Simultaneous Localization and Affordance Prediction of Tasks from Egocentric Video

Zach Chavis[1], Hyun Soo Park[1], and Stephen J. Guy[1]

https://appliedmotionlab.github.io/slap

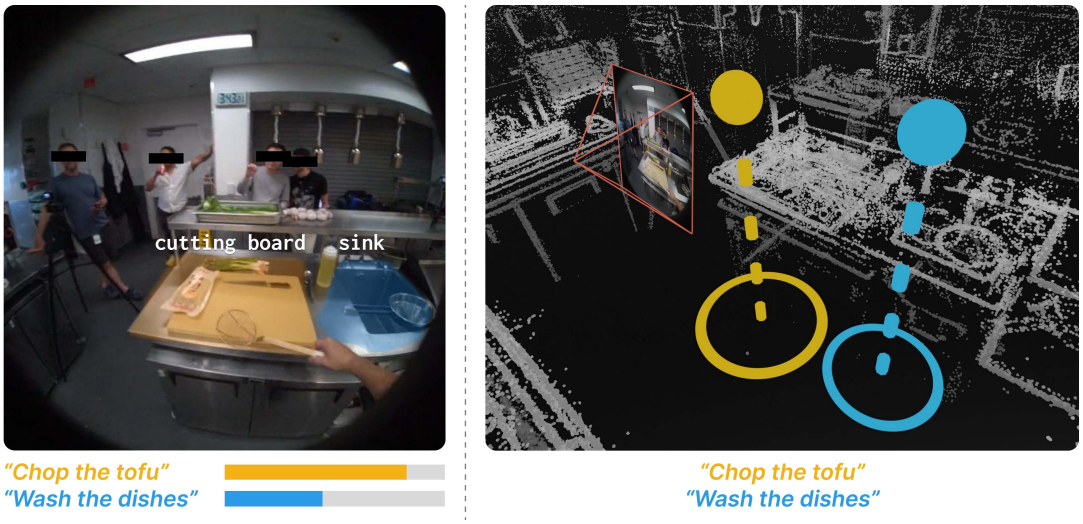

Fig. 1: **Predicting Spatial Task Affordance. (left)** Existing vision and language models (VLMs) are able to localize seen objects related to tasks within an image (colored regions) or reason about tasks through vision-language similarity (colored bars). However, existing VLMs are unable to predict into the 3D space outside of the given image itself. **(right)** We propose augmenting VLMs to make spatial predictions for where given tasks likely take place relative to an egocentric image. We refer to this region of a task's likely locations as the task's spatial affordance.

***Abstract*— Vision-Language Models (VLMs) have shown great success as foundational models for downstream vision and natural language applications in a variety of domains. However, these models are limited to reasoning over objects and actions currently visible on the image plane. We present a spatial extension to the VLM, which leverages spatially-localized egocentric video demonstrations to augment VLMs in two ways — through understanding *spatial task-affordances*, i.e. where an agent must be for the task to physically take place, and the localization of that task relative to the egocentric viewer. We show our approach outperforms the baseline of using a VLM to map similarity of a task's description over a set of location-tagged images. Our approach has less error both on predicting where a task may take place and on predicting what tasks are likely to happen at the current location. The resulting representation will enable robots to use egocentric sensing to navigate to, or around, physical regions of interest for novel tasks specified in natural language.**

I. INTRODUCTION

Understanding spatial affordances, i.e., the region of space in which a task can be accomplished in an environment, is a vital capability in any robotic or AI system that seeks to model or imitate how humans use the environment around them. Such affordances may be naturally learned from human demonstrations. Egocentric video demonstrations, i.e. first-person video captured from a head-mounted camera, is especially well-suited for learning spatial affordances as it simultaneously captures where a person is going and what they are seeing and using as they move through their environment [1]. In particular, recent large data collection efforts such as Ego4D [2] and EgoExo4D [3] provide high quality egocentric data with each frame localized in space together with annotations capturing narrative descriptions of the tasks being accomplished at each stage in the video.

Recent work has proposed systems that can provide high-quality reasoning about object's affordances, that is how objects can be used for tasks. For example, CLIP-Fields [4] allow robots to reason over semantic maps to find 3D locations of objects for tasks such as identifying the location of a microwave when given the task "warm up my lunch." However, these kinds of systems rely on access to a full 3D model at inference time (e.g., through a NeRF [5] or 3D point cloud).

