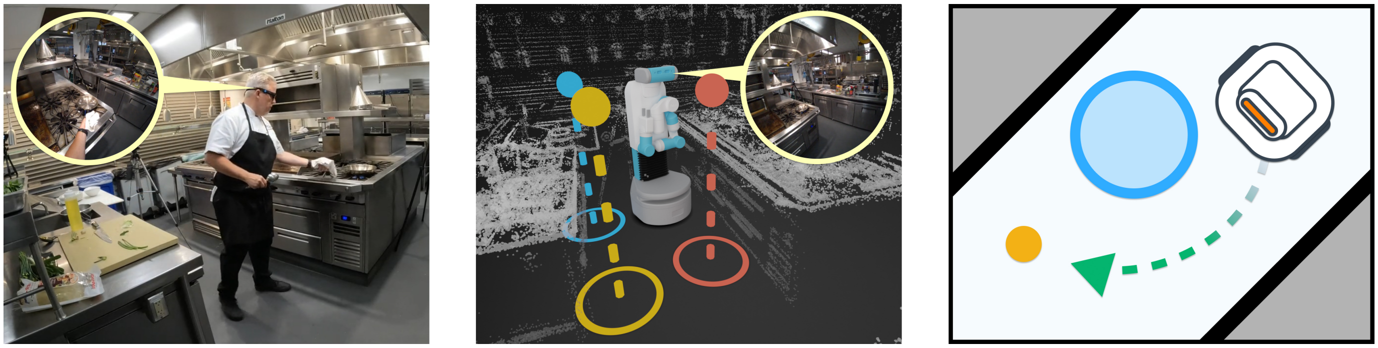
Spatial Affordance Prediction for Egocentric Task-Driven Navigation

Author Names Omitted for Anonymous Review.



(a) Ego-video Demonstration **(b)** Robot Task Inference **(c)** Affordance-aware Navigation

Fig. 1: **(a)** First-person video of human demonstrations is used to train a network to predict spatial affordances, which we define as locations where a task may take place in the current environment. **(b)** When a robot is navigating in the same environment, it uses these predictions to find likely locations for various tasks. **(c)** The robot can then navigate to one of these locations to accomplish a given task (orange dot), or avoid a location to give space for humans engaged in this task (blue circle) depending on the situation.

***Abstract*—We investigate the problem of spatial affordance prediction for egocentric task-driven navigation, that is, predicting locations in a environment where a given task is likely performed, using a single egocentric image and a natural language task query. Our end-to-end model encodes environment context and task semantics by fine-tuning a vision-language framework trained on egocentric human demonstrations from large-scale cooking activity videos. The resulting model outputs spatial regions representing task affordances relative to the egocentric camera pose. The resulting predictions outperform a nearest-neighbor baseline based on pretrained vision-language similarity, particularly on novel tasks and viewpoints. We incorporate these spatial affordance predictions for two robotic navigation applications: one, localizing goals for task completion, and two, defining task-based obstacles to avoid disturbing humans in a shared environment.**

I. Introduction

Egocentric video offers a unique view of human activities, capturing the tasks of daily life directly from an individual's perspective. This first-person viewpoint naturally aligns with the perspective of a robot, suggesting that the richness of human demonstration can be leveraged for various robotic tasks. To this end, large ego-video datasets of human activities have been released [3, 7, 14, 15], helping drive research in areas of human activity understanding and robotics. Egocentric data is especially valuable for applications to navigation, as the first-person camera does not restrict the camera-wearer's motion, allowing the capture of a person's natural movement around their environment.

Existing foundation models often used in robot perception, such as Vision-Language Models (VLMs), typically focus on representing the visible features of a scene, supporting queries relating to scene contents or object localization. While this aspect of VLMs is important for robotics, it's equally important to understand the meaning of the objects and space within the view; a VLM for robotics should not only understand "this is what I see," but also "what can be done here and there". Inspired by the need for this capability, we introduce the problem of *spatial affordance prediction* – identifying the locations in an environment where a given task can be completed.

We conceptualize this problem as first understanding the scene context from an ego-image, and then combining this context with a given task in order to predict the likely region where a person performing the task may be. We propose a deep end-to-end approach which solves both problems simultaneously. The resulting network is trained on a large set of tasks from a variety of cooking activities and kitchen environments and is able to predict new spatial task affordances given natural language descriptions.

