HandsOnVLM: Vision-Language Models for Hand-Object Interaction Prediction

Chen Bao[1], Jiarui Xu[2], Xiaolong Wang[2,†], Abhinav Gupta[1,†], Homanga Bharadhwaj[1,†]
[1] The Robotics Institute, Carnegie Mellon University
[2] University of California at San Diego

Abstract

How can we predict future interaction trajectories of human hands in a scene given high-level colloquial task specifications in the form of natural language? In this paper, we extend the classic hand trajectory prediction task to several tasks involving explicit and implicit language queries. Our proposed tasks require extensive understanding of human daily activities and reasoning abilities about what is happening next given cues from the current scene. We also develop new benchmarks to evaluate the proposed two tasks, Vanilla Hand Prediction (VHP) and Reasoning-Based Hand Prediction (RBHP). We enable solving these tasks by integrating high-level world knowledge and reasoning capabilities of Vision-Language Models (VLMs) with the auto-regressive nature of low-level ego-centric hand trajectories. Our model, HandsOnVLM is a novel VLM that can generate textual responses and produce future hand trajectories through natural-language conversations. Our experiments show that HandsOnVLM outperforms existing task-specific methods and other VLM baselines on proposed tasks, and demonstrates its ability to effectively utilize world knowledge for reasoning about low-level human hand trajectories based on the provided context.

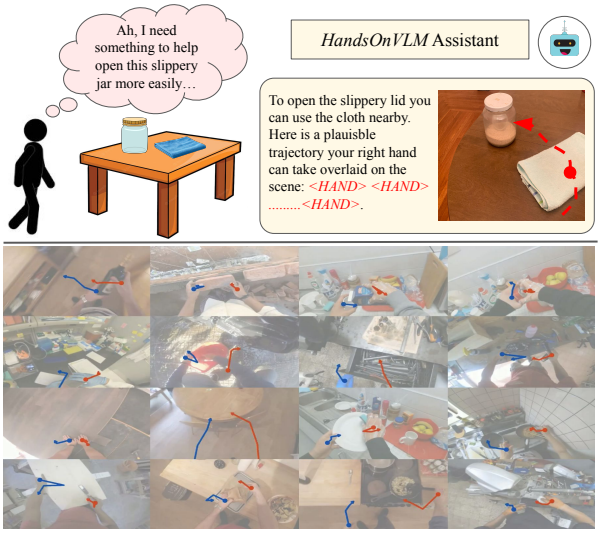

Figure 1. *HandsOnVLM* forecasts low-level actions in the form of hand trajectories in the user's egocentric view of a scene when queried with a question via natural language. It is also capable of handling indirect queries that require reasoning about *what* object to interact with and *how* to perform the interaction. [Bottom] We show a glimpse of left and right hand trajectory predictions from *HandsOnVLM* across diverse real-world scenarios.

1. Introduction

Humans interact with the everyday world and express themselves with informal and oftentimes vague language descriptions. Consider the example in Fig. 1 - when we try to open the jar, we might think, "Ah, I need something to help open this slippery jar more easily." We are uncertain about *what* we exactly want as well as about *how* to come up with a solution. Building a computational system addressing this need would require a good understanding of what tools we have lying around (visual scene understanding), general apriori experience of opening jars (reasoning ability and world knowledge priors), and the ability to actually execute the necessary actions for opening the jar (low-level trajectory). In this paper, we develop language-conditioned prediction tasks for tackling this problem, propose benchmarks for evaluating progress on these tasks, and build a vision-language model (VLM) for predicting low-level hand trajectories in a user's egocentric view of a scene given colloquial language queries.

Our approach *HandsOnVLM* casts hand trajectory prediction as an auto-regressive next token prediction conditioned on fused video and language tokens. We develop *HandsOnVLM* as an interactive chat assistant that we can query. The *HandsOnVLM* model first converts the RGB video context to visual tokens and fuses them with the language tokens through slow-fast pooling [8] for capturing temporal information from the context video

at a fine resolution. We extend the vocabulary to add a new <*HAND*> token, and output a sequence of text and and hand tokens. We finally have a trajectory decoder to convert the hand tokens to a sequence of 2D positions of the left and right hands over the prediction horizon.

