understsnding not only of the actions and objects involved but also of their processes, effects, and component-level changes. We built the benchmark on short video clips extracted from the Ego4D [13] dataset, which fearures diverse and natural hand-object interactions in real-world settings which may not be recorded in intentionally filmed footage. We provide six types of multi-choice question answering (MCQ) tasks totalling 11.7k carefully designed QA pairs that avoid trivial shortcuts, along with referred video object segmentation (RVOS) tasks for two question types (Objects and Object Parts) totalling 11k instances to directly evaluate spatial understanding.

We evaluate existing video-language models to quantify how well they capture various aspects of hand-object interactions. Our results indicate that even the latest foundation models struggle across all categories, achieving only around 61–77% accuracy in MCQ even with the powerful Qwen2.5-VL-72B model. Ablation studies on the number of input frames and image resolution reveals that spatiotemporally dense inputs are necessary to boost the performance. The results on the RVOS task also suggest that current models fail in referring local components finer than object-level.

Furthermore, to advance the understanding of dynamic HOI phenomena, we evaluate HanDyVQA to investigate whether explicitly feeding (i) hand pose, (ii) object tracking, and (iii) object features can enhance the performance or not. The results reveal additional modalities indeed improve performance in many categories, suggesting more sophisticated video encoder design to include local hand-object information towards understanding HOI dynamics.

In summary, our contribution is as follows: (a) We introduce HanDyVQA, a new comprehensive dataset for understanding fine-grained dynamics in HOIs. (b) We conduct an in-depth analysis of how latest video-language models struggle to capture spatiotemporal dynamics and pixel-level reasoning in HOI. (c) We show fine-tuning models with additional hand and object information can enhance the performance, showing the necessity of modeling fine-grained temporal evolution of hands, objects and their components towards further understanding of HOI dynamics.

2 Related Work

Video question answering benchmarks Conventional benchmarks [47, 52] focus on questions that identify human actions, events, or objects occurring within short video clips of a few seconds. As the field evolves, recent benchmarks have addressed more challenging tasks, such as long-form video understanding [44, 42, 59, 22]. Some works, such as NExT-QA [45] and TimeLogic QA [39], focus on temporal and causal relationships between multiple actions. MVBench [18] proposes a challenging set of temporal understanding tasks in a multiple-choice QA format that requires watching the entire video by curating major third-person video datasets. HD-EPIC [27] provides a wide variety of fine-grained QAs of egocentric video in a kitchen scenario. However, none of these benchmarks focuses on the fine-grained details of HOIs, including the local coordination between hands and objects, the subtle ways they are handled, and the resulting effects across diverse scenarios.

Hand-object interaction recognition benchmarks Various HOI recognition benchmarks have been proposed for applications such as AR/VR and robotics, with focuses on (i) low-level localization and (ii) high-level actions. For the former, benchmarks have focused on detecting hands and contact objects [35], estimating 3D hand and object poses [14, 3], reconstructing mesh representations [38], and object tracking [1, 12]. AMEGO [12] collects long-term hand and object tracks from the EPIC-KITCHENS dataset and curates a set of questions that require localizing the positions and moments of objects in interaction. For the latter, several benchmarks [22, 46, 2, 4, 12] have been built on egocentric video datasets such as EPIC-KITCHENS [7] and Ego4D [13], since egocentric videos capture close-ups of hands and objects in manipulation. EgoHOIBench [46] introduces an open-vocabulary HOI recognition task that addresses questions about the actions and objects involved in very short (1s) egocentric videos. HOI-QA [2] studies the task of referring to hands and objects in egocentric images, evaluating the relationships between entities and their locations. While these works cover some of the components crucial for HOI understanding, none address the fine-grained nuances of HOIs, such as processes, effects, and component-level spatiotemporal understanding.

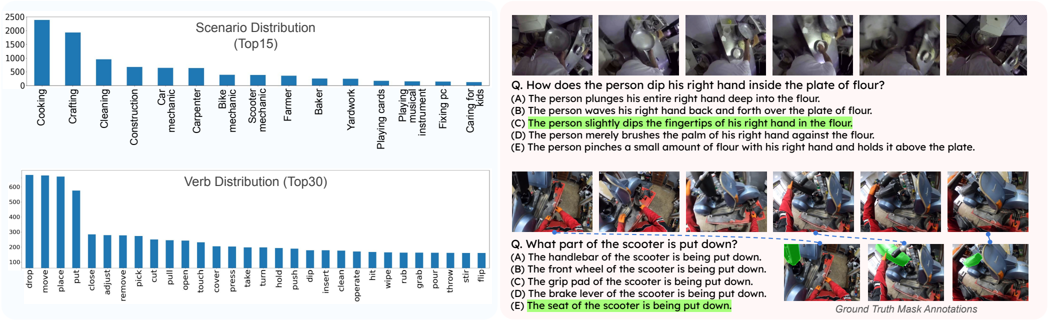
Vision-and-language models for video understanding The emergence of dual-encoder vision-language models trained on large-scale image-text pairs [31, 54] has spurred significant advancements in video understanding. Various approaches have been explored, including adapting image-based models for video tasks [21, 49], training models on instructional videos with web-based narrations [24, 23], and pretraining first-person, video-specific models [20, 30, 58]. Following the success of large language models (LLMs), recent video-language models integrate these pretrained visual encoders and LLMs to achieve general video comprehension capabilities across various downstream tasks [2, 57, 6, 51, 41]. Their primary efforts lie in increasing model parameter sizes and expanding training datasets by combining off-the-shelf image and video datasets. However, these multimodal LLM-based models typically employ simple frame-based architectures and rarely account for local entities and spatiotemporal dynamics such as hand poses, manipulated objects, and state or structural changes. HanDyVQA provides a new challenge for developing advanced visual encoders through demanding HOI recognition tasks.

