

## Supplementary Material Contents

A. Details on Experimental Results	1
B. Details on Data Collection	5
C. Details on Experimental Settings	15
D. Fine-tuning Qwen2.5-VL-7B	19
E. Broader Impacts	19

## A Details on Experimental Results

### A.1 Multi-Choice Questions

**Performance comparison by GPT-4o answerability** In multiple-choice questions, the GPT-4o (text/vision) models often declined to answer seemingly when the model evaluated there are insufficient information to respond to the question, even though we asked to answer the question. Reporting results only for the questions that GPT-4o answered risks an unfair comparison, as those questions might also be easy for the other models. To this end, we re-evaluated every model on the subset of questions answered by both the text and vision variants of GPT-4o, reducing the valid items from 9,942 to 6,970. Table 6 reports these subsampled scores together with their deltas from the original evaluation. Almost all metrics rose, implying that GPT-4o tends to skip the harder questions that rival models also frequently miss.

We divided the questions into four categories according to whether the GPT-4o text and vision variants produced an answer, then evaluated Qwen2.5-VL-72B on each category (Table 7). Accuracy peaked for questions answered by both GPT-4o variants and declined whenever either variant abstained. This pattern suggests that GPT-4o can recognize when textual or visual information is insufficient and refrain from answering, thereby avoiding errors. They also suggested that models specifically designed for videos perform well than the general-purpose GPT-4o model, showing the difficulty of our benchmark as a video QA benchmark.

Models	Action (Acc)	Process (Acc)	Location (Acc)	State (Acc)	Parts (Acc)	Avg. (Acc)	Objects (AP)
Random	20.3 (+1.0)	18.1 (-0.7)	20.5 (+0.0)	20.4 (+0.4)	19.5 (+0.1)	19.8 (+0.2)	28.6 (+0.1)
<i>Text only models</i>							
GPT-4o (text)	37.4 (+0.8)	53.6 (+3.3)	35.9 (+2.3)	39.9 (+0.6)	47.1 (+2.4)	42.8 (+1.9)	34.5 (+0.1)
<i>Open-source dual-encoder VLMs</i>							
LaViLa (TSF-L)	63.2 (+2.0)	41.4 (+1.4)	40.6 (+4.8)	39.9 (+1.4)	38.2 (+2.6)	44.7 (+2.5)	67.1 (+0.1)
InternVideo2-Stage2	41.7 (+0.9)	30.9 (+0.6)	33.1 (+3.9)	35.7 (+1.1)	31.4 (+0.7)	34.6 (+1.5)	36.9 (+0.1)
<i>Open source VLMs w/ LLMs</i>							
VideoLLaMA2.1-7B	43.7 (+2.6)	49.8 (+2.7)	39.0 (+4.6)	47.3 (+1.0)	44.5 (+4.5)	44.8 (+3.0)	52.4 (+0.3)
LLaVa-Video-7B	60.7 (+4.3)	57.8 (+4.2)	54.1 (+5.0)	59.5 (+1.6)	58.8 (+5.1)	58.2 (+4.1)	59.1 (+0.2)
mPLUG-Owl3-8B	54.6 (-1.6)	56.1 (+4.4)	49.0 (+4.1)	56.7 (+2.2)	52.0 (+4.2)	53.7 (+2.7)	59.8 (+0.1)
Qwen2.5-VL-7B	63.4 (+3.2)	58.2 (+3.2)	52.8 (+5.9)	56.9 (+1.4)	52.2 (+4.8)	56.7 (+3.7)	53.1 (+0.1)
Qwen2.5-VL-72B	79.7 (+2.4)	77.0 (+4.0)	68.0 (+6.6)	73.5 (+2.4)	65.4 (+4.2)	72.7 (+3.9)	73.7 (+0.2)
<i>Proprietary VLMs</i>							
GPT-4o (vision)	59.9 (-0.8)	64.3 (+0.2)	53.0 (+2.5)	58.8 (+0.4)	58.4 (+1.1)	58.9 (+0.7)	62.9 (+0.0)

Table 6: Comparison of different models on subset of questions answered by both the text and vision versions of GPT-4o. Only 6.9K questions (70.2%) are used for this evaluation. Differences from the full results in Table 2 are indicated in +X.X/-X.X.

Answered by GPT-4o?		Results of Qwen2.5-VL-72B (with number of questions)						
Text	Vision	Action	Process	Location	State	Parts	Avg. (Sum)	Objects
Yes	Yes	79.7 (980)	77.0 (1077)	68.0 (800)	73.5 (1375)	65.4 (1104)	72.7 (5336)	73.7 (1634)
Yes	No	67.6 (139)	65.4 (543)	53.9 (386)	60.5 (223)	53.5 (398)	60.2 (1689)	50.5 (16)
No	Yes	77.5 (454)	55.6 (9)	60.6 (310)	54.1 (37)	50.0 (70)	59.6 (880)	– (0)
No	No	67.3 (110)	71.4 (7)	49.7 (183)	50.0 (14)	47.3 (55)	57.1 (369)	– (0)

Table 7: Comparison of performance on questions grouped by whether the GPT-4o text/vision models provided an answer. The number in parentheses indicates the number of questions.

