

1 We provide additional details, extended experimental results, and further discussion in this supple-  
 2 mentary material, including:

- 3 • Implementation details: background of foundation models (Appendix A.1), dataset details (Ap-  
 4 pendix A.2), and experimental details (Appendix A.3).
- 5 • Additional quantitative results: analyses and evaluations (Appendix B).
- 6 • Additional qualitative results: Examples and visualizations to complement the main results (Ap-  
 7 pendix C).
- 8 • Discussion on related works: Insights and comparisons with prior research (Appendix D).

## 9 A Implementation Details

### 10 A.1 Background of Foundation Models

11 **HOI Detector [21].** HOID is a robust system for detecting human hands and interacting objects  
 12 in images. It is built upon Faster-RCNN [19], pretrained on 100K image dataset with hand-object  
 13 interaction annotations including hand bounding box, interacting object bounding box, hand side  
 14 (left or right), hand contact state (e.g., no contact, self-contact, other person contact, contact with  
 15 portable object, or contact with a non-portable object). To enhance its capabilities, HOID is further  
 16 trained with an additional 42K egocentric data samples, enabling it to better understand HOI from  
 17 egocentric view. We leverage HOID to generate spatial HOI boxes, hand side, hand contact state  
 18 for each video, analyzing 12 uniformly sampled frames per video. However, HOID often generates  
 19 inconsistent boxes across consecutive frames. To address this issue, we leverage the robust image and  
 20 video segmentation model to refine the detection results, ensuring greater consistency and accuracy.

21 **Segment Anything 2 [18].** SAM2 is a versatile segmentation model capable of segmenting objects in  
 22 both images and videos according to a given prompt, such as a point, box or mask, with remarkable  
 23 speed. It is trained on a large-scale SA-V dataset, comprising 50.9K videos and 35.5M high-quality  
 24 masks. SAM2 employs a hierarchical image encoder and a memory mechanism to handle streaming  
 25 frame input. In our approach, SAM2 is utilized to generate the spatial-temporal consistent HOI masks  
 26 by leveraging prompts derived from HOID outputs.

27 **Depth Anything 2 [22].** DAv2 excels in monoc-  
 28 ular depth estimation, offering fine-grained de-  
 29 tails, robust generalization and fast inference  
 30 speed. It is built upon the pretrained visual founda-  
 31 tion model DINOv2 [15] and a depth decoder  
 32 DPT [17]. Pretrained on 595K synthetic im-  
 33 ages and 62M pseudo-labeled real images, DAv2  
 34 demonstrates strong out-of-domain generaliza-  
 35 tion capabilities. For our work, we use DAv2  
 36 to generate depth maps for eight frames sam-  
 37 pled from egocentric videos, serving as super-  
 38 vision signals. Following the recommendations  
 39 of DAv2, we employ the DAv2-Large variant,  
 40 which produces more spatial-temporal consis-  
 41 tent depth maps.

42 **DeepSeek-LLM [3].** We employ the LLM to  
 43 interpret HOI information and enrich textual  
 44 descriptions with shape and movement details.  
 45 DeepSeek-LLM demonstrates exceptional abil-  
 46 ity to follow instructions and comprehend HOI  
 47 mask prompts. Specifically, we use the API from DeepSeek, which operates the LLM with 200B  
 48 parameters, to facilitates these tasks.

### 49 A.2 Dataset Details

50 **Ego4D [6].** Ego4D contains 3,670 hours of egocentric videos with dense narrations, covering diverse  
 51 scenarios and activities from worldwide. Each narration is timestamped and paired with a free-  
 52 form sentence. Following the approach in Zhao [24], we construct 4M video-text clip pairs for

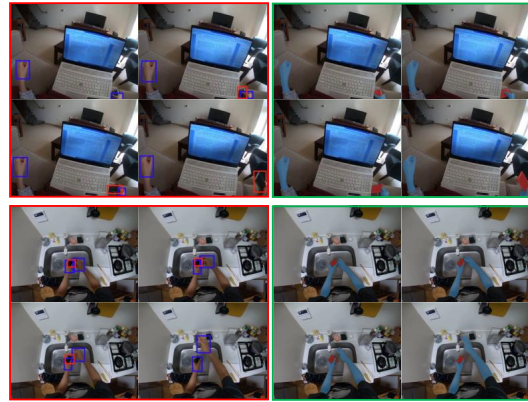


Figure 1: Comparisons of the noisy HOI bounding boxes (left) and the spatial-temporal consistent HOI masks (right).