In this work, we show how to incorporate spatial affordance prediction to improve robot navigation through two practical use cases:

- **Goal Localization**. Given a task specified in natural language (and the current egocentric robot perspective) compute the location in the environment where the task is likely to be completed.
- **Task Obstacles**. Given a set of tasks specified in natural language (and the current egocentric robot perspective) compute the locations a robot should avoid, to avoid

disturbing humans in potentially busy areas.

When deployed to new tasks from unseen views in known environments, our resulting system outperforms a non-parametric baseline. Because our system localizes tasks within a known environment, we propose its use in conjunction with motion planning strategies such as RRTs [20] or point navigation [44] for robot deployment.

II. Related Work

Deep learning has proven to be a powerful paradigm for understanding scene geometry from images, both in multi-image scene reconstruction as seen in NeRFs [29], and single-frame third-person body pose prediction [5], first-person navigation [34, 37, 33], and first-person body pose prediction [42] tasks. Beyond geometry, semantic reasoning through natural language over images has recently been enabled via Vision-Language Models (VLM) such as CLIP [38], BLIP [22], and EgoVLP [23]. However, these models on their own have limited spatial understanding [11].

Egocentric vision is a common representation for robots due to the prevalence of on-board cameras. As such, methods have been developed to leverage egocentric data for robotic tasks such as understanding activities [24, 12], shaping behavior [30], inferring goal locations [8], and understanding manipulation affordances [16]. To support these applications, specialized large-scale datasets of egocentric human demonstrations have been proposed, such as the Ego4D dataset [14], and the EgoExo4D dataset [15].

Recent work seeks to align geometry and semantics to enable robust navigation of mobile agents. Reinforcement learning approaches seek to understand how to reason about the environment given a pre-defined task from a robot's perspective [44, 28, 17, 35, 13, 21, 40]. CLIP has been integrated into mobile robot policies to allow natural language task augmentation [26, 10, 40]. VLMs have also been used to create flexible semantic maps a mobile robot can query using natural language, such as VLMaps [19], NLMap-SayCan [6], CLIP-Fields [39], and 3D-LLMs [18], using e.g. an RRT [20]. Particularly relevant to our proposed work are Vision-Language Navigation approaches [1, 36, 41]. These models seek enable language-conditioned robot navigation in real environments, but unlike our approach, focus on unknown environments where access to a navigational map is infeasible.

A closely related problem to spatial affordances (where a person stands for a task) is manipulation affordances (how to manipulate an object for a task). Manipulation affordances can be estimated from image segmentation [4, 9], from 3D object or scene representations [43, 27], or learned end-to-end [25]. Affordances can also be learned from human demonstration as in the Vision-Robotics Bridge [2] and its text-based extension [46] which learn to represent image-based affordances from egocentric human demonstrations, where affordance is defined as contact points and trajectories for robots to interact. R3M [31] uses egocentric human demonstrations to create a semantic representation well-suited as a foundational model for downstream robot tasks.

III. Affordance Localization

Given a task, our aim is to predict the locations where a person would be to execute the task. We refer to these locations as the environment's "spatial affordance" for the task.

To identify spatial affordances, we assume human demonstrations define a true per-task distribution $\mathcal{D}$, the region where a person performed a task, within an environment $\mathcal{E}$. These distributions can be extracted from an ego-video demonstration $\mathbf{V}$, defined by camera poses $\mathbf{X}$ under tasks $\mathcal{T}$. Given a first-person image $\mathbf{I}$ within $\mathcal{E}$ and a natural language task query $\mathbf{q}$, our goal is to predict the relative task distribution $\widetilde{\mathcal{D}}$, such that

$$\widetilde{\mathcal{D}}(\mathbf{q}, \mathbf{I}) = T(\mathcal{D}(\mathbf{q}, \mathcal{E})). \tag{1}$$

Importantly, because the image $\mathbf{I}$ is egocentric, each image carries with it an implied location within the camera's environment, explicitly represented by the transform T.

A. Model Architecture

We model the affordance prediction task with an encoder-decoder style deep neural network architecture, first encoding the egocentric image and task query as vectors $\mathbf{e}_V$ and $\mathbf{e}_L$ respectively, which individually capture the environmental and task semantics, which are then decoded into the task's location using a decoder $\mathbf{A}$.