Table 1 shows a comparison against previous datasets. HanDyVQA focuses on the components, processes, and effects of HOIs lasting several seconds, as opposed to instantaneous events (EgoThink, HOI-QA, EgoHOIBench) or long-form events (AMB), and it is the only dataset covering these aspects within the context of HOIs.

3 HanDyVQA Benchmark

Our goal is to create a systematic benchmark that evaluates the ability to recognize the spatiotemporal dynamics, processes, and effects present in HOIs. To this end, we define two tasks in our benchmark: (1) Multiple-Choice Question (MCQ) and (2) Referring Video Object Segmentation (RVOS). Given a video and a question, the goal of the MCQ task is to select the correct answer(s) from a set of options, while the RVOS task further requires predicting the segmentation masks corresponding to the correct objects or parts. We define six question categories: Action, Process, Location, Objects, State, and Parts. MCQ samples are provided for all question types, whereas RVOS samples are provided only for Objects and Parts questions.

We opted to adopt the MCQ format over open-ended answers because multiple valid responses can exist for certain types of questions, and MCQ enables quantitative evaluation of fine-grained differences in HOIs by presenting plausible, yet incorrect, alternatives. In this section, we describe details of our data collection process (Section 3.1) and its analysis (Section 3.2).



Q. How does the person dip his right hand inside the plate of flour?
(A) The person plunges his entire right hand deep into the flour.
(B) The person waves his right hand back and forth over the plate of flour.
(C) The person slightly dips the fingertips of his right hand in the flour.
(D) The person merely brushes the palm of his right hand against the flour.
(E) The person pinches a small amount of flour with his right hand and holds it above the plate.

Q. What part of the scooter is put down?
(A) The handlebar of the scooter is being put down.
(B) The front wheel of the scooter is being put down.
(C) The grip pad of the scooter is being put down.
(D) The brake lever of the scooter is being put down.
(E) The seat of the scooter is being put down.

Figure 2: Scenario distribution (left) and example QA pairs of HanDyVQA (right). Sentence with green highlights and green region in images denote correct answer and ground truth masks, respectively.

	Action	**Process**	**Location**	**State**	**Parts**	**Objects**
#Q	1978	1924	1974	1940	1913	1939
#Opt	5	5	5	5	5	5.7
#Ans	1	1	1	1	1	1.6
#Words	18.1	20.2	12.3	13.0	8.9	1.4

(a) Statistics of QA task. Q: Question, Opt: Options, Ans: Correct answers, Words: Words per option.

	#Frames	**Avg. Frames per Video**	**Avg. Centroid Shift** (px)	**Avg. IoU w/ Adjacent Frames**
Objects	5546	3.36	88.28	0.08
Parts	5492	2.89	94.13	0.17

(b) Statistics of segmentation task.

Figure 3: Overview of dataset statistics: (a) Question types and (b) segmentation annotations.

3.1 QA collection

We developed a collaborative framework that uses LLMs to propose initial QA candidates, that are carefully refined and verified by humans to ensure quality and diversity.

Data curation We build our benchmark on Ego4D [13] as it includes unscripted and realistic hand-object interactions that covers a variety of scenarios and recording locations. To find short video clips capturing moments of HOIs, we utilize the narrations and timestamp information provided in their annotation. Narrations concisely describe the actions performed by the camera wearer, allowing us to automatically determine whether an HOI event is occurring or not within the videos. We feed each narration into LLMs to infer the object in contact and second objects (objects contacted by an in-use tool [5]) for each hand. If we confirm that the camera wearer is manipulating at least one object, we retain the corresponding clip for further use. After curation, we sample 2,000 narrations per category that contain relevant verbs (primary action conducted in the clip) or information to generate questions for each question type. For each narration, we use a 5-second video segment centered around its timestamp, spanning 2.5 seconds before and after the narration. See supplementary for details.

Question candidate generation QA pair candidates are automatically generated from narrations using the following templates. **Action:** “What is the person doing with his/her hands?” **Process:** “How does the person `[verb] [object]`?” **Location:** “Where does the person `[verb]`” **Objects:** “What object is used by the hands?” **State Change:** “How did the state of `[object]` change?” **Object Parts:** “What part of `[object]` is `[effect]`?”

Verbs and objects are extracted from the narration and inserted into the corresponding placeholders, `[verb]` and `[object]`. For **Object Parts** questions, we ask LLMs to infer the plausible objects and effects to be inserted.