pre-training, with an average clip length of 1 second ( $\pm 0.9$ ). In our text enrichment process, we only keep those hand-object interaction clips performed by the camera wearer, where the text begins with ‘#C’ (denoting the wearer) and then follows HOI-related verbs and nouns. This strategy excludes clips that record other people’s activities, such as multi-person interactions [20] where the text begins with ‘#O’, and the videos like ‘#C C walks away’. For the natural language query task, it comprises 1,659 untrimmed videos, each averaging 500 seconds in duration. On average, each video contains 12 clip-query pairs. Following the official split from [6], we use 11,291 queries for training and 3,874 for validation.

**Epic-Kitchens [2].** Epic-Kitchens-100 (EK-100) consists of 100 hours of egocentric cooking videos divided into training (67,217 clips), validation (9,668 clips), and testing (13,092 clips) splits. Each clip includes start and end timestamps, a short textual narration, and a verb and noun class that correspond to the narration. There are 3805 action classes, 97 verb classes, and 300 noun classes. We evaluate our pre-trained model on the validation split.

**EGTEA [12].** EGTEA comprises 28 hours of egocentric cooking videos, annotated with 10,321 instances of fine-grained actions across 106 classes. The average action duration is 3.2 seconds. For our experiments, we use only the visual frames as input. We follow prior works [10, 24] and report top-1 accuracy and mean class accuracy on all three test splits, including 2,022 testing instances for each split.

**H2O [11].** H2O is a dataset capturing egocentric hand-object interactions in a laboratory, including 36 action classes. The egocentric data is captured from an Azure Kinect camera mounted egocentrically for recordings. Since our primary target is to evaluate the transfer learning capability of visual representation, the train/val splits have 7862/11638 frames.

### A.3 Experimental Details

**Zero-Shot Video-Text Retrieval in EK-100-MIR and EgoMCQ** We perform video-text matching using our dual encoders with 16 frames as input for EK-100-MIR and 4 frames for EgoMCQ, following [23].

**Zero-Shot Action Recognition in EGTEA.** We follow the evaluation protocol proposed by [12] to compute the mean performance across all evaluation splits. This involves performing video-text retrieval between video clips and their associated action text labels, which are prompted by prepending the prompt “#C C ...”. During inference, we apply three spatial crops of size  $224 \times 224$  from each  $256 \times 256$  frame of 10 video clip, averaging predictions across these crops to produce the final results.

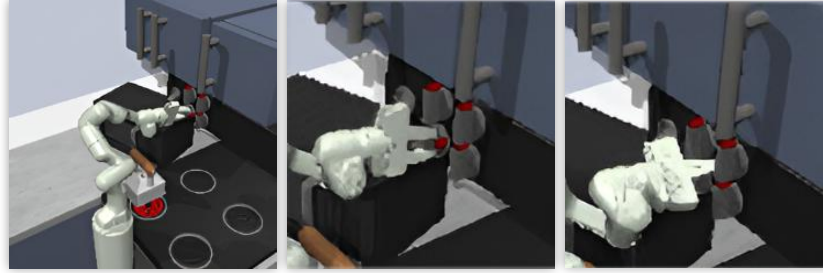
**Depth Estimation in H2O.** The frozen visual encoder produces feature maps of dimension 768, which are passed to a linear decoder to estimate depths with a resolution of  $720 \times 720$ . The model is trained for 10 epochs with a batch size of 64 and a learning rate of 0.0005, where the first 1.5 epochs serve as a warm-up phase. Our evaluation code is built upon Probing3D [4].