Environmental Context and Image Localization To encode the egocentric image $\mathbf{I}$ at the robot's current viewpoint, we can use pre-trained, foundational image models that have demonstrated a strong ability to capture the image's semantic information, such as CLIP [38]. However, such image encoding models typically capture the semantics of what is being viewed in the image rather than encoding the spatial affordance the image suggests. To address this, we finetine a pre-trained image encoder E_V, on ego-video demonstrations, intending to capture the relative spatial affordance given the image through the envoding vector $e_{\mathbf{v}} = E_V(\mathbf{I})$.

Task Encoding Unlike images, which need additional learned context, tasks described in natural language can be encoded directly with pretrained language models such as CLIP [38]. A task query is tokenized then passed to a pre-trained language encoder E_L, to obtain the encoding vector $e_{\boldsymbol{\ell}} = E_L(\mathbf{q})$. Unlike E_V, E_L is frozen during training, as learning the context on just the image information allows the network to learn environmental context separate from downstream language task queries.

Affordance Prediction Because a person may naturally move around as they accomplish a given task, each task may have a small range of positions where it was seen accomplished. We therefore represent the observed distribution of a task as being a normal distribution:

$$\mathcal{N}_{\mathbf{t}}(\mu_{\mathbf{t}}, \Sigma_{\mathbf{t}}) \approx \mathcal{D}(\mathbf{t}) \tag{2}$$

capturing the likely location for a person for that task across all the frames the task occurs.

The encoding vectors $e_{\mathbf{v}}$ and e_{ℓ} represent what is expected to be around the viewer, and what the goal task is, respectively.

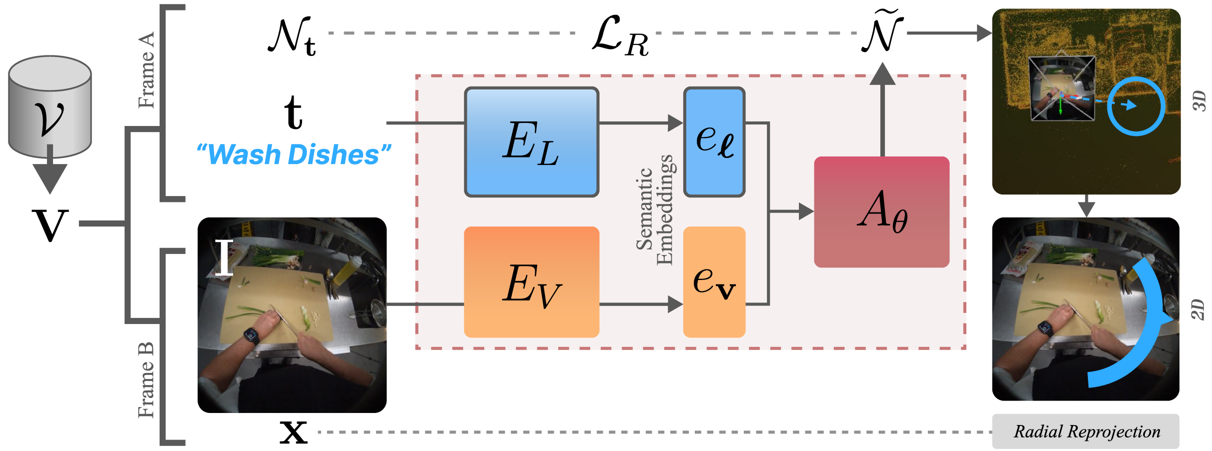


Fig. 2: **Model Architecture.** Given video demonstration, $\mathbf{V}$, of an activity containing several tasks, our model is trained over pairs of tasks and images selected from different times in the video. For example, a task at frame "A" is encoded via a (frozen) pretrained language model E_L and combined with an encoding of an image from frame "B". Images are encoded with a pretrained vision model E_V (unfrozen). This pair of encodings is finally passed to an affordance network A_θ which predicts an region where task "B" should take place relative to frame "A". The loss function $\mathcal{L}_R$ rectifies this position and compares it to the ground truth global position from task "A".