QA refinement by humans Given the generated questions, annotators verify their validity and revise or reject any that do not match the actual content. Then, they provide a correct answer for each question—listing all plausible objects in the **Objects** category where multiple answers may exist—while ensuring that each answer contains enough detail to be understood without watching the video. Next, wrong answer choice candidates are generated by LLMs based on the question and its correct answer. Annotators then refine these choices by removing overlaps, improving plausibility, and adding more challenging distractors when necessary. Overall, annotators ensure that all questions, answers, and choices are accurate, sufficiently confusing, and solvable by humans. Examples of challenging questions with confusing choices are shown in Figure 2 right and Figure 5.

Mask annotation by humans For the **Objects** and **Object Parts** questions, annotators sampled around three representative frames where the target regions were clearly visible from the video, and annotated the regions corresponding to the correct answer.

3.2 Dataset statistics

As a result, 11,668 QA pairs in total are curated for fine-tuning and evaluation. Figure 3 (a) shows the statistics for each question type. **Action** and **Process** exhibit longer descriptions than other categories to explain the nuance of the conducted HOIs. Because often more then one objects are being handled within a 5-seconds clip, an average of 1.6 objects are annotated in **Objects**.

Diversity in HOI scenarios As shown in Figure 2, our dataset covers a wide range of video scenarios, including cooking, gardening, cars, and more. We observe a relatively uniform frequency of verbs in the narration annotations, requiring the models to understand various actions and their underlying interactions.

Distribution of mask annotation Table 3 (b) shows the number of annotated frames, and the relative movement/spatial overlap between them. Due to the nature of object manipulation and moving cameras in egocentric videos, the segmentation masks shift dynamically over time and space, making them challenging to predict. See supplementary for further analysis.

Splits We divided the videos into training, validation, and test sets in a 10 : 5 : 85 ratio, yielding 1.2 K, 0.6 K, and 9.9 K questions, respectively. Only a small portion was set aside as a training/validation set for instruction tuning, allowing the models to learn the required output format while placing greater emphasis on evaluation rather than model training.

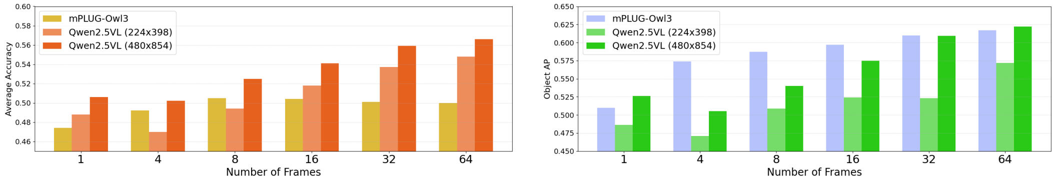
4 Experiments

Models	**Visual Backbone**	**Resolution**	**LLM**	**Action** (Acc)	**Process** (Acc)	**Location** (Acc)	**State** (Acc)	**Parts** (Acc)	**Avg.** (Acc)	**Objects** (AP)
Random	–	–	–	19.3	18.8	20.5	20.0	19.4	19.6	28.5
Text only models										
GPT-4o (text)[*1]	–	–	GPT-4o	36.6	50.3	33.6	39.3	44.7	40.9	34.4
Open-source dual-encoder video-language models										
LaViLa (TSF-L)	TimeSformer	224x224	–	61.2	40.0	35.8	38.5	35.6	42.2	67.0
InternVideo2-Stage2	Original	224x224	–	40.8	30.3	29.2	34.6	30.7	33.1	36.8
Open source video-language models w/ integrated LLMs										
VideoLLaMA2.1-7B	SigLip	384x384	Qwen2	41.1	47.1	34.4	46.3	40.0	41.8	52.1
LLaVa-Video-7B	SigLip	384x384	LLaVa-7B	56.4	53.6	49.1	57.9	53.7	54.1	58.9
mPLUG-Owl3-8B	SigLip	384x384	Qwen2	56.2	51.7	44.9	54.5	47.8	51.0	59.7
Qwen2.5-VL-7B	Original	384x384	Qwen2.5	60.2	55.0	46.9	55.5	47.4	53.0	53.0
Qwen2.5-VL-72B	Original	480x854	Qwen2.5	77.3	73.0	61.4	71.1	61.2	68.8	73.5
Proprietary vision and language models										
GPT-4o[*1] (vision)	Original	480x854	GPT-4o	60.7	64.1	50.5	58.4	57.3	58.2	62.9

Table 2: Comparison of different models across various question types. *1 GPT-4o text/vision refused to answer some questions, providing valid answers to around 87% and 79% of total questions. We report the numbers from valid responses.

Models	**#Frames**	**Key Features**	**Action**	**Process**	**Location**	**State**	**Parts**	**Avg.**	**Objects**
LaViLa-L	4	–	59.2	39.5	35.3	38.3	34.8	41.4	66.1
HelpingHands-L	4	Hand & Object BBox Inference	56.6 (-2.6)	36.6 (-2.9)	34.2 (-1.1)	39.1 (+0.8)	34.4 (-0.4)	40.2 (-1.2)	67.9 (+1.8)
LaViLa-L	16	–	61.2	40.0	35.8	38.5	35.6	42.2	67.0
EgoHOD-L	16	Rich Text & Motion Adapter	59.9 (-1.3)	37.3 (-2.7)	37.5 (+1.7)	41.8 (+3.3)	35.4 (-0.2)	42.4 (+0.2)	74.0 (+7.0)

Table 3: Comparison of models w/ explicit hand and object modeling



(a) Average accuracy except **Objects**.