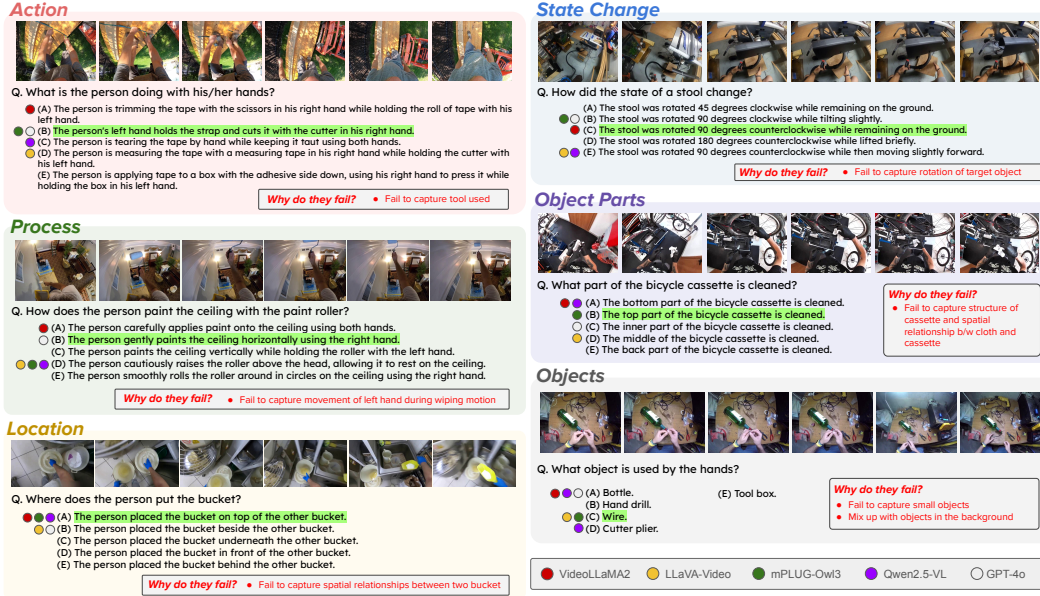


Figure 7: Additional qualitative results for multi-choice questions. Green highlights denote correct answers.



Figure 8: Ablation on mPLUG-Owl3-8B. Color indicates percentage change compared to baseline (frame number = 1).

**More Qualitative Results** More qualitative results per category are shown in Figure 7. We can see that current models struggle to accurately recognize objects being manipulated and their spatial relationship with hands or other objects, and their movement. The original video clips could be found in the supplementary materials.

**Detailed ablation on number of frames and resolution** Figure 8 and 9 show the detailed performance of mPLUG-Owl3-8B and Qwen2.5-VL-7B, respectively, varying the number of input frames and resolution (only in Qwen2.5-VL-7B).

For mPLUG-Owl3-8B, the performance with more than four frames is better than with a single frame, but it does not improve further beyond eight frames in the **Action**, **Process**, **State**, and **Parts** categories. No improvement is observed in the **Location** category. The **Objects** category shows a consistent improvement as the number of frames increases.

For Qwen2.5-VL-7B, raising either the frame count or the resolution usually improves performance across all categories—especially **Action** and **Objects**. The exception is the four-frame setting, which performs worse than the single-frame baseline in most categories, likely because four sampled frames still miss key moments required to solve the task.

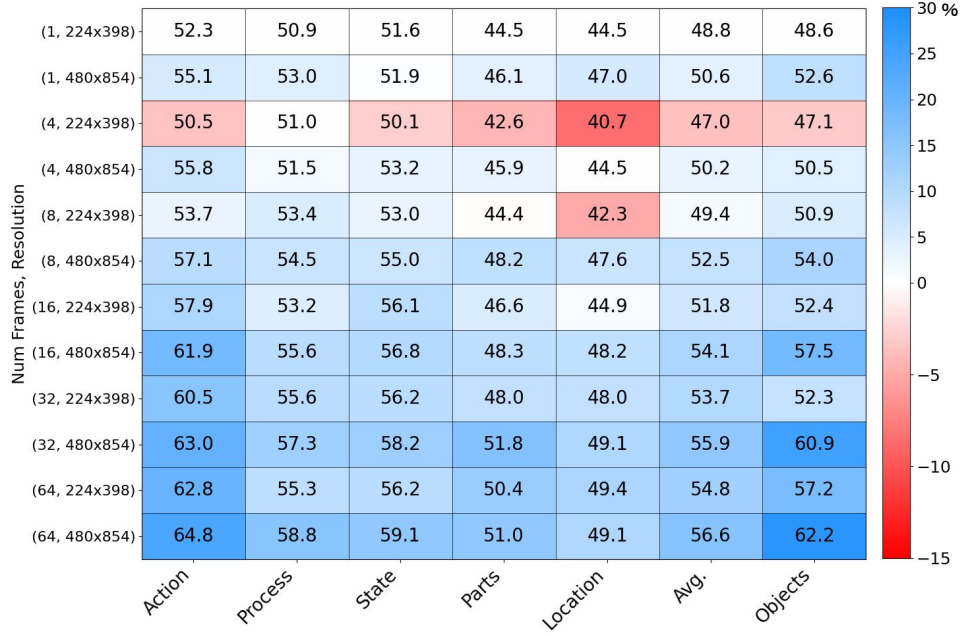


Figure 9: Ablation on Qwen2.5-VL-7B. Color indicates percentage change compared to baseline (frame number = 1, resolution = 224x398).

## A.2 Referring Video Object Segmentation

**More qualitative results** More qualitative results are shown in [10](#). The video baseline successfully exploits the context information to produce consistent output corresponding to the question, while the frame-wise baseline, including GT + Sa2VA, produces masks for similar objects/areas but not the correct ones regarding the question. The original video clips could be found in the supplementary materials.

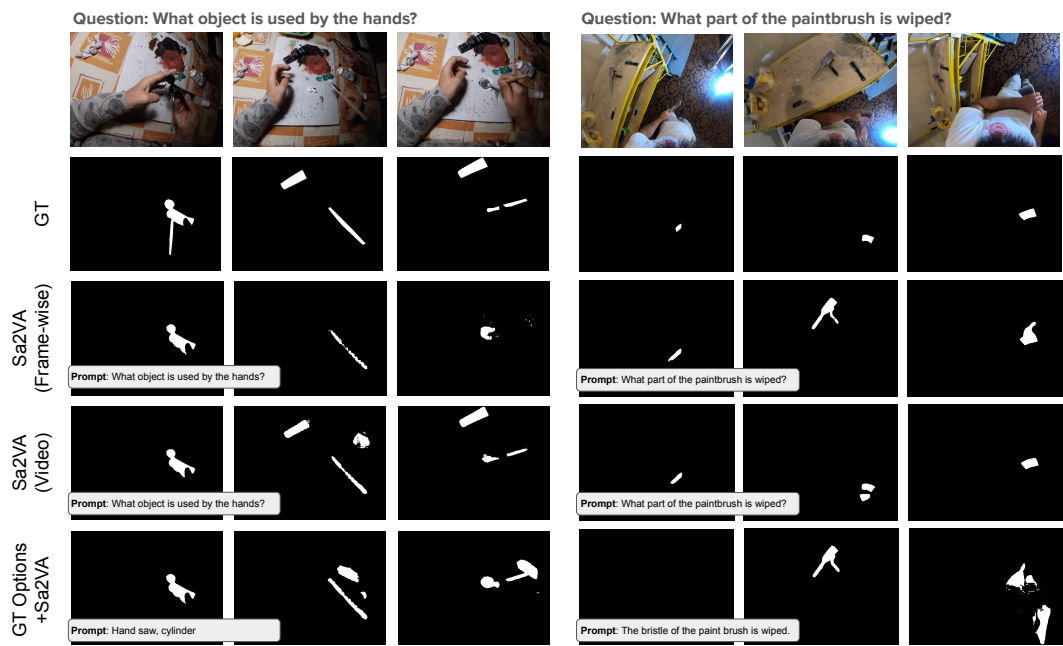


Figure 10: Qualitative results for referring video object segmentation. Text below each mask image sequence is textual information given to each model.