Taken together, this should provide sufficient information for spatial affordance prediction within the environment. An affordance prediction network, A_θ, is trained which takes as input these encoding vectors and produces a final 3D task region:

$$A_\theta(e_\mathbf{v}, e_\ell) := \widetilde{\mathcal{N}}(\mu_\theta, \Sigma_\theta) \tag{3}$$

whose mean is a 3D position and with 2D uncertainty constrained to lie along the ground plane with zero covariance (isotropic).

B. Loss Function

We can directly optimize Equation 1 by minimizing the difference between the task's predicted distribution and the target distribution. To ensure the predicted regions $\widetilde{N}$ are metrically meaningful, we use the Fréchet Distance, d_F. Because affordance predictions happen in an egocentric frame, the target task region must be rectified before the Fréchet Distance can be computed. We align the target task in the coordinate frame of the query image through the transform $R_\mathbf{x}$, and compute the error over all image-task pairs as follows:

$$\mathcal{L}_R = \sum_{\mathbf{x},\mathbf{I}\in\mathbf{V}} \sum_{\mathbf{t}\in\mathcal{T}} d_F(R_\mathbf{x}(\mathcal{N}_\mathbf{t}), \widetilde{\mathcal{N}}), \tag{4}$$

computed over all videos $\mathcal{V}$. The training scheme is shown alongside the architecture in Figure 2.

C. Training

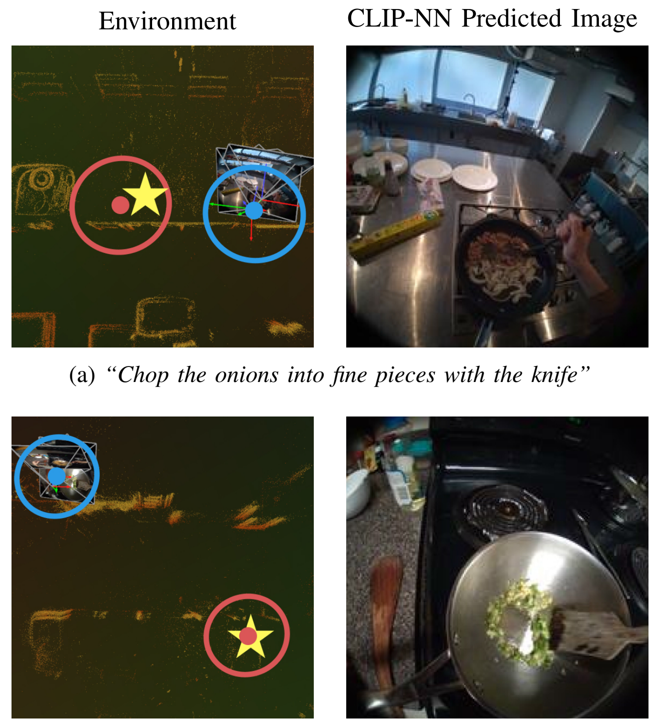
We curated a dataset consisting of egocentric videos of people demonstrating cooking tasks from the EgoExo4D dataset [15], where each task is a keystep from a larger cooking activity. For example, the activity "Making Noodles" includes tasks such as "Wipe hands with a kitchen towel" and "Add soy sauce to the noodles in the skillet." An LLM (GPT-4 [32]) was used during training to augment each task description with several rephrasings which preserve the meaning of the original task. When computing keysteps for training we only consider frames where the camera has a velocity below 0.1 m/s. To stabilize our predictions in our egocentric coordinate frame, we also correct for pitch and roll of the camera.

For the pretrained vision and language encoding networks, E_V and E_L, we used pretrained CLIP [38] as it has been shown successful in a wide variety of language tasks. The affordance predictor network A_θ is a 4-layer MLP with 1M trainable parameters, each with layer normalization.

We randomly split the dataset into training and testing tasks (80%/20%), and a training and testing image set (consecutive 10% held out), and train on a single V100 GPU and 10 CPU cores. Our base model was trained for 150 epochs over 7 hours of training.