**Robot Manipulation in Franka Kitchen.** The robot must learn to accurately perceive and interact in 3D spaces to accomplish daily tasks. Following previous works [14, 9], we evaluate the visual representations as frozen perception modules for downstream policy learning within the Franka Kitchens simulation environment [7]. Five tasks are adopted: turn knob (TK), open door (OD), flip switch (FS), open microwave (OM), and slide door (SD). We use success rate as the evaluation metric. In Franka Kitchen environment, all baselines apply imitation learning for visuomotor control. A policy network is trained for each task using observations from the environment and video representation from EgoDTM. To adapt EgoDTM and LaViLa, we repeat the image observation 4 times as video input. For each task, the experiment is conducted from two different camera viewpoints for two random seeds using 50 randomly sampled trajectories. The final result is the average of the success rate. Our codebase is built upon MPI [9].

**Natural Language Queries on Ego4D.** The task typically operates on a 6-minute video. Using the video compression technique from [1], we compress the original video frames by 6 times to save storage. Then we extract the features with the fps of 1.87 and sampling frame number 4. We take 256 dimension global video features and 768 dimension BERT features as input. Our codebase is built upon EgoVLP [13].

Table 1: Comparison of robot manipulation tasks in Franka Kitchen simulation to assess model’s 3D-awareness.

Method	TK	OD	OM	FS	SD	Average
R3M [14]	53.3%	50.7%	59.3%	86.3%	97.7%	69.4%
MPI [9]	83.3%	54%	44.5%	93.5%	100%	75%
ResNet [8]	28%	18%	26.7%	50%	75.5%	39.7%
CLIP [16]	26.3%	13%	24.7%	41.7%	86.3%	38.4%
LaViLa [24]	48%	26%	22.5%	69%	<b>94.5%</b>	52%
EgoDTM (ours)	<b>56%</b>	<b>28%</b>	<b>35.5%</b>	<b>81%</b>	92.5%	<b>58.6%</b>



Task: Turn Knob

EgoDTM ✓

LaViLa ✗

Figure 2: Qualitative results of robot manipulations.

## B Additional Quantitative Results

**Robot manipulation.** As shown in Table 1, EgoDTM consistently outperforms pretrained visual-language models such as CLIP [16] and LaViLa [24] by +20.2% and +6.6%, respectively, demonstrating stronger spatial perception in visual representations. Additionally, EgoDTM performs competitively with specialized robot learning methods on certain tasks, such as “turn knob” but underperforms on others, like “open microwave”. A possible reason is that our model, pretrained on depth and text, may overfit to real-world scenarios, whereas methods like R3M [14] leverage self-supervised pretraining, which could provide better generalization to diverse manipulation tasks.

**Predicted depth contains valuable information for action recognition.** Since ground-truth data is unavailable for directly evaluating the depth decoder, we demonstrate the utility of our predicted depth maps for multimodal action recognition, as shown in Table 2. We simply encode the depth map using an MLP at a 56p resolution. The action recognition accuracy improved by +1.4%, confirming that our predicted depths contain meaningful information for unseen egocentric data.

Table 2: Multi-modal action recognition on EGTEA using our predicted depths. We apply a simple MLP to encode depth maps and perform multi-modal feature fusion via late fusion.

Modality	mean-acc	top1-acc	top5-acc
RGB	61.6	68.2	87.4
RGB+Depth	62.5	69.5	89.2

## C Additional Qualitative Results

**Consistent HOI Masks.** We compare inconsistent HOI bounding boxes from HOID [21] (frame-by-frame detection) with spatial-temporal HOI masks generated by combining HOID [21] and SAM2 [18] in Figure 1. The HOI boxes detected by HOI detector often lose tracks, since they are detected by an image-based model. Our pipeline achieves consistent HOI tracking across frames, offering more precise HOI labels.

**LLM Prompts.** We use the LLM prompts in Figure 3 to enrich the texts with HOI shape and movement information. Specifically, the shape information is provided by the HOI mask areas, where the large, medium, small object occupies [0.1,1], [0.01, 0.1] and [0, 0.01] areas, respectively.

132 **Generated Data.** Examples of generated HOI boxes, HOI masks, depths, and enriched texts are  
133 illustrated in Figure 4.

134 **Case Study on Robot Manipulation.** In Figure 2, the learned policy based on EgoDTM visual  
135 representation enables the robot to approach the switch and turn it, while LaViLa successfully  
136 approaches but misses the switch.