(b) Average precision for **Objects**.

Figure 4: Ablations on number of input frames from 1 (≈0.2 fps) to 64 (≈12.8 fps) and resolution.

To reveal the challenges in recognizing the dynamic aspects of HOI in HanDyVQA, we compare the performance of major existing video-language models.

4.1 Multi-Choice Questions

We compare the zero-shot performance of representative off-the-shelf video-language models. In addition, we evaluate fine-tuning models with additional modalities of hand and object locations to find directions for future model development.

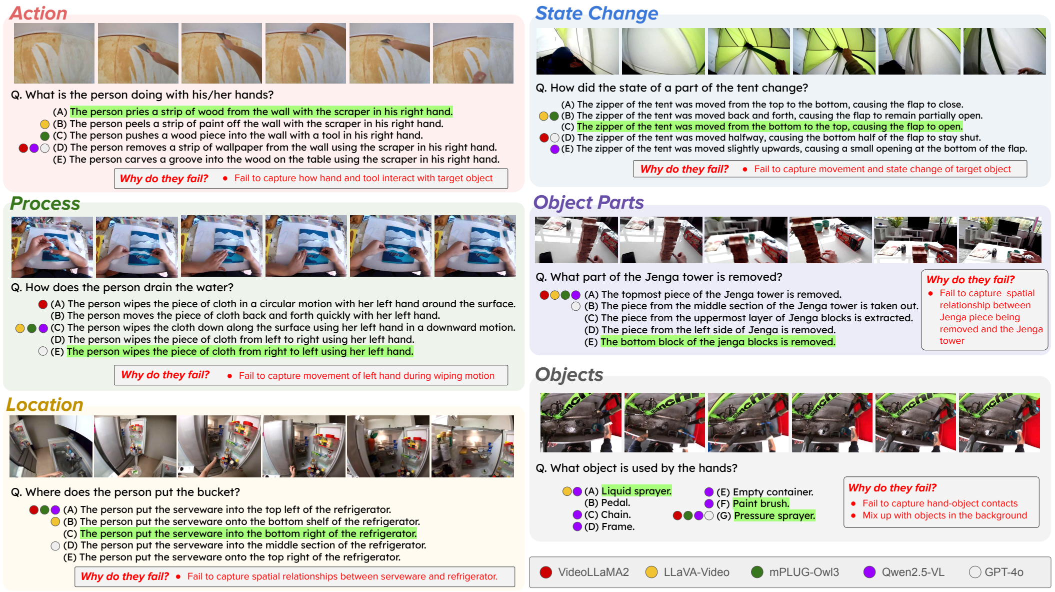
Baseline models We select six open-source video LLMs and one proprietary model, categorized into two types based on their architecture: Dual-Encoder models and LLM-integrated models. **Dual-Encoder models** include LaViLa [58], a video-language model trained on egocentric videos, and InternVideo2-Stage2 [43], whose visual encoder is pre-trained on large-scale video-text pairs. **LLM-integrated models** include VideoLLaMA2.1-7B [6], which specializes in spatio-temporal modeling; LLaVa-Video-7B [19], trained on general and egocentric video datasets; mPLUG-Owl3-8B [51], which efficiently processes long image sequences; and Qwen2.5-VL-7/72B [41], which accepts video inputs with arbitrary resolutions. We also evaluate GPT-4o [15], a proprietary model capable of processing image sequences, in both text-only and vision-enabled settings.

Implementation details We uniformly sample 16 frames from each video and use the default input resolution specified for each model. All models are evaluated in a zero-shot setting. Since Qwen2.5-VL supports arbitrary input resolutions, we aligned its input resolution with that of other 7B-scale LLM-integrated models for a fair comparison. For the 72B model, however, we use the full video resolution to showcase its full capability. For dual-encoder models such as LaViLa and InternVideo2, we compute the cosine similarity between the video feature and the text feature of each option, selecting the one(s) with the highest score. For the remaining video LLMs, we provide the video frames along with a prompt listing all options and infer the most probable option(s).

Evaluation metrics We report top-1 accuracy for all the categories except **Objects**. We report Average Precision (AP) for **Objects** because it has more than one answers per question.

Quantitative results Table 2 shows the quantitative results. Despite preparing answer options unsolvable from text solely, GPT-4o (text) showed moderate results (33–50 pts) than random chance, suggesting some textual bias exists but not enough to solve the task. The dual encoder-based LaViLa trained on the Ego4D dataset outperformed InternVideo2-Stage2, particularly in the **Action** and **Objects** categories, surpassing some LLM-integrated models without text decoder. However, its performance was weaker in other categories, suggesting that LaViLa is specialized to recognize actions and objects. Models with LLM decoders outperformed the text-only baseline, following similar trends observed in general video understanding tasks [33]. Among the 7B-scale models, LLaVA-Video-7B, which is fine-tuned on Ego4D, achieved the highest average accuracy of 54.1%, highlighting the benefits of domain-specific adaptation. Qwen2.5-VL-72B using high resolution images achieved the best overall performance with 68.8% average accuracy, even surpassing GPT-4o (vision) under the same input resolution. However, all models showed limited performance, with top-1 accuracies at most 61% to 77% across categories, suggesting that current large-scale video foundation models still struggle to capture the fine-grained nuances of hand-object interactions.