## B Details on Data Collection

**Details on video clip filtering** First, any narration from a stereo video is filtered out. If a narration is duplicated, it is counted as a single entry. A narration is also filtered out if it does not mention either “right hand” or “left hand.” Additionally, if the narration contains the word “unsure”, indicating an unknown object name, it is filtered out. We input each narration into large language models (LLMs) to infer the contact objects and secondary objects for each hand (The prompt is shown in Table 8). If we can confirm that the camera wearer is manipulating at least one object, we retain the narration for use. However, if the extracted contact objects include any of the wearer’s body parts or the camera itself, the narration is discarded, as such cases are likely to obscure the visibility of hand-object interactions.

<p><b>System Prompt:</b> You are a helpful assistant who understands interactions between human hands and objects.</p> <p><b>User Prompt:</b> Please analyze the narration and answer the contact object and the secondary object of each right/left hand. "contact object" means the object that a hand is contacting (manipulating). "secondary object" means the object that the contact object is contacting (the object that is manipulated by the contact object). If the information is not specified in the narration, fill with None. The answer format should be json: { "Contact object of right hand": &lt;obj/None&gt;, "Secondary object of right hand": &lt;obj/None&gt;, "Contact object of left hand": &lt;obj/None&gt;, "Secondary object of left hand": &lt;obj/None&gt; }</p> <p>Narration: {narration text}</p>
--

Table 8: Prompt to extract hand-object interaction information from narration. We replace {narration text} with narration from Ego4D.

**Sampling HOI narrations** To ensure diverse narration samples for each question, we first sort the remaining narrations chronologically. Then, for each question, we offset the starting index and select every sixth narration. This staggered sampling prevents identical questions from being generated for temporally adjacent HOI segments.

Next, we filter out unsuitable narrations specifically for **Location** and **Object Parts**. For **Location** questions, we select videos where the narration contains verbs indicating object movement. For **Object Parts** questions, we determine whether the object is partially affected and retain only those where partial impact is evident. This filtering process is done automatically using LLMs based on the narration (The prompt is shown in Table 9).

Finally, to ensure diversity in HOI samples, we extract verbs from the narrations and randomly select 2,000 samples while ensuring that no single verb appears more than  $N$  times. The  $N$  is determined for each question type to maintain diversity while ensuring at least 2,000 candidate samples are available.

**Human QA annotation** For the automatically generated questions, annotators perform the following tasks while referencing the video: (i) verifying the validity of the question, (ii) creating the correct answer, and (iii) generating wrong answer choices.

If an automatically generated question does not match the video, annotators either revise the question or reject the sample. Next, they create the correct answer, ensuring that it provides sufficient detail for the question to be understandable without watching the video.

Once the question and correct answer are prepared, an initial set of plausible wrong answer choices is generated using LLMs (The prompts are shown in Table 10–Table 15). Annotators then review these choices, filtering out inappropriate ones, such as those that overlap with the correct answer. They refine and add challenging distractors that effectively assess comprehension of the question.

Overall, human annotators verify all questions, correct answers, and wrong answer choices, ensuring that the dataset remains sufficiently challenging while still solvable by humans. The instructions

**System Prompt:**

You are a helpful assistant who understands interactions between human hands and objects.

**User Prompt:**

Create a question for VQA about Hand-Object Interaction. Specifically, the question should follow the format "What part of [Object] is [Verb]?" (e.g., "What part of the bicycle is replaced?").

Given a narration describing a Hand-Object Interaction:

1. First, determine if it's an appropriate scene to create a question.
2. An appropriate scene should include the following conditions:
  - The person is interacting with the object(s).
  - The action only affects a limited area of the object (e.g., replacing the tire of a bicycle or folding the left top corner of the paper) rather than the entire object (e.g., moving the speaker).

If the scene is appropriate for creating the question, create the question accordingly. If it's inappropriate to create a question, reply with None.

If possible, create the corresponding answer to the question. If it's difficult to create the answer, reply with Ambiguous.

Here is the narration:

{narration text}

Write reasoning and then output the json answer like this: <reasoning about whether it is appropriate to create a question>

```

{
  "question": <created question or None>,
  "answer": <created answer or Ambiguous>
}

```

Table 9: Prompt to extract hand-object interaction information from narration. We replace {narration text} with narration from Ego4D.

provided to the annotators are shown in Table 16–Table 21 for each category. The screenshot of the annotation tool interface is provided in Figure 11.

**User Prompt:**

You will be given a pair of a question and an answer about a hand-object interaction. Create 4 sentences that describe similar but incorrect answers, as per the examples.

Examples:

- Question: What is the person doing with his/her hands?
- Correct Answer: The person is cutting an apple on the chopping board using the knife in his right hand while holding the apple with his left hand.
- Incorrect Answer (object name or situation is wrong): The person is slicing an apple on the table using an apple slicer in his right hand while holding the apple with his left hand.
- Incorrect Answer (action is different): The person is removing the skin of an apple on the chopping board using the knife in his right hand while grasping the apple with his left hand.

The sentences should:

- Contain either different object names or situations (as in the first incorrect example)
- Or, contain the same object names but describe different actions (as in the second incorrect example)
- Maintain the same level of detail and be of a similar length to the correct answer
- Avoid overlapping with each other

The answer format should be json:

```
{"1": "<sentence1>", "2": "<sentence2>", "3": "<sentence3>", ..., "4": "<sentence4>"}
```

Make sure that you only answer json.

