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

Masatoshi Tateno[1,2,a)] Kato Gido[1,3,b)] Kensho Hara[1,c)]
Hirokatsu Kataoka[1,4,d)] Yoichi Sato[2,e)] Takuma Yagi[1,f)]

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 evaluate several video foundation models on our benchmark to identify existing challenges and reveal that current models struggle with component-level geometric understanding, achieving a average accuracy of only 68% in Qwen2.5-VL-72B.

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 [31], and motor function analysis [35].

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], [13], [17], [40], long-form actions [22], and procedural steps [29], [32], [44] or (ii) low-level localization such as hand-object localization [2], [5], [30] and hand pose estimation [9], [25], [49] while neglecting the semantically rich aspects of hand-object dynamics.

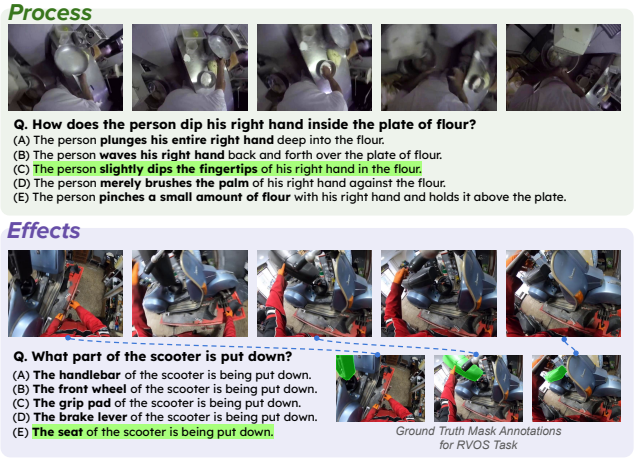

Fig. 1: Example questions from HanDyVQA Dataset.

We propose HanDyVQA (**Hand Dy**namics **V**ideo **QA**), a video question-answering benchmark designed to evaluate spatiotemporal reasoning in dynamics of HOI (see example in Figure 1). HanDyVQA aims to study the local spatio-temporal dynamics in HOI, requiring an understanding 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 from Ego4D [13] dataset, which features diverse hand-object interactions in real-world settings. We provide six types of multi-choice question-answering tasks totalling 11.7k carefully designed samples that avoid trivial shortcuts, along with segmentation tasks for two question types (objects and object parts) totaling 11k instances for fine-grained spatial understanding.

We first evaluate existing video recognition models to assess their ability to capture hand-object interactions. Even the best performing model struggles across all categories, achieving only 61% to 77% accuracy. This suggests current architectures fail to effectively encode sequential and dynamic information crucial for detailed HOI understanding. We also examine the impact of input frame count on performance, providing insights into models ' temporal reasoning

1 National Institute of Advanced Industrial Science and Technology (AIST)
2 The University of Tokyo
3 Waseda University
4 University of Oxford
a) masatate@iis.u-tokyo.ac.jp
b) katogido2018@fuji.waseda.jp
c) kensho.hara@aist.go.jp
d) hirokatsu.kataoka@aist.go.jp
e) ysato@iis.u-tokyo.ac.jp
f) takuma.yagi@aist.go.jp

	Objective	Source	View	Question Scope					Answer Type	#Questions	Avg. Duration
				HOI-Oriented?	Spatial	Temporal	Process	Effect			
Next-QA [39]	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]	Object / 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 [40]	Action / Objects	Ego4D	FPV	✓	✓	✓	✗	✗	MC	30K	1 s
AMB [12]	Long-Term Object Interactions	EK	FPV	✓	✓	✓	✗	✗	MC	21K	20 m
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, while OP represents open-ended question-answering. BBox indicates bounding box, and Seg refers to segmentation.

capabilities.

Overall, our new dataset and experimental results underscore the importance of understanding HOI spatio-temporal dynamics and offer valuable insights that can guide future research endeavors.

2. Related Work

Video question answering benchmarks. Video question answering (VideoQA) is challenging due to motion, events, and actions unfolding over time. Traditional benchmarks [41], [46] focus on identifying human actions, events, or objects in short clips. Recent efforts address long-form video understanding [22], [37], [38], [52], while works like NExT-QA [39] and TimeLogic QA [34] explore temporal and causal relationships. MVBench [18] tackles temporal understanding in a multiple-choice QA format. However, none of these benchmarks emphasize fine-grained hand-object interactions (HOIs), including hand-object coordination, handling manner, and resulting effects.