Qualitative results Figure 5 shows examples that most of the models struggled in each category. Major failure cases include (i) missing objects or hand movement mentioned in the question, (ii)



Q. What is the person doing with his/her hands?

(A) The person pries a strip of wood from the wall with the scraper in his right hand.
🟡 (B) The person peels a strip of paint off the wall with the scraper in his right hand.
🟢 (C) The person pushes a wood piece into the wall with a tool in his right hand.
🔴🟣⚪ (D) The person removes a strip of wallpaper from the wall using the scraper in his right hand.
(E) The person carves a groove into the wood on the table using the scraper in his right hand.

Why do they fail?
- Fail to capture how hand and tool interact with target object

Q. How does the person drain the water?

🔴 (A) The person wipes the piece of cloth in a circular motion with her left hand around the surface.
(B) The person moves the piece of cloth back and forth quickly with her left hand.
🟡🟢🟣 (C) The person wipes the cloth down along the surface using her left hand in a downward motion.
(D) The person wipes the piece of cloth from left to right using her left hand.
⚪ (E) The person wipes the piece of cloth from right to left using her left hand.

Why do they fail?
- Fail to capture movement of left hand during wiping motion

Q. Where does the person put the bucket?

🔴🟢🟣 (A) The person put the serveware into the top left of the refrigerator.
🟡 (B) The person put the serveware onto the bottom shelf of the refrigerator.
(C) The person put the serveware into the bottom right of the refrigerator.
⚪ (D) The person put the serveware into the middle section of the refrigerator.
(E) The person put the serveware onto the top right of the refrigerator.

Why do they fail?
- Fail to capture spatial relationships between serveware and refrigerator.

Q. How did the state of a part of the tent change?

(A) The zipper of the tent was moved from the top to the bottom, causing the flap to close.
🟡🟢 (B) The zipper of the tent was moved back and forth, causing the flap to remain partially open.
(C) The zipper of the tent was moved from the bottom to the top, causing the flap to open.
🔴⚪ (D) The zipper of the tent was moved halfway, causing the bottom half of the flap to stay shut.
🟣 (E) The zipper of the tent was moved slightly upwards, causing a small opening at the bottom of the flap.

Why do they fail?
- Fail to capture movement and state change of target object

Q. What part of the Jenga tower is removed?

🔴🟡🟢🟣 (A) The topmost piece of the Jenga tower is removed.
⚪ (B) The piece from the middle section of the Jenga tower is taken out.
(C) The piece from the uppermost layer of Jenga blocks is extracted.
(D) The piece from the left side of Jenga is removed.
(E) The bottom block of the jenga blocks is removed.

Why do they fail?
- Fail to capture spatial relationship between Jenga piece being removed and the Jenga tower

Q. What object is used by the hands?

🟡🟣 (A) Liquid sprayer.
(B) Pedal.
🟣 (C) Chain.
🟣 (D) Frame.
🟣 (E) Empty container.
🟣 (F) Paint brush.
🔴🟢🟣⚪ (G) Pressure sprayer.

Why do they fail?
- Fail to capture hand-object contacts
- Mix up with objects in the background

🔴 VideoLLaMA2 🟡 LLaVA-Video 🟢 mPLUG-Owl3 🟣 Qwen2.5-VL ⚪ GPT-4o

Figure 5: Qualitative results. Sentence with green highlights denote correct answer.

failing to capture the spatiotemporal dynamics between several objects or parts, and (iii) confusing objects spatially close to each other. In summary, the proposed challenging MCQ reveal that large video models exhibit some shortcuts ignoring the fine-grained aspects of HOIs.

Ablations on number of frames and resolution We measured the effect on number of input frames and spatial resolution on mPLUG-Owl3 and Qwen2.5-VL. We changed the number of input frames for mPLUG-Owl3 while keeping the resolution at 384×384, and varied both the number of input frames and resolutions for Qwen2.5-VL.

As shown in Figure 4, increasing both the number of input frames and resolution was effective. However, for mPLUG-Owl3-8B, performance plateaued beyond 8 frames in all categories except **Objects**, possibly because it has been primarilly trained by 8 frames per video clip. Qwen2.5-VL-7B consistently benefited from incrasing the number of frames across all categories. Increasing input resolution from 224×398 to 480×854 led to a consistent improvement in average accuracy of 3.2%–6.8% across all frame settings. The largest gains were seen in the Objects category (6.1%–16.5%), followed by **Action**, **Location**, and **Parts**. The impact was relatively smaller for **Process** and **State** (see Supplemental for full results). These findings suggest that our benchmark requires both spatially and temporally fine-grained details to answer the questions to find the exact moments and locations of the HOI events, compared to the typical settings (224×224, 16 frames per clip).