D. Non-parametric Baseline

An alternative to our proposed approach is to represent the entire scene either directly by retaining all images, or through a learned field representation. However, without additional training to consider affordances, these methods will still have limitations despite their larger data requirements. As a representative baseline, we introduce a nearest-neighbor based approach which leverages pretrained CLIP [38] as a task-similarity measure that can be applied over all images captured per environment in the dataset. This baseline approach, referred to as CLIP-NN, takes a CLIP text encoding of the task description $\mathbf{q}$, and a CLIP image encoding of every image in the demonstration $\mathbf{V}$. We can predict the best fitting image as the frame c for which the cosine encoding similarity between the egocentric image and the task query is maximized. The



(a) *"Chop the onions into fine pieces with the knife"*

(b) *"Acquire the scallion from the countertop"*

Fig. 3: Affordance prediction from the baseline approach (blue) and our proposed approach (red) on two task queries. **Left:** The spatial predictions along with the ground truth as a yellow star. On both queries, our proposed method predicts the correct locations while the baseline predicts a pose far away. **Right:** The image/location pair with the highest CLIP similarity. These capture elements of the scene (e.g., onions or scallions) but do not capture the full task and miss the context of the cutting board and countertop, which are needed for the actions of 'chopping' or 'acquiring'.

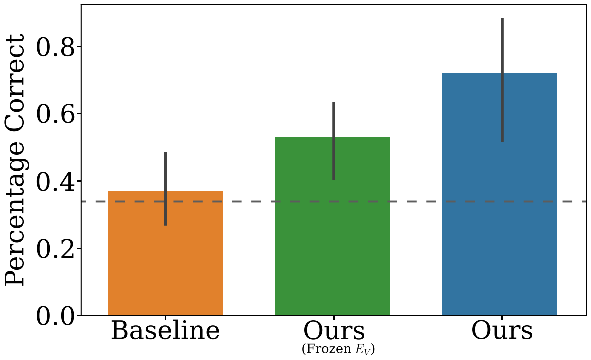


Fig. 4: **Affordance grounding.** When predicting from which in a set of three tasks is the most likely for a given image, the baseline (orange) performs similarly to random guessing (dashed line). Our model with a frozen language encoder (green) significantly outperforms the baseline, and our model with an unfrozen image encoder (blue) nearly doubles the baseline.

task position prediction is then $\mathbf{x}_c$, the corresponding position of the camera-wearer at time c. That is, we predict the location where the view best matches the task as evaluated by the CLIP encoding similarity. To compute the region uncertainty, we compute per-task uncertainty from all task positions, and average over all tasks in $\mathbf{V}$.

An immediate limitation of the baseline, and similar approaches based directly on CLIP descriptions, is that CLIP only captures the content of the image itself, rather than information about the kinds of tasks and activities that the scene affords, as shown in Figure 3.

This affordance grounding capability can be directly measured through a multiple-choice paradigm, where the model is used to predict which of three randomly selected task queries is most likely to take place at a given image, either the highest CLIP similarity for the baseline, or the lowest predicted distance for our method. The CLIP-NN baseline only does slightly better than random guessing (37%), while our model has nearly double the performance of the baseline (63%) as seen in Figure 4. We hypothesize this is due to CLIP encodings capturing the content of the image, rather than the activities afforded by the scene viewed from the image. Our model's ability to capture affordances comes in part by fine-tuning the vision encoder E_V. Even with a frozen E_V (unmodified CLIP), the model still outperforms the baseline, but by a less significant margin.