2. Approach

HandsOnVLM is a video-based VLM with the capability of predicting future hand trajectories given a video context and language instructions. There are three key components of *HandsOnVLM*'s architecture: (1) SlowFast tokens to capture temporal information at fine temporal resolution, (2) hand representation using an augmented vocabulary of <*HAND*> token, and (3) iterative hand decoding to enable auto-regressive trajectory training and inference. In the training stage, we fine-tune a pre-trained VLM by combining next-token prediction loss and trajectory loss.

2.1. Architecture

HandsOnVLM takes a sequence of T frames X_v and a language instruction X_q as input and predicts future hand trajectories $\mathcal{H} = \{h_{T+i}\}_1^N$, where N is the future horizon. At each future time step $T+i$, the future hand location h_{T+i} consists of the 2D location of the center of the left and right hands projected to the last observation frame $X_v\,[-1]$. The key components of the architecture include a visual backbone $\mathcal{F}_{enc}$, a vision-to-language projection layer f, a Large Language Model(LLM) $\mathcal{F}$ and a trajectory decoder $\mathcal{F}_{dec}$.

SlowFast Token Compression. To obtain a capable video-conditioned VLM we need to be able to interpret temporal information at a fine resolution. Following [8], given X_v, we embed them into $T \times M$ visual tokens using a visual backbone, where M is the number of tokens in each frame. Then we apply slow-fast pooling to get $T + M$ visual tokens. Then we embed and align those visual tokens to the language space through a vision-to-language projector $f(\cdot)$.

Hand as Embedding. To represent hand in the language space, we extend the existing vocabulary with a new <*HAND*> token. However, a typical embedding layer would encode each <*HAND*> token identically, resulting in individual <*HAND*> token being indistinguishable from one another. To overcome this limitation, we embed ground truth hand positions into the <*HAND*> tokens during the tokenization process. We feed them into the Large Language Model backbone and get the embedding of the last layer H, where $H = \mathcal{F}(X_q, f(\mathcal{F}_{enc}(X_v)))$.

Iterative Hand Decoding. For i-th token in the sequence, let H_i be the last-layer embedding of this token from the Large Language Model. *HandsOnVLM* decode it to predict the $(i+1)$-th token as LLMs do. When $(i+1)$-th token is a <*HAND*> token, we input H_i into a hand trajectory decoder $\mathcal{F}_{dec}$ to predict the hand position of the $(i+1)$-th token $h_{i+1} = \mathcal{F}_{dec}(H_i)$. During inference, this decoded position is then encoded into the corresponding <*HAND*> token embedding for following prediction rounds. In this way, we ensure that each subsequent prediction is conditioned on all previously predicted hand positions, maintaining temporal consistency and spatial awareness throughout the inference process and mitigating compounding errors.

2.2. Training Objectives

The model is trained end-to-end using a text generation loss $\mathcal{L}_{\text{txt}}$ and a hand trajectory prediction loss $\mathcal{L}_{\text{hand}}$. The overall objective $\mathcal{L}$ is the weighted sum of both losses, determined by λ_{txt} and λ_{hand}:

$$\mathcal{L} = \lambda_{\text{txt}}\mathcal{L}_{\text{txt}} + \lambda_{\text{hand}}\mathcal{L}_{\text{hand}} \tag{1}$$

Specifically, $\mathcal{L}_{\text{txt}}$ is the auto-regressive cross-entropy loss for text generation, and $\mathcal{L}_{\text{hand}}$ is the hand prediction loss, which encourages the model to generate high-quality hand trajectories as well. Following [13], we employ a reconstruction loss over future timesteps and a KL-Divergence Regularization loss as $\mathcal{L}_{\text{hand}}$:

$$\mathcal{L}_{\text{hand}} = \sum_{t=1}^{N} \mathcal{L}_{\text{recon}} \left(h_{T+t}, \hat{h}_{T+t}\right) + \mathcal{L}_{kl}\left(\mu_h, \sigma_h\right). \tag{2}$$

3. Reasoning and Predicting Hand Trajectories

In this section, we introduce two tasks: the Vanilla Hand Prediction (VHP) task, which extends the classic hand motion prediction [13], and the proposed Reasoning-based Hand Prediction (RBHP) task. Finally, we describe a two-step annotation-generating pipeline to build the corresponding RBHP dataset.