Evaluation on hand/object-aware models Besides the generic video baselines, we also tested HelpingHands [56] and EgoHOD [26], two models expressly designed to capture hand/object-aware features. The former extends the LaViLa visual encoder, while the latter builds on CLIP of a similar size. Both incorporate auxiliary supervision from hand and object bounding box locations—an approach likely well suited to our dataset. The results are shown in Table 3. Both models boosted performance in the **State** and **Objects**, and EgoHOD additionally improving in **Location**. EgoHOD—designed to generate textual descriptions of hand–object motions—outperformed HelpingHands using hand/object location supervision. However, both models reduced accuracy in the remaining categories, suggesting that hand-object location supervision helps with queries about object types and positions but does little for more dynamic aspects.

4.2 Referring Video Object Segmentation

Baseline models We compare three baselines: **Sa2VA** [53] **(Frame-wise)**: A multimodal large language model (MLLM) capable of solving both referring image/video segmentation. We input each frame into the Sa2VA model along with the question as a prompt. **Sa2VA (Video)**: We input the

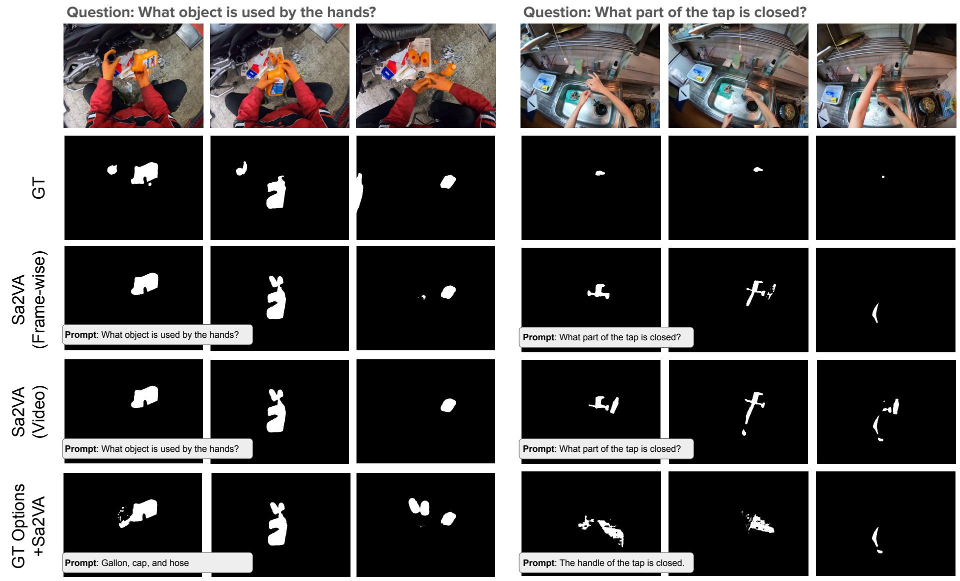

Question: What object is used by the hands?
Question: What part of the tap is closed?
GT
Sa2VA (Frame-wise)
Sa2VA (Video)
GT Options +Sa2VA
Prompt: What object is used by the hands?
Prompt: What part of the tap is closed?
Prompt: What object is used by the hands?
Prompt: What part of the tap is closed?
Prompt: Gallon, cap, and hose
Prompt: The handle of the tap is closed.

Figure 6: Qualitative results on RVOS. Each black and white image in the bottom shows the ground truth/predicted masks, where white region denotes active regions. The text below each mask image sequence is the textual prompt given to each model.

entire sequence of video frames at once along with the question as a prompt. **Ground Truth Option + Sa2VA**: Use the frame-wise Sa2VA but input the correct answer text instead of questions.

Evaluation Metrics Following standard VOS evaluation protocols [28, 48], we use the Jaccard Index ($\mathcal{J}$) and Boundary F-measure ($\mathcal{F}$) computed for each frame and report their average over annotated frames. Furthermore, we categorize the videos into three size-based groups (S/M/L) based on the area of the ground truth masks averaged over each video, and report their average.

Implementation details We input 16 frames per video for all the models, ensuring to include the annotated frames while maintaining the frames to be uniformly sampled from the entire video.

Results As shown in Table 4, all the models performed significantly worse than those in prior VOS tasks (*e.g.,* 70+ $\mathcal{J}$ in [8]), especially for **Parts** which requires referring to specific parts of objects. We observed different trends in each size group. While giving the GT option led to better scores for larger ground truth masks, single-stage models (Sa2VA frame-wise and video) were better against smaller masks (groups S and M). This is possibly because the ground truth text used in the two-stage approach may be not sufficient to describe the precise region in fine-grained HOIs, often leading to over-segmentation of the target area (*e.g.,* the gallon in Figure 6, left). Sa2VA (frame-wise) achieved slightly higher $\mathcal{J}$ than Sa2VA (video), suggesting that video models struggle to track regions in rapidly moving egocentric videos (*e.g.,* the tap handle in Figure 6, right).

Models	Objects ($\mathcal{J}$)				Objects ($\mathcal{F}$)				Parts ($\mathcal{J}$)				Parts ($\mathcal{F}$)			
	S	M	L	All	S	M	L	All	S	M	L	All	S	M	L	All
Sa2VA (Frame-wise)	0.215	**0.425**	0.439	**0.359**	0.226	0.349	**0.306**	0.294	**0.019**	**0.089**	0.270	**0.126**	0.036	0.097	0.165	0.099
Sa2VA (Video)	**0.223**	0.380	0.355	0.319	**0.277**	**0.360**	0.297	**0.312**	0.017	0.080	0.230	0.109	**0.040**	**0.101**	0.160	**0.101**
GT option + Sa2VA	0.076	0.239	**0.464**	0.259	0.088	0.185	0.268	0.182	0.011	0.060	**0.284**	0.119	0.024	0.070	**0.172**	0.089

Table 4: Results of RVOS for **Objects** and **Parts** categories. S/M/L denotes groups of videos categorized by the average size of annotated masks within each video.