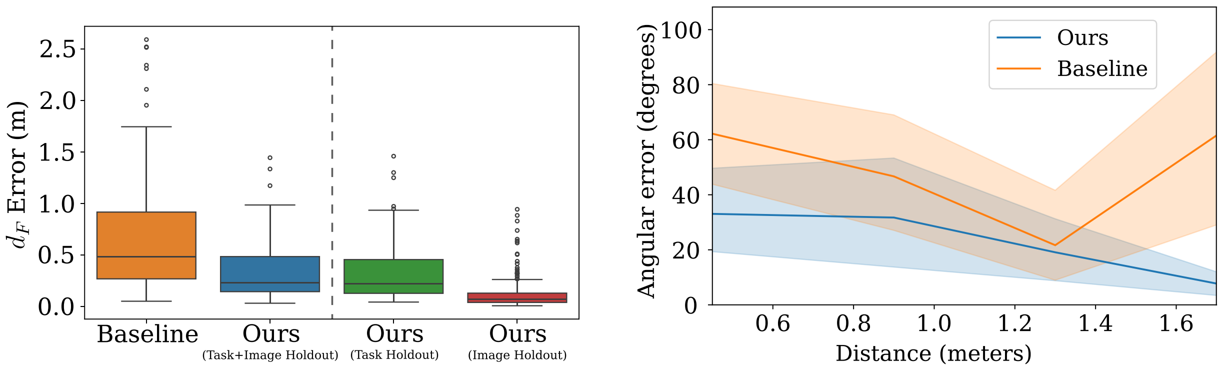
IV. Robot Applications

A. Navigation

Task localization ability is directly required by a home assistive robot to accomplish natural language directions such as "turn on the stove". When compared to the baseline, our approach is significantly more accurate at predicting where a given task will take place relative to an arbitrary egocentric viewpoint. Our approach shows statistically significant gain over the baseline [t(82) = 4.683, p <0.001] (Figure 5a left of dashed line) even when testing on both unseen tasks from held-out viewpoints. The right side Figure 5a shows two additional breakdowns of the task localization results tested on either only known images or known tasks. When tested on seen images and unseen tasks, the performance is nearly the same. When tested on seen tasks and unseen images, our model has almost no error.

In many cases, it is not possible to accurately predict where a task may place from a single viewpoint, especially if the task happens far away from the robot and out of its view. However, in these cases, it's often possible to establish a reasonable guess of what general direction a task is relative to the viewer. Then, as a robot moves towards the predicted direction, it can refine its estimate of the task location. Figure 5b captures this angular error. Unlike the baseline, our approach has the lowest angular error for far away tasks, highlighting its utility in egocentric navigation.

We demonstrate applications to navigation in a custom simulator to allow a robot to use testing images to navigate



(a) Task Localization (Frechet Distance)

(b) Task Localization (Direction)

Fig. 5: Results for task localization. **(a)** When predicting which location a given task may take place at, the baseline approach does well in some cases (median error of 0.48m) but has many cases with high error or significant outliers. Our approach has much lower error when testing on unseen images (red), unseen tasks (green), or both unseen images and tasks (blue). **(b)** When predicting task direction from current viewpoint, our approach is able to better estimate the angle for far-away tasks compared to baseline.

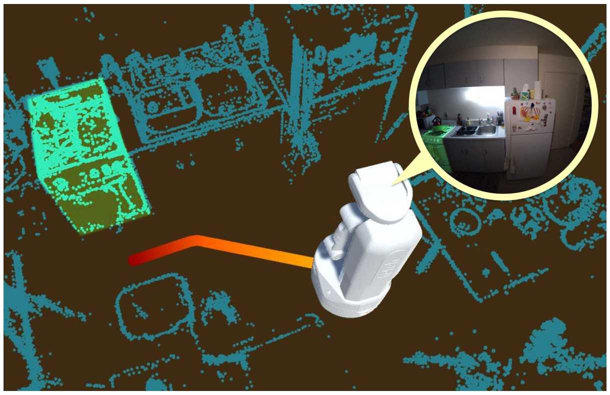


Fig. 6: Trajectory to "Heat the Food" (stove highlighted).

to positions appropriate for new tasks unseen in training. We based the simulation on the Fetch robot [45] as it has similar physical affordance to humans. An example of navigation is shown in Figure 6, where a robot is given a novel view (shown in the inset bubble) and asked to navigate to the task "*Heat the Food*". Given this egocentric view, the robot is able to predict the tasks' relative location. A navigation mesh of estimated free space is used to avoid collision during motion.

B. Collision Avoidance

In shared robot-human environments, it can be important for a robot to proactively avoid regions where a person may need to be be while doing a set of tasks. We can use the trained model to define a *Task Obstacle* covering a set of locations a robot should avoid while a person is doing a set of related tasks as detailed in Algorithm 1. For a given set of tasks a person may do, we first bound a safety radius of σ_{bound} standard deviations around the predicted task regions and then encompass the entire set of bounded regions by their convex hull. The resulting task obstacle contains both the likely regions a person would be in during tasks and the areas they will likely travel between tasks, allowing a robot to plan accordingly.