3.1. Vanilla Hand Prediction Task

This task uses explicit language (e.g., "cut the paper") to predict hand motion. We use Epic-Kitchens [4, 5], H2O [11], and FPHA [6]. Hand trajectories are generated by detecting hands [17], tracking via SURF+RANSAC [1], smoothing with Hermite splines, and applying filtering.

Datasets are reformatted for VQA with the template: *"USER: <images>, can you give me the future hand trajectory for action? ASSISTANT: Sure, it is <HAND>...<HAND>."*

3.2. Reasoning-based Hand Prediction Task

RBHP requires predicting hand motion from colloquial, implicit instructions. We extract such queries using a two-step GPT-4-based pipeline applied to Epic-Kitchens[5] and Ego4D [7], yielding 7.5k and 8k examples respectively.

Step 1: GPT-4 generates a scene description with object and spatial context.

Step 2: It rewrites the action in an implicit manner using that context.

4. Experiments

We perform experiments to answer the questions:

- How plausible are the hand trajectories produced by *HandsOnVLM*?
- Does *HandsOnVLM* exhibit reasoning abilities for implicit language queries?
- Does *HandsOnVLM* generalize zero-shot to unseen scenes from new datasets?

4.1. Experiment Details

Architecture. We use CLIP-L/14 [16] as the visual encoder, Vicuna [3] as the LLM, and CVAE [18] as the decoder. The projector is from LLaVA [12].

Datasets. We sample 10 video frames and predict 4 future hand positions at 4 FPS. For *HandsOnVLM*†, we co-train on five additional tasks: ActivityNet-Captions [10], YouCook2 [20], NExT-QA [19], LLaVA-150K [12], and ActivityNet-RTL [8].

Training. *HandsOnVLM* is trained for 40 epochs with batch size 128 and learning rate 2e-5. We sample 24k VHP examples per epoch; *HandsOnVLM*† uses 6k VHP, 6k RBHP, and 12k from the other tasks. Visual backbone is frozen; other modules are fine-tuned. Training on 8 H100 GPUs takes 18 hours.

4.2. Metrics and Baselines

Following previous works [13, 15] we use Average Displacement Error (ADE), Final Displacement Error (FDE) and Weighted Displacement Error (WDE) as metrics to evaluate VHP and RBHP tasks.

Vanilla Hand Prediction. For the VHP task, we choose Kalman Filter(KF) and Object-centric Transformer(OCT) [13] as the baselines. Since OCT still requires the bounding box feature of the hand and object as input, to get a fairer comparison with other end-to-end methods, we implement a version without the requirement of the bounding box, which we call OCT-global.

Reasoning-based Hand Prediction. To evaluate *HandsOnVLM*'s performance on the RBHP task, we perform baseline comparisons with several VLM-based methods. We describe these basleines below:

- **LLaVA-Traj.** Note that the hand trajectories are a sequence of pixel positions, we can represent them in text directly. In this case, we can directly fine-tune the LLaVA without any modification.
- **LLaVA-Pixel2Seq.** An alternative approach to representing hand positions involves quantizing the image into discrete spatial bins [2], each corresponding to a unique token. We can extend the existing vocabulary with those discrete tokens.
- **Language-conditioned Image-to-Video Models.** We also compare our model to baselines of the language-conditioned image-to-video generation followed by hand-tracking. We use commercial state-of-the-art language-conditioned image-to-video systems such as LumaLabs [14], Kling 1.5 [9] and generate videos conditioned on the last observation frame and the language description. Following the hand label generation process in Sec. 3.1, we track and extract the hand trajectories of the generated video.