Here, we define a task's spatial affordance as the area in free space where a person would stand in order to perform the task. This type of knowledge is important for robots in human environments, as it will help them better understand where people will likely be as they are doing tasks. We consider the problem of predicting 3D regions of spatial affordances from a single, egocentric image.

[1]Department of Computer Science and Engineering (CS&E), University of Minnesota, Minneapolis, MN 55414 US. (chavi014|hspark|sjguy)@umn.edu

We conceptualize this problem as first understanding the scene context from an image, and then combining this context with a given task in order to predict the likely region where a person may be. We propose a neural-network based approach which solves both problems simultaneously with an encoder-decoder style architecture. 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.

Our problem is closely related to the use of Vision-Language Models (VLMs) in robotics. Figure 1(left), outlines common uses of VLMs in robotics, such as segmenting objects a robot may interact with for its task [6], [7], or measuring the similarity of a current ego image with a task the robot is interested in [8], [9]. In contrast, our proposed framework provides a new capability: Given a single egocentric image, rather than identifying items or measuring similarities, our model produces spatial 3D regions of task's location Figure 1(right).

When deployed to new tasks and views in known environments (seen in training), our resulting system outperforms baselines, even when baselines are provided with many images (entire demonstration) at inference time rather than the single image used in our approach. We build on the proposed spatial task affordance predictions to introduce the concept of a task obstacle. These are regions over sets of related tasks which can be used to guide robot avoidance in human environments.

In summary, our main contributions are:

- An extension of VLMs to predict 3D spatial regions representing task location likelihood.
- Using the EgoExo4D dataset to learn spatial affordances from egocentric human demonstrations in real kitchen environments.
- A training approach which allows for optimizing a model on spatial demonstrations of tasks on views across the full environment.
- Task obstacles to enable robots to avoid potential collisions in human environments.

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 [5], and single-frame third-person body pose prediction [10], first-person navigation [11]–[13], and first-person body pose prediction [14] tasks. Beyond geometry, semantic reasoning through natural language over images has recently been enabled via Vision-Language Models (VLM) such as CLIP [8], BLIP [15], and EgoVLP [16]. However, these models on their own have limited spatial understanding [17].

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 identifying activities [18], shaping behavior [19], and inferring goal locations [20]. To support these applications, specialized large-scale datasets of egocentric human demonstrations have been proposed, such as the Ego4D dataset [2], and the EgoExo4D dataset [3].

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 [21]–[27]. Vision-Language Navigation approaches [28] seek enable robot navigation in human environments, but focus on objects as opposed to tasks. CLIP has been integrated into mobile robot policies to allow natural language task augmentation [27], [29], [30]. VLMs have also been used to create flexible semantic maps a mobile robot can query using natural language, such as VLMaps [31], NLMap-SayCan [32], CLIP-Fields [4], and 3D-LLMs [33], using e.g. an RRT [34].

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 [6], [35], from 3D object or scene representations [36], [37], or learned end-to-end [38]. Affordances can also be learned from human demonstration as in the Vision-Robotics Bridge [39] and its text-based extension [40] 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 [9] uses egocentric human demonstrations to create a semantic representation well-suited as a foundational model for downstream robot tasks.

III. SIMULTANEOUS LOCALIZATION AND AFFORDANCE PREDICTION

Given an egocentric image and a natural language distribution of a task (e.g., "turn on the stove"), our goal is to predict the region where someone would likely be when performing this task. We assume that tasks, images, and viewpoints are new and unseen during training. However, the environment is expected to either be seen in training or adapted to via fine tuning (Sec V-D). We conceptualize this task affordance prediction as two related aims: first, hypothesize the environmental context induced by the image, and second, predict the region in which someone may go within this hypothesized environment to perform the given task. We refer to this region of free space where a task should take place as the task's "spatial affordance," and the end-to-end prediction of a task's relative region given a single egocentric image as simultaneous localization and affordance prediction.

A. Problem Formulation

Formally, given a first-person image $\mathbf{I}$ and a natural language task query $\mathbf{q}$, we would like to predict a distribution of positions where the task is performed, $\widetilde{\mathcal{D}}$, that matches the true human distribution, $\mathcal{D}$, in environment $\mathcal{E}$:

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

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.