## 137 D Discussions

138 **Comparison with Related Works that Pretrained with Depth.** ImageBind [5] and Language-  
139 Bind [25] are the most relevant depth-based vision-language pretraining works. These methods  
140 employ multiple encoders to align various modalities within a unified feature space through con-  
141 trastive learning. While both methods utilize depth for pretraining, their application may be less  
142 impactful when applied to conventional third-person datasets. In contrast, depth is essential for  
143 egocentric perception, where spatial awareness is critical for understanding human indoor activities.  
144 Furthermore, their pretraining processes treat depth as an input rather than a prediction target, re-  
145 sulting in depth features that lack pixel-level 3D information and video representations that remain  
146 unaware of 3D structure. In our work, we aim to enable video representations to predict depth maps,  
147 thereby embedding 3D awareness directly into the representations.

148 **Potential for Real-World Applications.** While our model demonstrates improvements over text-  
149 pretrained models in both video understanding and robotic manipulation tasks, it falls short of  
150 state-of-the-art performance of manipulation models. However, our model predicts more meaningful  
151 depth maps in real-world settings than in simulations, offering promising potential for real-world  
152 deployment.

## Egocentric Video with HOI Masks



### System Prompt

#### ## Background

1. The user will provide information about a short egocentric video (12 frames) captured by one person using VR/AR, approximately one second long, with continuous frame box annotations. The annotations represent the center point and size of objects using the format: <length x, width y, area s, contact state (optional)>, where values range from [0, 1] indicating percentages of the length or width. The area is the product of length and width. The annotations may include up to four elements: the left hand, the right hand, an object related to the left hand, and objects being manipulated by the hands.
2. About the directions, smaller x means more left, larger x means more right, smaller y means more higher, larger y means more lower.
3. If the area is larger than 0.1, then the object is large object; if the area is larger than 0.01 but smaller than 0.1, then the object is medium size; if the area is smaller than 0.01, then the object is small object.
4. There are five possible contact states: no contact, self-contact (between the user's hands), contact with another person, contact with a portable object (e.g., an apple), or contact with a stationary object (e.g., furniture).
5. Note that I can only assure the hands, but the types of left/right can not be guaranteed. Typically, if there exists two hands, the hand with lower x is the left hand, the hand with larger x is the right hand. Another you should notice is that, if there are mostly left hand but exists few right hand data, you should ignore the right hand data, vice versa.

#### ## Response Requirements

1. You should interpret the hand-object interaction (HOI) in the video: Use the given text information to describe the interaction process, such as relevant relations, human actions and objects. Describe the hand positions, movement directions, and speed, as well as the sizes and positions of any objects.
2. Remember that the information in original text must be contained in your response.
3. Your response should be a precise, fluent and unified natural language summary, restrictly using less than two sentences. I denote C as the user, please start your response with '#C C ...'.
4. Avoid using pronouns like 'their', 'his', 'her'. Never mention 'in the video', just express what happens.
5. Don't express the same thing twice. Never use parentheses i.e., (), to explain your meaning.

I will provide you the above mentioned information. The information will keep empty if the video does not have that type of object.

### Input Prompt

Given the HOI-related information below, respond me with a new sentence following the above requirements.