4.3. Comparisons with Baselines

We evaluate *HandsOnVLM* on both the VHP task and the proposed RBHP task and report the results and comparisons with baselines in Table 1 and Table 2 respectively. All models except *HandsOnVLM*† are trained on VHP datasets. *HandsOnVLM*† is trained on all available datasets.

VHP Task. We evaluate all the baselines on the VHP datasets as described in section 4.1. Here, the FPHA and H2O datasets serve as unseen datasets to test zero-shot generalization capabilities. Among all the VHP datasets, *HandsOnVLM* outperforms both the task-specific methods as well as the VLM-based methods, which demonstrates its strong ability to produce plausible trajectories corresponding to how a real human hand would move given explicit instructions. We also find that *HandsOnVLM* can generalize to completely unseen scenes (for example scenes from H2O and FPHA datasets), which demonstrates it can effectively leverage the world knowledge of the pre-trained VLM.

RBHP Task. For evaluations on the RBHP task shown in Table 2, *HandsOnVLM* achieves state-of-the-art performance in all three metrics. This suggests that *HandsOnVLM* is able to reason based on implicit cues of the scene and be applied to complicated scenarios involving everyday natural language conversations. However, we observe that LumaLabs [14] achieves the smallest ADE in the Ego4D RBHP benchmark but relatively higher FDE and WDE. This may be because the commercial text-conditioned image-to-video generation models have realistic video generation capabilities but cannot understand reasoning-based language prompts which is necessary for generating plausible videos maintaining temporal consistency. Since the training dataset compositions of these video models are not disclosed, there may also be some data leakage issues of the evaluation datasets in this paper being a part of their training corpora.

4.4. Qualitative Results

In Fig. 2 we show qualitative results for *HandsOnVLM* and the strongest baseline LLaVA-Pixel2Seq. The section above the horizontal line shows visualization from the validation split of RHBP datasets.

Approach	BBox Input	On Validation Split						Zero-shot					
		EK55			EK100			H2O			FPHA		
		ADE ↓	FDE ↓	WDE ↓	ADE ↓	FDE ↓	WDE ↓	ADE ↓	FDE ↓	WDE ↓	ADE ↓	FDE ↓	WDE ↓
KF	✓	0.392	0.386	0.199	0.317	0.318	0.168	-	-	-	-	-	-
OCT	✓	0.216	0.199	0.105	0.209	0.187	0.102	-	-	-	-	-	-
OCT-global		0.232	0.218	0.115	0.216	0.193	0.105	-	-	-	-	-	-
LLaVA-Pixel2Seq		0.156	0.139	0.076	0.254	0.224	0.124	0.150	0.121	0.032	0.214	0.189	0.043
LLaVA-Traj		**0.126**	0.142	0.073	0.201	0.191	0.103	**0.133**	0.130	0.031	0.191	0.167	0.041
HandsOnVLM		0.136	**0.106**	**0.062**	**0.194**	**0.157**	**0.090**	0.135	**0.108**	**0.028**	**0.175**	**0.151**	**0.034**

Table 1. Comparison of VHP task with different baselines. We reported the performance on the validation split of Epic-Kitchen dataset. For the RBHP baselines, we also evaluate them on two unseen datasets, H2O and FPHA.