4.3 Integration of HOI cues

We also investigate additional factors to better capture the hand-object dynamics in our dataset. We hypothesize that explicitly feeding spatio-temporally local information about hand manipulations and interacting objects can improve performance compared to relying solely on frame-level features.

Input	**Fine-tune**	**Action**	**Process**	**Location**	**State**	**Parts**	**Avg.**	**Objects**
Zero-shot (RGB)	No	40.8	30.3	29.2	34.6	30.7	33.1	36.8
RGB	Yes	50.0	63.6	43.4	44.8	44.7	49.3	37.1
RGB + BBox	Yes	48.1 (-3.8%)	68.2 (+7.2%)	47.1 (+8.5%)	47.9 (+6.9%)	47.5 (+6.3%)	51.8 (+5.1%)	37.5 (+1.1%)
RGB + Hand	Yes	49.6 (-0.8%)	67.1 (+5.5%)	47.2 (+8.8%)	49.0 (+9.4%)	47.0 (+5.2%)	52.0 (+5.5%)	37.0 (-0.3%)
RGB + Object Feats	Yes	44.4 (-11.2%)	68.2 (+7.2%)	47.1 (+8.5%)	48.5 (+8.3%)	45.4 (+1.6%)	50.7 (+2.8%)	38.1 (+2.7%)
RGB + Hand + BBox	Yes	50.2 (+0.4%)	69.1 (+8.6%)	45.8 (+5.5%)	48.4 (+8.0%)	48.2 (+7.8%)	52.3 (+6.1%)	37.0 (-0.3%)
RGB + Hand + Object Feats	Yes	42.3 (-15.4%)	69.5 (+9.3%)	47.0 (+8.3%)	46.6 (+4.0%)	45.8 (+2.5%)	50.2 (+1.8%)	38.2 (+3.0%)
RGB + Hand + BBox + Object Feats	Yes	51.5 (+3.0%)	68.1 (+7.1%)	46.2 (+6.5%)	49.4 (+10.3%)	47.8 (+6.9%)	52.6 (+6.7%)	37.8 (+1.9%)

Table 5: Comparison between different input information. Percentages in Red/Green color indicate performance drop/gain relative to the RGB input.

To test this hypothesis, we chose InternVideo2-Stage2 [43] as our baseline model and fine-tune the model using the training split (1.2 K questions) which could be regarded as a small-scale instruction tuning set. We trained different models by appending addtional branches that input combination of additional cues. Specifically, we considered (i) 3D hand pose information, (ii) bounding box tracklets of manipulated objects, and (iii) their object features.

Implementation details In addition to the Internvideo2 visual/text encoder, separate encoders for each modality consisting of frame-wise MLP and LSTM are introduced. These cues are concatenated with the video feature and passed to a projection layer to match the embedding space. The visual encoder and text encoder of InternVideo2 remain frozen during training, and only the added layers are trained. We input 16 frames per video with a resolution of 224×224. 63-dimensional 3D hand poses are extracted using WiLoR [29]. 4-dimensinal bounding box tracklets of manipulated objects are obtained using AMEGO [12]. 768-dimensional object features are extracted using CLIP [31].

Results Table 5 shows the comparison across different modalities. First, we observe significant improvement by applying fine-tuning (49.3 vs. 33.1 avg. accuracy), especially in the **Process** category that requires answering the detailed process on HOIs. In contrast to the results of EgoHOD and HelpingHands, additional cues boosted performance also in **Process** and **Parts**, suggesting that the standard ViT encoder is suboptimal solving fine-grained tasks in the challenging HanDyVQAdataset.

5 Discussion

What is missing towards understanding HOI dynamics? Experimental results show that state-of-the-art models still struggle to capture fine-grained hand–object interactions across categories. Ablation studies and qualitative analyses indicate that these models often miss the precise locations and motions of local components and the interactions between hands and objects, details that are essential for distinguishing key events. Most recent video MLLMs rely on frame-level Vision Transformers that remain frozen during video-text training. However, the improvements observed when fine-tuning the visual encoder with additional modalities suggest considerable room for progress. Modeling below the frame level—by tracking the spatio-temporal evolution of hand movements and object transformations at higher frame rates (*e.g.,* [32]) may further enhance HOI understanding.

Limitations Although we use a multiple-choice format for quantitative evaluation, crafting convincing distractors becomes increasingly difficult as questions grow more specific and detailed. This raises the risk that models exploit subtle textual biases or general commonsense instead of genuine comprehension. A hybrid evaluation that also incorporates free-form answers may be necessary. Moreover, extending the task to predict geometric properties of HOIs (*e.g.,* the positions and shapes of hands, objects and their components) appears to be a promising next step.