**Original text:** #C C Cuts a cucumber on a chopping board with a knife.

**Possible left hand:** [].

**Possible right hand:** [[0.59, 0.92, 0.01, 'portable object contact'], [0.96, 0.96, 0.0, 'portable object contact'], [0.98, 0.88, 0.0, 'portable object contact'], [0.59, 0.92, 0.01, 'portable object contact'], [0.96, 0.96, 0.0, 'portable object contact'], [0.98, 0.87, 0.0, 'portable object contact'], [0.59, 0.93, 0.01, 'portable object contact'], [0.97, 0.95, 0.0, 'portable object contact'], [0.98, 0.86, 0.0, 'portable object contact'], [0.59, 0.93, 0.01, 'portable object contact'], [0.97, 0.94, 0.0, 'portable object contact'], [0.98, 0.84, 0.0, 'portable object contact'], [0.58, 0.93, 0.01, 'portable object contact'], [0.98, 0.93, 0.0, 'portable object contact'], [0.98, 0.82, 0.0, 'portable object contact'], [0.58, 0.93, 0.01, 'portable object contact'], [0.98, 0.92, 0.01, 'portable object contact'], [0.98, 0.81, 0.0, 'portable object contact'], [0.58, 0.93, 0.01, 'portable object contact'], [0.98, 0.91, 0.01, 'portable object contact'], [0.98, 0.8, 0.0, 'portable object contact'], [0.58, 0.93, 0.01, 'portable object contact'], [0.98, 0.91, 0.01, 'portable object contact'], [0.98, 0.8, 0.0, 'portable object contact'], [0.58, 0.93, 0.01, 'portable object contact'], [0.98, 0.9, 0.01, 'portable object contact'], [0.98, 0.78, 0.0, 'portable object contact'], [0.59, 0.93, 0.01, 'portable object contact'], [0.98, 0.89, 0.01, 'portable object contact'], [0.98, 0.78, 0.0, 'portable object contact'], [0.59, 0.93, 0.01, 'portable object contact'], [0.98, 0.88, 0.01, 'portable object contact'], [0.98, 0.77, 0.0, 'portable object contact'], [0.59, 0.93, 0.01, 'portable object contact'], [0.99, 0.88, 0.01, 'portable object contact'], [0.97, 0.76, 0.0, 'portable object contact']]].

**Hand (not sure which hand):** [].

**Object1:** [[0.45, 0.77, 0.0], [0.46, 0.78, 0.0], [0.46, 0.78, 0.0], [0.46, 0.79, 0.0], [0.47, 0.79, 0.0], [0.48, 0.8, 0.0], [0.48, 0.8, 0.0], [0.49, 0.8, 0.0], [0.49, 0.8, 0.0], [0.5, 0.8, 0.0], [0.51, 0.79, 0.0]].

**Object2:** [].

**Your response:**

### LLM Response

#C C holds a knife with the right hand, moving it downward to cut a small cucumber on a chopping board.

Figure 3: LLM prompt strategy for generating enriched text from HOI masks and the original text.