Question: {question}

Correct Answer: {answer}

Table 10: Prompt to generate options for **Action** question. We replace {question} and {answer} with created question and answer, respectively

**User Prompt:**

You will be given a pair of a question and an answer about a hand-object interaction. Create 4 sentences that describe similar but incorrect answers, as per the examples.

Examples:

- Question: How does the person loosen the fastener?
- Correct Answer: The person loosens the fastener by holding the piece firmly in one hand while using the wrench in the other hand to turn it.
- Incorrect Answer (hand movement or object handling is slightly different but plausible): The person loosens the fastener by shaking the piece in one hand while holding the wrench in the other hand.

The sentences should:

- Contain same objects but different hand movements or ways of doing the action compared to the correct answer
- Be incompatible with the correct answer
- Maintain the same level of detail and be of a similar length to the correct answer
- Avoid overlapping with each other

The answer format should be json:

```
{"1": "<sentence1>", "2": "<sentence2>", "3": "<sentence3>", ..., "4": "<sentence4>"}
```

Make sure that you only answer json.

Question: {question}

Correct Answer: {answer}

Table 11: Prompt to generate options for **Process** question. We replace {question} and {answer} with created question and answer, respectively

**User Prompt:**

You will be given a pair of a question and an answer about a hand-object interaction.  
Create 4 sentences that describe similar but incorrect answers, as per the examples.

Examples:

- Question: Where does the person place the cup?
- Correct Answer: On the left bottom corner of the table.
- Incorrect Answer: On the right top corner of the table.
- Incorrect Answer: On the chair under the table.

The sentences should:

- Include a specific location indicating where the object is moved to, but it should be a different location from the correct answer
- Be close to but different from the correct location (e.g., "On the right top corner of the table.")
- Maintain the same level of detail and be of a similar length to the correct answer
- Avoid overlapping with each other

The answer format should be json:

```
{"1": "<sentence1>", "2": "<sentence2>", "3": "<sentence3>", ..., "4": "<sentence4>"}
```

Make sure that you only answer json.

Question: {question}

Correct Answer: {answer}

Table 12: Prompt to generate options for **Location** question. We replace {question} and {answer} with created question and answer, respectively

**User Prompt:**

You will be given a pair of a question and an answer about a hand-object interaction.  
Create 4 sentences that describe similar but incorrect answers, as per the examples.

Examples:

- Question: How did the state of the screw change?
- Correct Answer: The screw was picked up, put on the hole, and turned clockwise.
- Incorrect Answer (a little bit different state change): The screw that had already been in the hole was turned counter-clockwise.
- Question: How did the state of the camera change?
- Correct Answer: The camera was moved right by the right hand.
- Incorrect Answer (a little bit different state change): The camera was moved left and slightly shifted.

The sentences should:

- Describe very similar but different state changes of the object.
- Be incompatible with the correct answer.
- Not change the object (tools) used from the correct answer.
- Maintain the same level of detail and be of a similar length to the correct answer.
- Avoid overlapping with each other.

The answer format should be json:

```
{"1": "<sentence1>", "2": "<sentence2>", "3": "<sentence3>", ..., "4": "<sentence4>"}
```

Make sure that you only answer json.

Question: {question}

Correct Answer: {answer}

Table 13: Prompt to generate options for **State** question. We replace {question} and {answer} with created question and answer, respectively



**User Prompt:**

You will be given a pair of a question and an answer about a hand-object interaction.

Create 4 sentences that describe similar but incorrect answers, as per the examples.

Examples:

- Question: What part of the hammer is being held?
- Correct Answer: The bottom of the hammer handle.
- Incorrect Answer: Close to the head of the hammer.

The sentences should:

- Contain nonexistent parts or incorrect parts of the object (e.g., Close to the head of the hammer)
- Maintain the same level of detail and be of a similar length to the correct answer
- Avoid overlapping with each other

The answer format should be json:

```
{"1": "<sentence1>", "2": "<sentence2>", "3": "<sentence3>", ..., "4": "<sentence4>"}
```

Make sure that you only answer json.

Question: {question}

Correct Answer: {answer}

Table 14: Prompt to generate options for **Parts** question. We replace {question} and {answer} with created question and answer, respectively

**User Prompt:**

You will be given a pair of a question and an answer about a hand-object interaction.

Create 4 sentences that describe similar but incorrect answers, as per the examples.

Examples:

- Question: What object is used by the hands?
- Correct Answer: knife,apple.
- Incorrect Answer: chopping board.

The sentences should:

- Contain incorrect object names that are likely to be in the same scene as the correct answer (e.g., chopping board)
- Even if the correct answer contains multiple objects, the incorrect answer should contain only one object per option
- Avoid overlapping with each other

The answer format should be json:

```
{"1": "<sentence1>", "2": "<sentence2>", "3": "<sentence3>", ..., "4": "<sentence4>"}
```

Make sure that you only answer json.

Question: {question}

Correct Answer: {answer}

Table 15: Prompt to generate options for **Objects** question. We replace {question} and {answer} with created question and answer, respectively

<b>Purpose</b>	To ask how the person is manipulating objects and to understand the dynamically changing relationships between the hands and the objects.
<b>Required answer information</b>	<ul style="list-style-type: none"> <li>• Details of manipulation by both hands (left and right, if distinguishable)</li> <li>• All objects involved in the action</li> </ul>
<b>Example question</b>	What is the person doing with his/her hands?
<b>Example correct answer</b>	The person is slicing an apple on the chopping board using the knife in his right hand while holding the apple with his left hand.
<b>Examples of incorrect answers</b>	<ul style="list-style-type: none"> <li>• The person is slicing an apple on the table using an apple slicer in his right hand while holding the apple with his left hand.</li> <li>• The person is removing the skin of an apple on the chopping board using the knife in his right hand while grasping the apple with his left hand.</li> </ul>
<b>Notes</b>	Do not omit the subject ("The person").

Table 16: Instruction to annotators to annotate **Action** category

<b>Purpose</b>	To ask about the manner, procedure, technique, or skill involved in a hand action or its interaction with an object.
<b>Required answer information</b>	<ul style="list-style-type: none"> <li>• Which hand is used</li> <li>• How the hand moves or interacts with the object, including the steps and state changes</li> </ul>
<b>Example question</b>	How does the person drop the toy on the table?
<b>Example correct answer</b>	The person drops the action figure gently on the table by holding it with the index finger and thumb in the left hand.
<b>Incorrect example</b>	The person released his grip and let it fall with force.
<b>Notes</b>	Do not omit the subject ("The person") or the verb (action).