Hand-object interaction recognition benchmarks. Various HOI recognition benchmarks focus on (i) low-level localization, such as detecting hands and objects [30], estimating 3D poses [3], [14], mesh reconstruction [33], and object tracking [1], [12], and (ii) high-level actions, leveraging egocentric datasets like EPIC-KITCHENS [7] and Ego4D [13][2], [4], [12], [22], [40]. AMEGO[12] curates long-term object tracks and spatiotemporal questions, while EgoHOIBench [40] and HOI-QA [2] explore open-vocabulary recognition and entity relationships. However, these works lack fine-grained HOI understanding, including processes, effects, and component-level spatiotemporal reasoning.

Vision-and-language models for video understanding. The rise of dual-encoder vision-language models trained on large-scale image-text pairs [27], [48] has advanced video understanding. Approaches include adapting image models for video [21], [43], leveraging instructional videos [23], [24], and pretraining first-person models [20], [51]. Inspired by LLMs, recent models integrate visual encoders and LLMs for general video comprehension [2], [6], [36], [45], [50], scaling parameters and datasets. However, they rely on frame-based architectures, often ignoring local entities and spatiotemporal dynamics like hand poses and state changes. HanDyVQA introduces a new challenge, enhancing HOI recognition.

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.

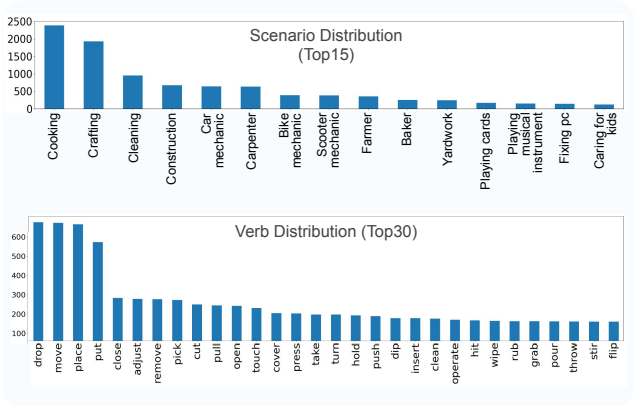

Fig. 2: Scenario distribution in HanDyVQA. Sentence with green highlights and green region in images denote correct answer and ground truth masks, respectively.

	Action	Process	Location	State	Parts	Objects
#Questions	1978	1924	1974	1940	1913	1939
#Options	5	5	5	5	5	5.7
#Correct Answers	1	1	1	1	1	1.6
#Words	18.1	20.2	12.3	13.0	8.9	1.4

Table 2: Statistics for each question type.

	#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

Table 3: Statistics for segmentation annotation for Parts and Objects questions.

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.

In this section, we describe details of our data collection process (Section 3) and its analysis (Section 3).

QA collection. To generate challenging and diverse QAs on HOI dynamics, we developed a collaborative framework that combines AI-generated QAs with human verification.

Data curation. We build our benchmark on Ego4D [13] due to its diverse, unscripted hand-object interactions across various scenarios. Using provided narrations and timestamps, we identify short clips capturing these interactions. Narrations help determine if an HOI event is occurring, which we verify using LLMs to infer contacted objects and secondary objects [5]. Clips where no object is manipulated are filtered. We then sample 2,000 narrations per category based on relevant verbs to generate questions. Each narration corresponds to a 5-second video segment centered around its timestamp. See supplementary for details.

Automatic question generation. QA pairs are automatically generated from narrations following the templates below:

- **Action:** What is the person doing with his/her hands?
- **Process:** How does the person `[verb] [object]`?
- **Location:** Where does the person `[verb] [object]`?
- **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. For **Object Parts** questions, we ask LLMs to infer the plausible objects and effects to be inserted.

QA annotation by humans. Annotators verify, revise, or reject generated questions, ensuring they match the video content. They then provide detailed correct answers, with a list of plausible objects only for Objects. LLMs generate initial wrong choices, which annotators refine by removing overlaps, improving plausibility, and adding challenging distractors. This ensures all questions and choices are accurate, confusing, and human-solvable. See Figure 1 for an example.

Mask annotation by humans. For the **Objects** and **Parts** questions, annotators selected three clear frames and annotated the answer regions.

Dataset statistics. As a result, 11,668 QA pairs in total are curated for fine-tuning and evaluation. Table 2 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 1.4 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.

Mask annotation. Table 3 shows the number of annotated frames per each category, and the relative movement/spatial overlap between annotated frames. Due to the nature of object manipulation and moving cameras in egocentric videos, the segmentation masks move across frames, suggesting that they are not easy to predict.