6 Conclusion

We have proposed HanDyVQA, a new video QA benchmark for evalutating abundant spatiotemporal dynamics in HOIs. Experimental results reveal that strong video-language models struggles in capturing the fine-grained details of HOIs, only achieving at most 61–77% top-1 accuracy in MCQ, and showing poor performance in referring local regions. Ablation studies and modality analysis suggested the need of improvements to model the local spatiotemporal dynamics b/w local components. We hope that HanDyVQA opens up new directions towards future development.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing HOT3D: An egocentric dataset for 3D hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024.

[2] Siddhant Bansal, Michael Wray, and Dima Damen. Hoi-ref: Hand-object interaction referral in egocentric vision. *arXiv preprint arXiv:2404.09933*, 2024.

[3] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021.

[4] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024.

[5] Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. *Advances in Neural Information Processing Systems*, 36:30453–30465, 2023.

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

[8] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022.

[9] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023.

[10] Alessandro Flaborea, Guido Maria D'Amely Di Melendugno, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, and Fabio Galasso. PREGO: online mistake detection in procedural egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18483–18492, 2024.

[11] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[12] Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. AMEGO: Active memory from long egocentric videos. In *European Conference on Computer Vision*, pages 92–110. Springer, 2024.

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

[14] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3D annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020.

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[16] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022.

[17] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021.

[18] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[19] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[20] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[21] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

[22] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[23] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889, 2020.

[24] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[25] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3D hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023.

[26] Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, et al. Modeling fine-grained hand-object dynamics for egocentric video representation learning. *Proceedings of the International Conference on Learning Representations*, 2025.

[27] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, To Appear*, 2025.

[28] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[29] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, To Appear*, 2025.

[30] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[32] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18622–18632, 2024.

[33] Mohammadreza Salehi, Jae Sung Park, Tanush Yadav, Aditya Kusupati, Ranjay Krishna, Yejin Choi, Hannaneh Hajishirzi, and Ali Farhadi. Actionatlas: A videoqa benchmark for domain-specialized action recognition. *arXiv preprint arXiv:2410.05774*, 2024.

[34] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[35] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[36] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pages 654–665. PMLR, 2023.

[37] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4D Goal-Step: Toward hierarchical understanding of procedural activities. *Advances in Neural Information Processing Systems*, 36:38863–38886, 2023.

[38] Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. Showme: Benchmarking object-agnostic hand-object 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1935–1944, 2023.

[39] Sirnam Swetha, Hilde Kuehne, and Mubarak Shah. Timelogic: A temporal logic benchmark for video qa. *arXiv preprint arXiv:2501.07214*, 2025.

[40] Meng-Fen Tsai, Rosalie H Wang, and José Zariffa. Recognizing hand use and hand role at home after stroke from egocentric video. *PLOS Digital Health*, 2(10):e0000361, 2023.

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[42] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.

[43] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[44] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

[45] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[46] Boshen Xu, Ziheng Wang, Yang Du, Zhinan Song, Sipeng Zheng, and Qin Jin. Do egocentric video-language models really understand hand-object interactions? In *Proceedings of the International Conference on Learning Representations*, 2025.

[47] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[48] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018.

[49] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[50] Takuma Yagi, Misaki Ohashi, Yifei Huang, Ryosuke Furuta, Shungo Adachi, Toutai Mitsuyama, and Yoichi Sato. FineBio: a fine-grained video dataset of biological experiments with hierarchical annotation. *arXiv preprint arXiv:2402.00293*, 2024.

[51] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

[52] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[53] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images and videos. *arXiv*, 2025.

[54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[55] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024.

[56] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware ego-centric video recognition model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13901–13912, 2023.

[57] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[58] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023.

[59] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We create a new benchmark to evaluate detailed hand-object interaction recognition. The novelty (Section 1) and the characteristics of the dataset (Section 3) are properly detailed. In addition, the evaluation results on current video-language models and detailed analysis are provided in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our dataset in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide model details, evaluation metrics, and implementation procedures for all experiments (Section 4). The experiments are reproducible by referring to the existing documentation for the models and following our described usage.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide a Hugging Face link to the dataset. In addition, we specify all text prompts and implementation details used with the LLMs/VLMs in our experiments in supplementary material, enabling others to reproduce the results using publicly available models.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details regarding the use of our newly proposed dataset are provided at the beginning of each experiment (Sections 4.1, 4.2, and 4.3). Fine-grained information, such as text prompts and training settings, is included in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars due to the high computational cost of each experiment, which made it challenging to run them multiple times.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For each experiment we report the type of GPU and the approximate number of GPU hours used in supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres to the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss positive societal impacts in the introduction, and further discussions including their risks in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our dataset is built upon the existing Ego4D dataset, which has already established strict safeguards and responsible data usage policies. As our dataset only provides additional annotations on top of Ego4D, we rely on Ego4D's safeguards for responsible release.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Ego4D is properly cited in our paper, and we comply with its license. Our annotations are released under CC-BY-SA 4.0.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Details of our dataset is documented in Section 3 and the supplementary material, and is released with code for data processing.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Annotations were outsourced to a company. Details of the instructions, interface, and compensation are provided in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve research that requires IRB approvals or equivalent.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We used LLMs to assist human annotators. The usage is detailed in Section 3 and the supplementary material.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.