Table 17: Instruction to annotators to annotate **Process** category

<b>Purpose</b>	To ask where the manipulated object was moved to, or where it ended up as a result of the action.
<b>Required answer information</b>	A specific description of the location where the object was placed or moved.
<b>Example question</b>	Where does the person place the cup?
<b>Example correct answer</b>	The person places the cup on the left bottom corner of the table.
<b>Incorrect example</b>	The person places the cup on the top right corner of the table.
<b>Notes</b>	Do not omit the subject ("The person"), the verb (action), or the object.

Table 18: Instruction to annotators to annotate **Location** category

<b>Purpose</b>	To ask how the state, structure, composition, or spatial arrangement of the object changed during the video (or remained unchanged).
<b>Required answer information</b>	A description of how the object's state, structure, composition, or placement changed or did not change in the video.
<b>Example question</b>	How did the state of the apple change?
<b>Correct answer</b>	The apple was cut into small slices.
<b>Example question 2</b>	How did the state of the camera change?
<b>Correct answer 2</b>	The camera is divided into two parts: the body and the lens.
<b>Incorrect examples</b>	<ul style="list-style-type: none"> <li>• The apple was crushed.</li> <li>• The apple was sliced. (when slicing did not occur)</li> </ul>
<b>Notes</b>	Do not omit the subject (the object) or the verb.

Table 19: Instruction to annotators to annotate **State** category

<b>Purpose</b>	To identify the specific part of the object that is being affected, considering the object's structure, function, and spatial relation to the hands.
<b>Required answer information</b>	A detailed verbal description of the affected region and its position (including segmentation mask if applicable).
<b>Example question</b>	What part of the hammer is being held?
<b>Correct answer</b>	The bottom of the hammer handle is being held.
<b>Incorrect example</b>	Close to the head of the hammer is being held.
<b>Notes</b>	Do not omit the subject (the part) or the verb (effect).


Table 20: Instruction to annotators to annotate **Parts** category

<b>Purpose</b>	To identify the types and positions of objects being manipulated by the hands.
<b>Required answer information</b>	<ul style="list-style-type: none"> <li>A verbal description of all manipulated objects and their positions (including segmentation mask if applicable)</li> <li>Objects that are merely touched and not clearly manipulated should not be included as correct or incorrect options</li> </ul>
<b>Example question</b>	What object is used by the hands?
<b>Correct answer</b>	knife, apple
<b>Incorrect example</b>	The chopping board (present but not manipulated)
<b>Notes</b>	Only include object names in the answer. Separate multiple correct objects with commas.

Table 21: Instruction to annotators to annotate **Objects** category

how\_0000

Annotated



**Question: How does the person loosen the bolt?**

※ 質問文の修正が必要な場合はボックスに記入  
※ 修正できない問題がある場合は当てはまるものを全てをチェック (Check all that apply)

How does the person loosen the bolt?

☐ 質問として成立していない (The question doesn't make sense)

☐ 動画からは質問に回答出来ない (Unable to answer question from this video)

☐ 正答が一意に定まらない (Unable to create uniquely determined answer)

☐ 既にほぼ同じ動画-質問の組があった (I have already annotated almost the same video-question pairs.)

[Save Question](#)

**Answer: The person loosens the bolt by holding the T-shaped screw tool with his right hand, applying pressure, and rotating it counterclockwise.**

※ 解答は質問に対して内容が適切かつ、解答から当該部分の映像が想像できる程度に詳細である必要があります。  
(The answers must be appropriate to the questions and detailed enough for the reader to visualize the relevant part of the scene.)

例) Q. What is the person doing with his hands?  
A. The person slices an apple using a knife with the right hand while holding the apple with the left hand.

The person loosens the bolt by holding the T-shaped screw tool with his right hand, applying pressure, and rotating it counterclockwise.

[Save New Answer](#)

**Other Incorrect Options:** [Regenerate Options from QA](#)

- ☒ 誤回答として適切 (Appropriate as incorrect answer)
- ☐ 正解と意味が重複 (Overlap with the correct answer)
- ☐ 正解と異なる別解 (Another correct answer)
- ☐ 誤回答として不適 (Not Appropriate as incorrect answer)

The person holds the bolt with his fingers in the left hand, twisting it back and forth instead of using a tool. ☒

Figure 11: Screenshot of the annotation tool interface. Annotators proceed from top to bottom, sequentially annotating question, correct answer, and distractor options.

**Final QA post-processing using LLMs** After human QA annotation, we refined each option, including the correct answer, using LLMs to correct their grammar and ensure a consistent tone across all choices, without changing their meaning. This was especially done for the **Action**, **Process**, and **State** categories, where longer sentences tend to introduce textual biases (e.g., the correct answer being more likely to contain grammatical errors than the distractors). All the prompts used for each question category is shown in Table 22 and Table 23.

<p>You will be given a triplet consisting of a question, a correct answer, and a set of distractors. First, revise the answer if it includes grammatical errors or is not in a natural tone without changing its meaning. If the answer is already correct, please keep it as is. Then, rephrase each distractor to make it more plausible and similar in tone to the correct answer, without significantly changing its original meaning, since the distractors are currently written in a biased way, making them too easy to eliminate. To do this, you may:</p> <ul style="list-style-type: none"> <li>- Use words or phrasing that commonly appear in the correct answer, or avoid words frequently used in distractors.</li> <li>- Tone down any exaggeration to make the distractors sound more natural and believable.</li> <li>- Remove or rephrase strong negations (such as “without” or “instead”) if they clearly oppose the correct answer, unless they are essential to the meaning.</li> </ul> <p>The answer (rephrased distractors) format should be json:</p> <pre>{   "answer": "&lt;revised_answer&gt;",   "options": {"1": "&lt;sentence1&gt;", "2": "&lt;sentence2&gt;", "3": "&lt;sentence&gt;", ..., "4": "&lt;sentence4&gt;" } }</pre> <p>Make sure that you only answer json. Note that all the sentences should start with in the same way as the original sentences (mostly "The person...").</p> <p>Question: {question} Correct Answer: {answer} Distractors: {options}</p>
---

Table 22: Prompt to refine options for **Action** and **Process** question. We replace {question}, {answer}, {options} with created question, answer, and options respectively

**Choice of LLMs** We used gpt-4o-mini-2024-07-18 to generate the initial QA pairs. For final refinement, we used gpt-4o-2024-08-06. We note that generated questions/options are for reference and all the pairs are thoroughly reviewed and corrected to form the final QA pairs.