4. Experiments

To reveal the challenges in recognizing the dynamic aspects of HOI in HanDyVQA, we compare the performance of major existing Video-Language Models. We conduct separate evaluations for MCQ and RVOS.

4.1 Multi-Choice Questions

We first compare the zero-shot performance across off-the-shelf Video-Language Models, and conduct an ablation study changing the the number of input frames and input resolution.

Baseline models. We choose the following six open-source video LLMs and one proprietary model that have different visual backbone, language models, and training data:

LaViLa [51]: Video LLM trained on egocentric videos without an LLM. **VideoLLaMA2.1-7B** [6]: Video LLM specializing in spatio-temporal modeling of video information. **LLaVa-Video-7B** [19]: Video LLM trained on video datasets, including egocentric videos. **mPLUG-Owl3-8B** [45]: Video LLM capable of efficiently processing longer image sequences in a single prompt. **Qwen2.5-VL-7B/72B** [36]: Video LLM that accepts videos with arbitrary resolutions. **GPT-4o (text-only / vision)** [15]: Proprietary model capable of processing image sequences.

Implementation details. We uniformly sample 16 frames from each video and use the default input resolution for each model. All models are evaluated in a zero-shot setting. For dual-encoder models such as LaViLa, 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** which have more than one answers.

Quantitative results. Table 4 shows the results. First, GPT-4o (text) showed better results (34–50pts) compared to the random baseline, suggesting come textual bias exists but not enough to solve the task. The dual encoder-based LaViLa did not achieve more than a GPT-4o (text) baseline except **Action** and **Object** category even they are fine-tuned by the Ego4D dataset, suggesting that they primarily learns actions and objects appearing in videos. Models with LLM decoders performed better than text-only model, following similar trends compared to general videos [28]. It is worth mentioning that LLaVA-Video-7B which is fine-tuned on Ego4D achieved the best score of 54.1 % across 7B-sized models, suggesting the merits of adaptation to the target

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
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 4: 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.

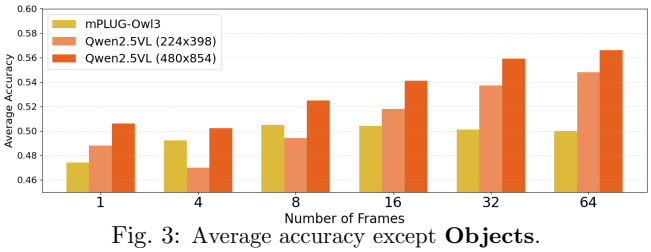

Fig. 3: Average accuracy except **Objects**.

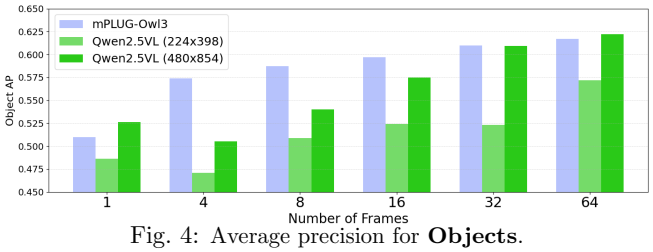

Fig. 4: Average precision for **Objects**.

Models	Objects		Parts	
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Sa2VA (Frame-wise)	0.359	0.294	0.126	0.099
Sa2VA (Sequentially)	0.319	0.312	0.109	0.101
GT option + Sa2VA	0.259	0.182	0.119	0.089

Table 5: Results of RVOS.

domain. Qwen2.5-VL-72B achieved the best results among all models, even better than the proprietary GPT-4o vision model. However, all the models showed unsatisfying results of at most 61–77% top-1 accuracies across categories, suggesting that large video foundation models does not fully capture the fine-grained aspects of HOIs. We show qualitative results in the supplementary material.

Ablations on number of frames and resolution. We measure the effect of input frame count and resolution on mPLUG-Owl3, Qwen2.5-VL. For mPLUG-Owl3, we changed the number of input frames. For Qwen2.5-VL, we evaluated both different frame counts and resolutions.

Figure 3 and 4 present the quantitative results. Performance improved with more frames and resolutions, highlighting the importance of temporal context and spatial resolution. However, mPLUG-OWL3 showed no gains beyond eight frames in all categories except **Object**. This suggest it does not fully leverage increased temporal information, underscoring the need for advanced architectures.