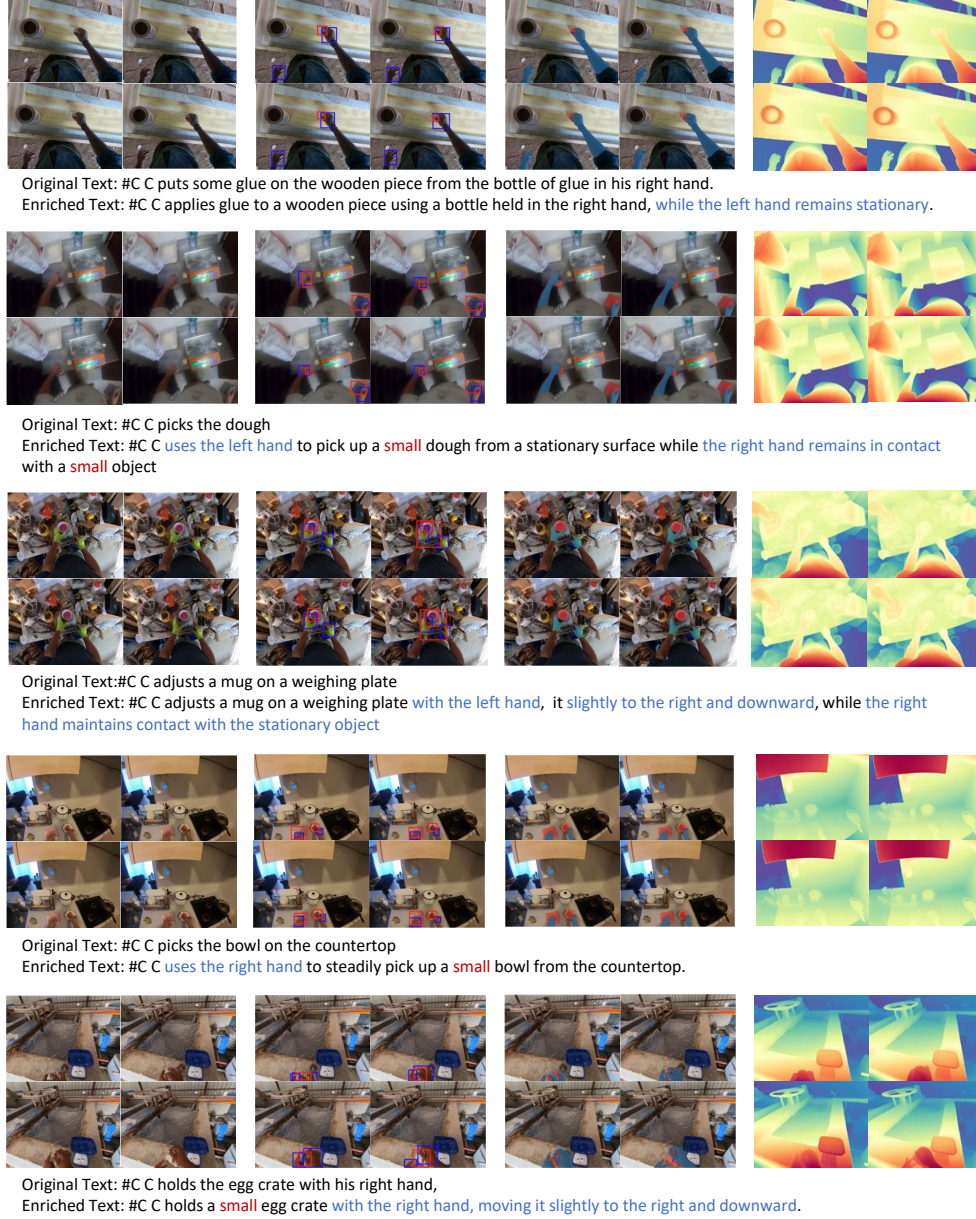


Figure 4: Illustration of data generated by our data generation pipelines, including intermediate HOI boxes, masks, and the enriched texts and depth maps used as supervision signals. The text that includes HOI movements is marked **blue**, while the contents that include HOI spatial information are marked **red**.

## References

- [1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 2
- [2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021. 2
- [3] DeepSeek-AI. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 1
- [4] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2
- [5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 4
- [6] 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. 1, 2
- [7] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long horizon tasks via imitation and reinforcement learning. *Conference on Robot Learning (CoRL)*, 2019. 2
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [9] Zeng Jia, Bu Qingwen, Wang Bangjun, Xia Wenke, Chen Li, Dong Hao, Song Haoming, Wang Dong, Hu Di, Luo Ping, Cui Heming, Zhao Bin, Li Xuelong, Qiao Yu, and Li Hongyang. Learning manipulation by predicting interaction. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 2, 3
- [10] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. In *British Machine Vision Conference (BMVC)*, 2021. 2
- [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, October 2021. 2
- [12] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 2
- [13] 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. 2
- [14] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2022. 2, 3
- [15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1
- [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

- 203 [17] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust  
204 monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on*  
205 *Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1
- 206 [18] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr,  
207 Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas  
208 Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment  
209 anything in images and videos. In *The Thirteenth International Conference on Learning Representations*,  
210 2025. 1, 3
- 211 [19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object  
212 detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*,  
213 39(6):1137–1149, 2016. 1
- 214 [20] Fiona Ryan, Hao Jiang, Abhinav Shukla, James M. Rehg, and Vamsi Krishna Ithapu. Egocentric auditory  
215 attention localization in conversations. *2023 IEEE/CVF Conference on Computer Vision and Pattern*  
216 *Recognition (CVPR)*, pages 14663–14674, 2023. 2
- 217 [21] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at  
218 internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,  
219 pages 9869–9878, 2020. 1, 3
- 220 [22] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao.  
221 Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1
- 222 [23] Yue Zhao and Philipp Krähenbühl. Training a large video model on a single machine in a day. *arXiv*  
223 *preprint arXiv:2309.16669*, 2023. 2
- 224 [24] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from  
225 large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*  
226 *Recognition*, pages 6586–6597, 2023. 1, 2, 3
- 227 [25] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu  
228 Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-  
229 language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International*  
230 *Conference on Learning Representations*, 2024. 4