**Human mask annotation.** Egocentric videos often include severe blurring that harms the visual quality of the video clip. To this end, we opted to annotate three representative frames from each 5-seconds video clip. Annotators manually selected one frame each from the front, middle, and last thirds of the video clip, where the target regions were clearly visible and sampled them from different parts of the video whenever possible. However, the above condition was relaxed when the frames from some of the intervals are unusable.

For **Object Parts** questions, clips often involve object state changes, which change the appearance of the components. In such case, both the frames before and after the state change were selected.

**Detailed dataset statistics** Figure 12 shows the detailed distribution of the scenarios and the primary verbs included in HanDyVQA. While the scenarios reflecting the distribution of the original Ego4D video clips, the verbs are more uniformly selected to ensure diverse HOIs are covered.

Figure 13 shows the spatial and temporal distribution of segmentation mask annotations for **Objects** and **Object Parts** questions. While selected frames are biased towards the beginning and the end of a video the remaining annotations are evenly distributed throughout the video. Regarding the

You will be given a triplet consisting of a question, a correct answer, and a set of distractors. First, revise the answer if it includes grammatical errors or is not in a natural tone without changing its meaning.

Note that since the answer is mainly written in the passive voice to focus on the state of the object, please make sure to keep it in passive form. If the answer is already correct, please keep it as is.

Then, rephrase each distractor to make it more plausible and similar in tone to the correct answer, without significantly changing its original meaning, since the distractors are currently written in a biased way, making them too easy to eliminate. Also, it's better to avoid using adverbs (e.g., "gently") in the distractors, since adverbs typically describe human action rather than the state change of the object.

The answer (rephrased distractors) format should be json:

```
{
  "answer": "<revised_answer>",
  "options": {"1": "<sentence1>", "2": "<sentence2>", "3": "<sentence>", ..., "4": "<sentence4>" }
}
```

Make sure that you only answer json.

Note that all the sentences should start with in the same way as the original sentences (mostly "The person...").

Question: {question}

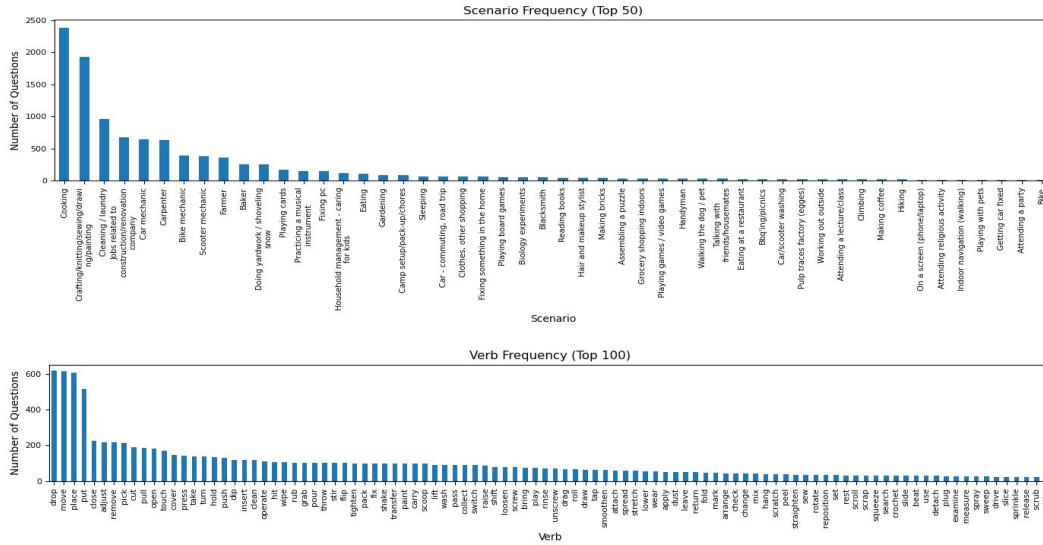
Correct Answer: {answer}

Distractors: {options}

Table 23: Prompt to refine options for **State** question. We replace {question}, {answer}, {options} with created question, answer, and options respectively

spatial direction, the segmentation showed a tendency to concentrate in the center of a frame, but also appeared to spread around the center.

**Details on compasation** We outsourced the annotation of MCQs and segmentation masks to an agency at a cost of 3,180,000 JPY (approximately 22,000 USD). The company is responsible for managing payments to annotators, ensuring compliance with the minimum wage regulations in the annotator's country.



## C Details on experimental settings

### C.1 Multi-Choice Questions

**Frame sampling strategy** We sample  $n$  frames uniformly from a video of length  $L$  by dividing it into  $n$  equal segments and selecting one frame from the center of each segment. Specifically, the sampling index  $i_k$  for the  $k$ -th frame ( $k = 0, 1, \dots, n-1$ ) is computed as:

$$i_k = \left\lfloor k \cdot \text{gap} + \frac{\text{gap}}{2} \right\rfloor, \quad \text{where } \text{gap} = \frac{L}{n}$$

This approach ensures that the sampled frames are evenly distributed over the entire video, while avoiding bias toward the start or end. By choosing the center of each interval, we obtain a more representative snapshot of the temporal progression.

**Prompts for zero-shot evaluation** The textual prompt fed to video-language models with integrated LLMs to solve MCQs is shown in Table 24 and Table 25.

**Computational cost** The 7B-size video-language models fit on a single NVIDIA H200 (141GB) GPU and completed inference for each question category in under an hour. The 72B-size models were able to run on a single node with eight H200 GPUs, requiring approximately 1.7 hours per category.

### C.2 Referring Video Object Segmentation

**Frame sampling strategy** Given a set of annotated frame indices  $\mathcal{A} \subseteq [0, n]$  and a target number of samples  $l$ , we construct a set  $\mathcal{S}$  of  $l$  indices that are both representative and temporally balanced.