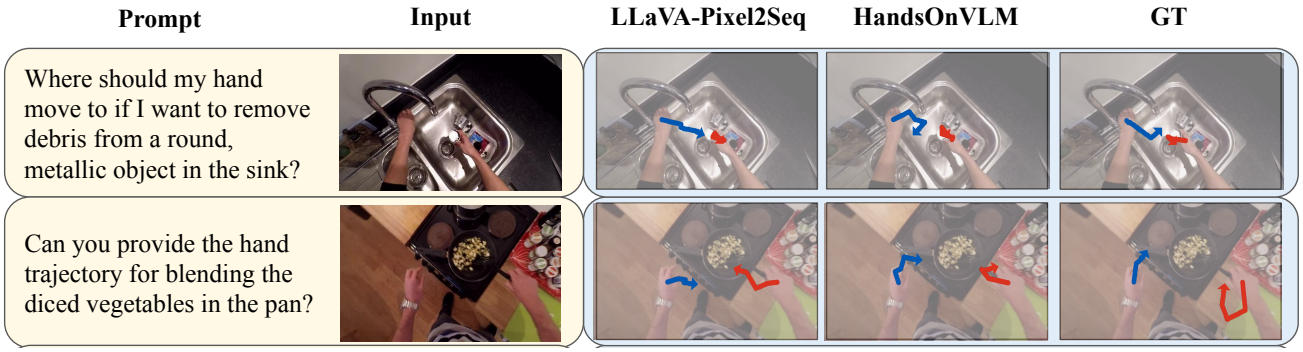

Figure 2. Qualitative results for different samples from the validation split of our RBHP dataset. The left-hand trajectory is visualized in blue and the right-hand trajectory is in red. The arrows denote the direction of each trajectory. GT trajectories are provided for reference.

Approach	RBHP (Epic-K)			RBHP (Ego4D)		
	ADE ↓	FDE ↓	WDE ↓	ADE ↓	FDE ↓	WDE ↓
Kling 1.5	0.31	0.35	0.19	0.27	0.41	0.18
LumaLabs	0.29	0.37	0.18	**0.21**	0.28	0.13
LLaVA-P2S	0.27	0.24	0.13	0.31	0.28	0.14
LLaVA-T	0.19	0.18	0.10	0.38	0.35	0.17
HandsOnVLM	0.19	0.16	0.09	0.22	0.19	0.10
HandsOnVLM†	**0.18**	**0.15**	**0.08**	0.22	**0.18**	**0.09**

Table 2. Comparison of *HandsOnVLM* on the RBHP task with different baselines. †means fine-tuned on the RBHP dataset.

4.5. Human Evaluation

Going beyond automated metrics, to determine plausibility of the generated hand trajectories for various scenarios, we also perform human evaluations. Results in Table 3 show that the predictions from *HandsOnVLM*† (the model fine-tuned on reasoning tasks) is more plausible for both VHP and RBHP evaluations, suggesting that model is able to effectively leverage world knowledge from other reasoning tasks to reason about low-level hand-object interaction predictions in diverse scenarios.

Method	VHP	RBHP
HandsOnVLM	$30 \pm 3\%$	$28 \pm 5\%$
HandsOnVLM†	$70 \pm 3\%$	$72 \pm 5\%$

Table 3. Human study showing the % mean and SE of trials where participants consider hand trajectory predictions from one method more plausible than the other.

5. Conclusion

In this work, we propose *HandsOnVLM*, a novel video-based VLM to predict future hand motion from ego-centric video contexts. We also propose different prediction tasks, including Vanilla Hand Prediction (VHP) and Reasoning-based Hand Prediction (RBHP) to benchmark low-level trajectory prediction and reasoning. We demonstrate the effectiveness of our approach through extensive evaluations in diverse real-world scenarios.

References

[1] Herbert Bay. Surf: Speeded up robust features. *Computer Vision—ECCV*, 2006. 2

[2] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 3

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhang-

hao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[6] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 2

[7] 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. 2

[8] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. *arXiv preprint arXiv:2403.19046*, 2024. 1, 2, 3

[9] KlingAI. Kling ai. https://klingai.com/, 2024. 3

[10] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[11] 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 (ICCV)*, pages 10138–10148, 2021. 2

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[13] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[14] LumaLabs. Dream machine. https://lumalabs.ai/dream-machine, 2024. 3

[15] Junyi Ma, Jingyi Xu, Xieyuanli Chen, and Hesheng Wang. Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos. *arXiv preprint arXiv:2405.04370*, 2024. 3

[16] 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. 3

[17] 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. 2

[18] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 3

[19] 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. 3

[20] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3