Algorithm 1: Task Obstacle Generation

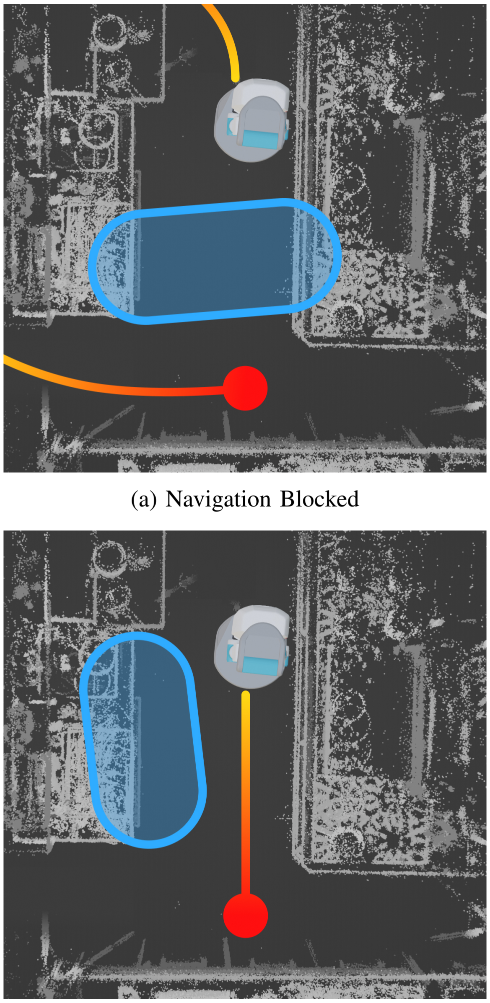
```
Load $\mathbf{A} \coloneqq E_V,\ E_L,\ A_\theta,$
Given $\mathbf{I}_{\text{current}},\ \mathcal{T}_{\text{set}},\ \sigma_{\text{bound}}$
distributions $=$ $\mathbf{A}(\mathcal{T}_{\text{set}}, \mathbf{I}_{\text{current}})$
regions $=$ [ region($\mathcal{D}$, $\sigma_{\text{bound}}$) for $\mathcal{D}$ in distributions ]
points $=$ [ discretize($\mathbf{r}$) for $\mathbf{r}$ in regions ]
task_obstacle $=$ convex_hull(points)
```

Figure 7b shows two examples of task obstacle-aware navigation. In the first example, a task obstacle is comprised of the two tasks of "slice tofu" and "cook meat" which the robot expects a person to be completing. The resulting task obstacles spans the kitchen galley, blocking the direct path between the robot and its goal task of "get soy sauce", requiring the robot to take a less-direct path. In the second example, the task obstacle is comprised of "slice tofu" and "wash dishes". Here, the expected motion of a person hugs the countertop closely, allowing the robot to freely pass by without interfering.

V. Conclusion

We presented a framework to predict spatial affordances of where people perform tasks within a robot's environment, and demonstrated it's applications in robot navigation tasks. Our system is trained on egocentric video demonstrations and shows generalizability to new tasks (not seen in training) described in natural language.

Limitations Though our approach shows generalization to new tasks and novel viewpoints, this generalization is limited to scenes very similar to those seen at train time. This limitation could be alleviated via online learning where the model is continuously updated based on live observations. Likewise, the affordances from human demonstrations may not map one-



(a) Navigation Blocked

(b) Navigation Unblocked

Fig. 7: Example of task obstacles for navigation. In (a), two tasks defining the task obstacle span across the workspace, representing an expectation of a busy area, blocking the robot's direct path to its goal (red dot). In (b), predicted task locations span along the countertop, allowing the robot to pass freely without disturbing the person in that area.

to-one with various types of robots, and online learning or other approaches could be used to adapt between the robot and the demonstrations. Another important limitation of our work is that all examples were taken from cooking activities in kitchens, and more environments should be considered. Lastly, we currently assume each task region is approximated by a unimodal distribution. In the future, we would like to explore alternative forms of spatial affordance prediction, for example predicting heatmaps, or full-body poses.

Future Work As a point of future research, we would like to explore the video input, as opposed to static images, as video input may stabilize predictions over time. We would also like to explore the use of spatial affordances directly within a robot policy as seen in vision-language navigation tasks. Lastly, we wish to explicitly ground the predictions in the environmental geometry via training loss function, which may help give more accurate predictions.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Introducing hot3d: An egocentric dataset for 3d hand and object tracking, 2024.