- We initialize the set  $\mathcal{S}$  as the sorted, unique subset of annotated indices  $\mathcal{A}$  within their valid range:

$$\mathcal{S} \leftarrow \text{sorted}(\{x \in \mathcal{A} \mid 0 \leq x \leq n\})$$

- While  $|\mathcal{S}| < l$ , we iteratively identify the largest temporal gap between consecutive elements in  $\mathcal{S}$ , including gaps at the start ( $[0, \mathcal{S}_1]$ ) and end ( $[\mathcal{S}_{|\mathcal{S}|}, n]$ ), and insert the midpoint of the largest such gap:

$$\text{midpoint} = \left\lfloor \frac{i+j}{2} \right\rfloor \quad \text{for each gap } [i, j]$$

- This process continues until  $|\mathcal{S}| = l$ , or no more meaningful midpoints can be added.
- If the final size  $|\mathcal{S}| > l$  (e.g., due to duplicate insertions), we resample  $l$  indices from  $\mathcal{S}$  to evenly cover  $[0, n]$ . Specifically, we define  $l$  ideal positions:

$$t_i = \text{round}\left(\frac{i \cdot n}{l-1}\right), \quad i = 0, 1, \dots, l-1$$

and for each  $t_i$ , we select the closest available index in  $\mathcal{S}$  without duplication.

This strategy ensures that manually annotated indices are selected while interpolating additional indices to maximize temporal coverage.

**Prompts for zero-shot evaluation** The prompt used for the frame-wise and video baseline using Sa2VA is provided in Table 26. The prompt used for GT + Sa2VA baseline is provided in Table 27.

**Grouping for evaluation** We group videos into three size bins—small (S), medium (M), and large (L)—based on the average area of their ground-truth masks, so we can examine performance as a function of target size. The S/M/L thresholds differ between the **Area** and **Object** settings:

- **Area:** S→M at 372, M→L at 2,127 (pixels)
- **Object:** S→M at 3,612, M→L at 13,231 (pixels)

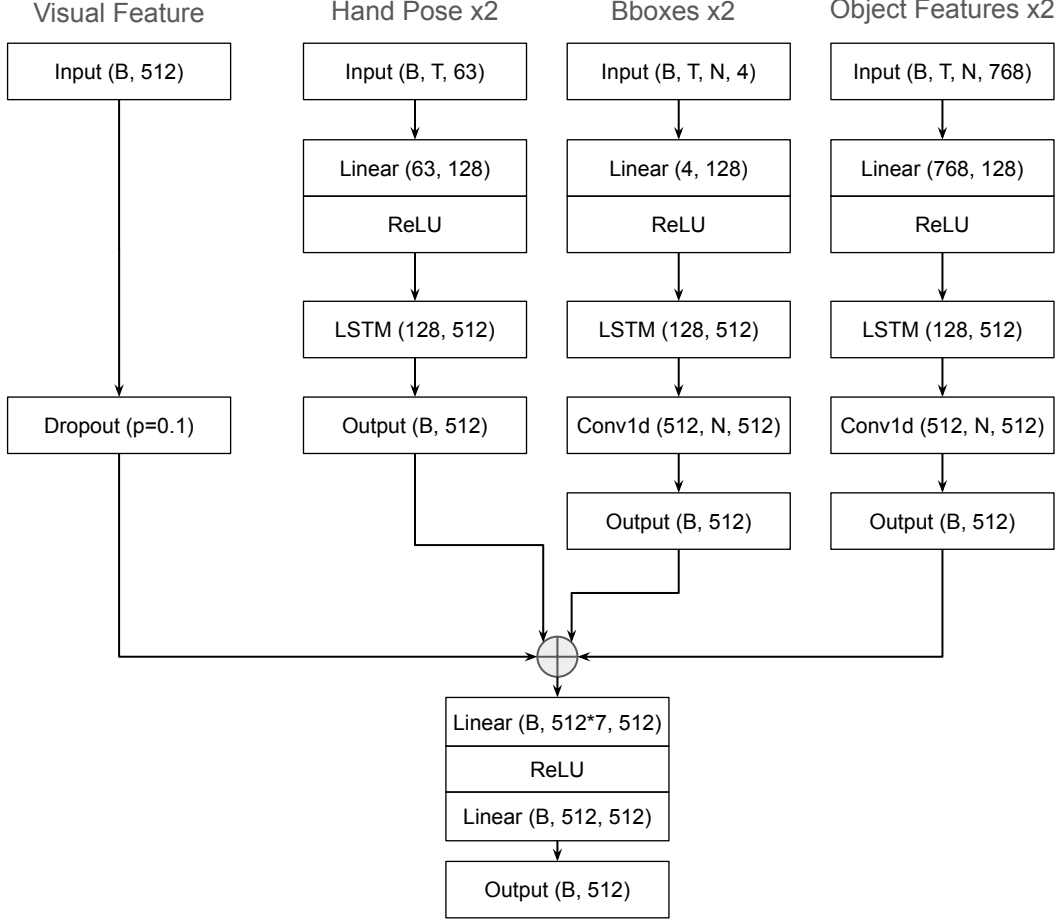


Figure 14: Architecture of additional head used for fine-tuning experiment. Branches are omitted for unused modalities.

**Computational cost** Sa2VA model fit on a single NVIDIA H200 GPU and took approximately 2 hours for the video baseline and 1.5 hours for the frame-wise baseline per category.

### C.3 Integration of HOI cues

**Implementation details** Figure 14 illustrates the architecture of our model, which integrates multiple modalities: frame-level rgb visual features, hand poses, bounding boxes of manipulated objects, and visual features of the manipulated objects.

Hand poses are extracted using an off-the-shelf 3D hand pose detector [31], resulting in a tensor of shape  $[B, T, 21, 3]$  for each hand, where  $B$  is the batch size and  $T$  is the number of frames in the video. The bounding boxes of manipulated objects for each hand are obtained using AMEGO [13] and represented as  $[B, T, N, 4]$ , where  $N$  is the maximum number of objects per hand. We set  $N = 8$  in our experiments. For each detected bounding box, we crop the corresponding image region and extract CLIP features [33]. For the visual feature, we add a dropout layer to encourage the model to use other modalities.

For each modality (excluding the global rgb visual feature), we use separate processing modules for the left and right hands, resulting in a total of seven feature vectors, each with a shape of  $[B, 512]$ . These features are concatenated and passed through a multi-layer perceptron (MLP) to produce the final fused representation of shape  $[B, 512]$ .