4.2 RVOS

Baseline models. We compare two models in four settings: **Sa2VA** [47]: An MLLM model for referring image and video segmentation, applied **frame-wise** (no temporal context) and **sequentially** (leverages temporal context). **Ground truth option + Sa2VA**: Uses the ground truth answer as input for segmentation.

Evaluation Metrics. Following the standard VOS evaluation protocols [26], [42], we use the Jaccard Index ($\mathcal{J}$) and Boundary F-measure ($\mathcal{F}$) for each frame and take the average of all the annotated frames.

Implementation details. We input 16 frames for the video segmentation baseline for Sa2VA uses, including the annotated frames, ensuring they are sampled as evenly as possible across the entire video. Frame-wise methods are applied only to the annotated frames each.

Results. As shown in Table 5, all models performed significantly worse than prior egocentric segmentation tasks (*e.g.*, 70+ $\mathcal{J}$ in [8]), especially for **Parts**. The single-stage model (**Sa2VA**) outperformed two-stage models. Frame-wise results had a slightly higher $\mathcal{J}$ than the sequential method, suggesting current models struggle with the temporal context in fast-moving egocentric videos. Results from **GT option + Sa2VA** show that textual answers alone are insufficient for precise region reference in HOIs. Qualitative results are in the supplementary material.

5. Conclusion

We have proposed HanDyVQA, a new video QA benchmark for evalutating abundant spatio-temporal dynamics, process, and effects contained in hand-object interactions. Experimental results show that strong video-language models struggle with fine-grained details of HOIs, only achieving at most 61–77% top-1 accuracy in MCQ setting, and showing poor performance in referring local interacting regions. Ablation study suggests a space for improvements to model the local spatiotemporal dynamics.

References

[1] Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R. et al.: Introducing HOT3D: An Egocentric Dataset for 3D Hand and Object Tracking, *arXiv preprint arXiv:2406.09598* (2024).

[2] Bansal, S., Wray, M. and Damen, D.: HOI-Ref: Hand-Object Interaction Referral in Egocentric Vision, *arXiv preprint arXiv:2404.09933* (2024).

[3] Chao, Y.-W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y. S., Van Wyk, K., Iqbal, U., Birchfield, S. et al.: DexYCB: A benchmark for capturing hand grasping of objects, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9044–9053 (2021).

[4] Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H. and Liu, Y.: Egothink: Evaluating first-person perspective thinking capability of vision-language models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14291–14302 (2024).

[5] Cheng, T., Shan, D., Hassen, A., Higgins, R. and Fouhey, D.: Towards a richer 2d understanding of hands at scale, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 30453–30465 (2023).

[6] Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D. et al.: VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs, *arXiv preprint arXiv:2406.07476* (2024).

[7] Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W. et al.: The epic-kitchens dataset: Collection, challenges and baselines, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 11, pp. 4125–4141 (2020).

[8] Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D. and Damen, D.: Epic-kitchens visor benchmark: Video segmentations and object relations, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 13745–13758 (2022).

[9] Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M. J. and Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954 (2023).

[10] Flaborea, A., Di Melendugno, G. M. D., Plini, L., Scofano, L., De Matteis, E., Furnari, A., Farinella, G. M. and Galasso, F.: PREGO: online mistake detection in PRocedural EGOcentric videos, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18483–18492 (2024).

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

[12] Goletto, G., Nagarajan, T., Averta, G. and Damen, D.: AMEGO: Active Memory from long EGOcentric videos, *European Conference on Computer Vision*, Springer, pp. 92–110 (2024).

[13] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X. et al.: Ego4d: Around the world in 3,000 hours of egocentric video, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012 (2022).

[14] Hampali, S., Rad, M., Oberweger, M. and Lepetit, V.: Honnotate: A method for 3D annotation of hand and object poses, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3196–3206 (2020).

[15] Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A. et al.: Gpt-4o system card, *arXiv preprint arXiv:2410.21276* (2024).

[16] Jia, B., Lei, T., Zhu, S.-C. and Huang, S.: Egotaskqa: Understanding human tasks in egocentric videos, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 3343–3360 (2022).

[17] Kwon, T., Tekin, B., Stühmer, J., Bogo, F. and Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10138–10148 (2021).

[18] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P. et al.: Mvbench: A comprehensive multi-modal video understanding benchmark, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206 (2024).

[19] Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P. and Yuan, L.: Video-LLaVA: Learning United Visual Representation by Alignment Before Projection, *arXiv preprint arXiv:2311.10122* (2023).

[20] Lin, K. Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, E. Z., Gao, D., Tu, R.-C., Zhao, W., Kong, W. et al.: Egocentric video-language pretraining, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 7575–7586 (2022).