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[5] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, 2020.

[6] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *arXiv preprint arXiv:2209.09874*, 2022.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022.

[8] Samyak Datta, Oleksandr Maksymets, Judy Hoffman, Stefan Lee, Dhruv Batra, and Devi Parikh. Integrating egocentric localization for more realistic point-goal navigation agents. *ArXiv*, abs/2009.03231, 2020.

[9] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *International Conference on Robotics and Automation (ICRA)*, 2018.

[10] Vishnu Sashank Dorbala, Gunnar A. Sigurdsson, Robinson Piramuthu, Jesse Thomason, and Gaurav S. Sukhatme. Clip-nav: Using clip for zero-shot vision-and-language navigation. *ArXiv*, abs/2211.16649, 2022.

[11] 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 *CVPR*, 2024.

[12] 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.

[13] Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 8(79): eadf6991, 2023. doi: 10.1126/scirobotics.adf6991.

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, June 2022.

[15] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2023.

[16] Marvin Heidinger, Snehal Jauhri, Vignesh Prasad, and Georgia Chalvatzaki. 2handedafforder: Learning precise actionable bimanual affordances from human videos, 2025.

[17] David Hoeller, Lorenz Wellhausen, Farbod Farshidian, and Marco Hutter. Learning a state representation and navigation in cluttered and dynamic environments. *IEEE Robotics and Automation Letters*, 6:5081–5088, 2021.

[18] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 20482–20494. Curran Associates, Inc., 2023.

[19] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[20] Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning. *The international journal of robotics research*, 30(7):846–894, 2011.

[21] Ashish Kumar, Saurabh Gupta, and Jitendra Malik. Learning navigation subroutines from egocentric videos. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 617–626. PMLR, 30 Oct–01 Nov

2020.

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[23] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z. XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, WANG HongFa, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7575–7586. Curran Associates, Inc., 2022.

[24] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M. Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 621–638, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19778-9.

[25] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv: Arxiv-2310.12931*, 2023.

[26] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[27] Abhijit Makhal and Alex K. Goins. Reuleaux: Robot base placement by reachability analysis. *2018 Second IEEE International Conference on Robotic Computing (IRC)*, pages 137–142, 2017.

[28] Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech Galuba, Stefan Lee, and Dhruv Batra. Thda: Treasure hunt data augmentation for semantic navigation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15354–15363, 2021.

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, dec 2021. ISSN 0001-0782. doi: 10.1145/3503250.

[30] Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. *ArXiv*, abs/2110.07692, 2021.

[31] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[32] OpenAI. Gpt-4 technical report, 2023.

[33] Boxiao Pan, Bokui Shen, Davis Rempe, Despoina Paschalidou, Kaichun Mo, Yanchao Yang, and Leonidas J. Guibas. Copilot: Human collision prediction and localization from multi-view egocentric videos. *ArXiv*, abs/2210.01781, 2022.

[34] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] Ruslan Partsey, Erik Wijmans, Naoki Yokoyama, Oles Dobosevych, Dhruv Batra, and Oleksandr Maksymets. Is mapping necessary for realistic pointgoal navigation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17232–17241, June 2022.

[36] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[37] Jianing Qiu, Lipeng Chen, Xiao Gu, Frank P.-W. Lo, Ya-Yen Tsai, Jiankai Sun, Jiaqi Liu, and Benny P. L. Lo. Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion. *IEEE Robotics and Automation Letters*, 7:8799–8806, 2021.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[39] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory, 2023.

[40] Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. LM-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *6th Annual Conference on Robot Learning*, 2022.

[41] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR, 2023.

[42] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. *CVPR*, 2023.

[43] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. *ICRA 2022*, 2022.

[44] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *ICLR*, 2020.

[45] Melonee Wise, Michael Ferguson, Derek King, Eric

Diehr, and David Dymesich. Fetch and freight: Standard platforms for service robot applications. In *Workshop on autonomous mobile service robots*, pages 1–6, 2016.

[46] Tomoya Yoshida, Shuhei Kurita, Taichi Nishimura, and Shinsuke Mori. Text-driven affordance learning from egocentric vision, 2024.