**Training** For each integrated feature  $\mathbf{v}_i$ , we use one corresponding positive text feature and  $B_N$  negative text features that are sampled from the distractors of the same video and from options of the other videos in the batch to calculate contrastive loss in the following procedure:

- Normalize visual feature:  $\hat{\mathbf{v}}_i = \text{normalize}(\mathbf{v}_i)$
- Normalize text features:
  - Positive:  $\hat{\mathbf{p}}_i$
  - Negatives:  $\hat{\mathbf{n}}_{i,1}, \dots, \hat{\mathbf{n}}_{i,B_N}$
- Compute logits:  $\mathbf{s}_i = [\hat{\mathbf{v}}_i^\top \hat{\mathbf{p}}_i / \tau, \hat{\mathbf{v}}_i^\top \hat{\mathbf{n}}_{i,1} / \tau, \dots, \hat{\mathbf{v}}_i^\top \hat{\mathbf{n}}_{i,B_N} / \tau]$
- Define labels:  $\mathbf{y}_i = [1, 0, \dots, 0]$
- Compute loss:  $\mathcal{L}_i = \text{BCEWithLogits}(\mathbf{s}_i, \mathbf{y}_i)$

Total loss:  $\mathcal{L} = \frac{1}{B} \sum_{i=1}^B \mathcal{L}_i$

We trained separate models for each category. The batch size was set to 16, the learning rate to  $1 \times 10^{-5}$ , the weight decay to  $1 \times 10^{-4}$ , the number of negative samples  $B_N$  to 16, and the temperature  $\tau$  to 0.07. Each model was trained for 500 epochs, and the best-performing checkpoint was selected based on validation performance.

**Computational cost** Each training job fit on a single NVIDIA H200 GPU and finished in roughly 0.05–2.5 hours per category, depending on the amount of input data.

Carefully watch the first-person view video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons.

Question: {question}

Choose **only one** option from the following list.

Options:

- (A) {option1}
- (B) {option2}
- (C) {option3}
- (D) {option4}
- (E) {option5}

Answer format:

(A) <Description of Option A>

Table 24: Prompt for zero-shot evaluation of video-language models integrated with LLMs when there is only one correct answer. {question} is replaced with question and {option $n$ } is replaced with  $n$ -th option.

Carefully watch the first-person view video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons.

Question: {question}

Choose **\*\*all\*\*** options that apply from the following list.

Options:

(A) {option1}

(B) {option2}

(C) {option3}

(D) {option4}

(E) {option5}

...

Answer format:

(A) <Description of Option A>

Table 25: Prompt for zero-shot evaluation of video-language models integrated with LLMs when there are multiple correct answers. {question} is replaced with question and {option $n$ } is replaced with  $n$ -th option.

Segment the area that corresponds to the answer to the question.

Question: {question}

Table 26: Prompt for Sa2VA (frame-wise/video) baseline in referring video object segmentation. {question} is replaced with question.

Segment all the mentioned area:

{GT}

Table 27: Prompt for GT + Sa2VA baseline in referring video object segmentation. {GT} is replaced with ground truth.

## D Fine-tuning Qwen2.5-VL-7B

We conducted instruction tuning on the training split of HanDyVQA to explore the effectiveness of fine-tuning. Specifically, we used Qwen2.5-VL-7B as the base model and trained LoRA adapters using QLoRA [9].

**Implementation details** We trained separate models for each category. We only trained the LoRA parameters injected into the query and value projection layers (q\_proj and v\_proj) of the attention modules, while keeping all other model weights frozen. The number of input video frames is 16, and the resolution is  $480 \times 854$ . We set the batch size to 4, learning rate to  $1 \times 10^{-4}$ . Each model was trained for 150 epochs, and the best-performing checkpoint was selected based on the validation performance.

**Computational cost** Training can be performed on a single NVIDIA H200 GPU and takes roughly two hours per category.

**Results** Table 28 shows the results of fine-tuning the model. The **Action**, **Process**, and **State** categories show gains of around 10 points, while improvements in the **Location** and **Parts** categories are less than 5 points. A performance drop was observed in the **Objects** category. These results suggest that the model has learned textual biases at some extent—particularly those that appear more strongly in longer sentences—and that fine-tuning the language decoder does not improve visual recognition abilities, even for relatively simple tasks like identifying manipulated objects.

Models	Action	Process	Location	State	Parts	Avg.	Objects
Qwen2.5-VL-7B zero-shot	61.9	55.6	56.8	48.3	48.2	54.1	<b>57.5</b>
Qwen2.5-VL-7B fine-tuned	<b>71.5</b>	<b>68.4</b>	<b>61.5</b>	<b>59.5</b>	<b>51.1</b>	<b>62.4</b>	49.7

Table 28: Results of fine-tuning Qwen2.5-VL-7B model on HanDyVQA.

## E Broader Impacts

The proposed HanDyVQA dataset provides a detailed evaluation of fine-grained hand-object interactions. As such, it serves as a valuable benchmark for systems designed to assist human workers using visual information captured by wearable cameras in diverse real-world scenarios [29]. This enables the development of systems that can better understand subtle interactions and deliver more accurate and context-aware feedback to the users.

Such recognition capabilities are also essential for applications in Augmented Reality (AR) and Virtual Reality (VR), where systems must respond to user actions and changes in the environment in real time. Unlike previous datasets that focus primarily on action recognition or object detection, HanDyVQA offers a unique benchmark that evaluates a model’s ability to comprehend nuanced hand-object interactions, pushing the boundaries beyond conventional video recognition tasks.

## 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] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.
- [10] 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.
- [11] 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.
- [12] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.

- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, 132(11):4880–4936, 2024.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.

- [35] 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.
- [36] Fadime Sener, Dibyadip Chatterjee, Daniel Sheleпов, 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.
- [37] 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.
- [38] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pages 654–665. PMLR, 2023.
- [39] 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.
- [40] 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.
- [41] Sirnam Swetha, Hilde Kuehne, and Mubarak Shah. Timelogic: A temporal logic benchmark for video qa. *arXiv preprint arXiv:2501.07214*, 2025.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] 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.

- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] 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.
- [60] 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.
- [61] 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.