[21] Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N. and Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval, *arXiv preprint arXiv:2104.08860* (2021).

[22] Mangalam, K., Akshulakov, R. and Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 46212–46244 (2023).

[23] Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J. and Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9879–9889 (2020).

[24] Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I. and Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640 (2019).

[25] Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L. and Keskin, C.: AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12999–13008 (2023).

[26] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A. and Van Gool, L.: The 2017 davis challenge on video object segmentation, *arXiv preprint arXiv:1704.00675* (2017).

[27] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning transferable visual models from natural language supervision, *International conference on machine learning*, PMLR, pp. 8748–8763 (2021).

[28] Salehi, M., Park, J. S., Yadav, T., Kusupati, A., Krishna, R., Choi, Y., Hajishirzi, H. and Farhadi, A.: ActionAtlas: A VideoQA Benchmark for Domain-specialized Action Recognition, *arXiv preprint arXiv:2410.05774* (2024).

[29] Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R. and Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106 (2022).

[30] Shan, D., Geng, J., Shu, M. and Fouhey, D. F.: Understanding human hands in contact at internet scale, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9869–9878 (2020).

[31] Shaw, K., Bahl, S. and Pathak, D.: Videodex: Learning dexterity from internet videos, *Conference on Robot Learning*, PMLR, pp. 654–665 (2023).

[32] Song, Y., Byrne, E., Nagarajan, T., Wang, H., Martin, M. and Torresani, L.: Ego4D Goal-Step: Toward hierarchical understanding of procedural activities, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 38863–38886 (2023).

[33] Swamy, A., Leroy, V., Weinzaepfel, P., Baradel, F., Galaaoui, S., Brégier, R., Armando, M., Franco, J.-S. and Rogez, G.: Showme: Benchmarking object-agnostic hand-object 3D reconstruction, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1935–1944 (2023).

[34] Swetha, S., Kuehne, H. and Shah, M.: TimeLogic: A Temporal Logic Benchmark for Video QA, *arXiv preprint arXiv:2501.07214* (2025).

[35] Tsai, M.-F., Wang, R. H. and Zariffa, J.: Recognizing hand use and hand role at home after stroke from egocentric video, *PLOS Digital Health*, Vol. 2, No. 10, p. e0000361 (2023).

[36] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W. et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, *arXiv preprint arXiv:2409.12191* (2024).

[37] Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y. et al.: Lvbench: An extreme long video understanding benchmark, *arXiv preprint*

arXiv:2406.08035 (2024).

[38] Wu, H., Li, D., Chen, B. and Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding, *arXiv preprint arXiv:2407.15754* (2024).

[39] Xiao, J., Shang, X., Yao, A. and Chua, T.-S.: Next-qa: Next phase of question-answering to explaining temporal actions, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786 (2021).

[40] Xu, B., Wang, Z., Du, Y., Song, Z., Zheng, S. and Jin, Q.: Do Egocentric Video-Language Models Really Understand Hand-Object Interactions?, *Proceedings of the International Conference on Learning Representations* (2025).

[41] Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X. and Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion, *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653 (2017).

[42] Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S. and Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation, *Proceedings of the European conference on computer vision (ECCV)*, pp. 585–601 (2018).

[43] Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H. and Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment, *arXiv preprint arXiv:2209.06430* (2022).

[44] Yagi, T., Ohashi, M., Huang, Y., Furuta, R., Adachi, S., Mitsuyama, T. and Sato, Y.: FineBio: a fine-grained video dataset of biological experiments with hierarchical annotation, *arXiv preprint arXiv:2402.00293* (2024).

[45] Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F. and Zhou, J.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, *arXiv preprint arXiv:2408.04840* (2024).

[46] Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y. and Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 9127–9134 (2019).

[47] Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J. and Yang, M.-H.: Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos, *arXiv* (2025).

[48] Zhai, X., Mustafa, B., Kolesnikov, A. and Beyer, L.: Sigmoid loss for language image pre-training, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986 (2023).

[49] Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K. and Lu, C.: Oakink2: A dataset of bimanual hands-object manipulation in complex task completion, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456 (2024).

[50] Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z. and Li, C.: Video instruction tuning with synthetic data, *arXiv preprint arXiv:2410.02713* (2024).

[51] Zhao, Y., Misra, I., Krähenbühl, P. and Girdhar, R.: Learning video representations from large language models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597 (2023).

[52] Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T. and Liu, Z.: MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding, *arXiv preprint arXiv:2406.04264* (2024).