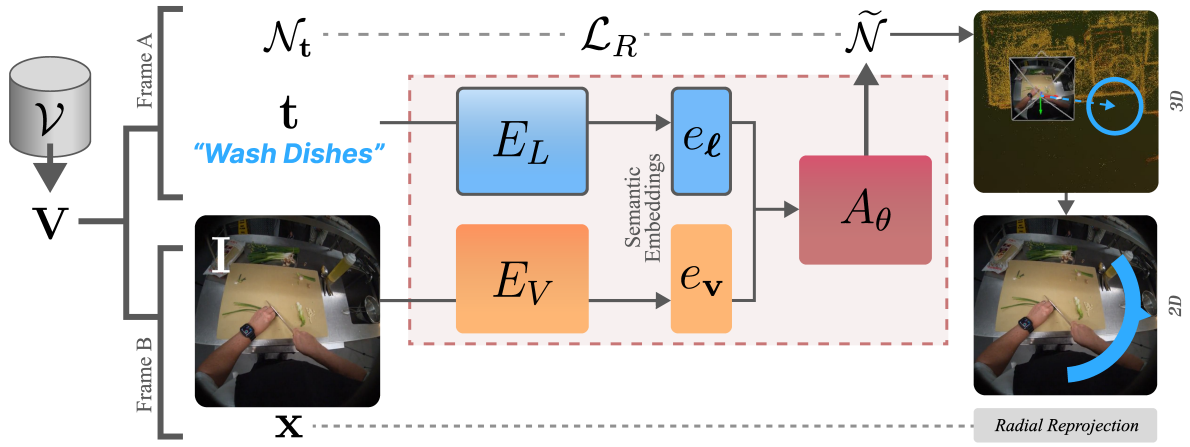

Fig. 2: **Model Architecture.** Given video demonstration, $\mathbf{V}$, of an activity containing several tasks, our model is trained over pairs of tasks and images selected from different times in the video. For example, a task at frame "A" is encoded via a (frozen) pretrained language model E_L and 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".

Data Assumptions We assume we have a dataset of videos $\mathcal{V}$ over environments $\mathcal{E}$; each video comprised of a collection of images $\mathcal{I}$, with corresponding tasks $\mathcal{T}$, and the corresponding pose where each image was observed $\mathcal{X}$. We refer to the associated collection as an annotated, localized egocentric video $\mathbf{V} \in \mathcal{V}$. Formally, $\mathbf{V} \equiv (\mathcal{X}, \mathcal{T}, \mathcal{I})$, where elements of each set $\mathcal{X}$, $\mathcal{I}$, $\mathcal{T}$ are indexed by a frame i linking the three sets together in time. In practice, such a dataset could be defined with narrated demonstrations from a subject wearing a first person camera with $\mathcal{I}$ consisting of images from the camera, $\mathcal{T}$ consisting of tasks gathered from the self-narration, and $\mathcal{X}$ consisting of poses determined by a post-collection reconstruction process such as SLAM [41].

B. Model Architecture

We model the affordance prediction task with an encoder-decoder style deep neural network architecture, first encoding the egocentric image as a vector capturing the image's semantics, and then convert this encoding into a task-conditioned prediction of the given task's performance (Figure 2).

Environmental Context and Image Localization To encode the egocentric image at the robot's current viewpoint $\mathbf{I}$, we can use pre-trained, foundational image models that have demonstrated a strong ability to capture the image's semantic information. However, such image encoding models typically capture the semantics of what is being viewed in the image rather than capture information about what to expect of the (unseen) scene surrounding the image. To address this, we fine-tune the weights of the image model, which is intended to capture the expected context given the image. The image encoder, E_V, is a large foundation model pre-trained on a large variety of images (such as CLIP [8]), and will be fine-tuned on a dataset of images related to spatial affordance prediction contained in the video dataset $\mathcal{V}$. That is

$$e_\mathbf{v} = E_V(\mathbf{I}). \tag{2}$$

Task Encoding Unlike images, which need additional learned context, tasks can be encoded directly with pretrained language models such as CLIP [8]. A task query is tokenized then encoded as a vector with a frozen, pre-trained language encoder E_L:

$$e_\boldsymbol{\ell} = E_L(\mathbf{q}) \tag{3}$$

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:

$$\mathcal{N}_\mathbf{t}(\mu_\mathbf{t}, \Sigma_\mathbf{t}) \approx \mathcal{D}(\mathbf{t}) \tag{4}$$

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

$$A_\theta(e_\mathbf{v}, e_\boldsymbol{\ell}) \coloneqq \widetilde{\mathcal{N}}(\mu_\theta, \Sigma_\theta) \tag{5}$$

whose mean is a 3D position and with 2D uncertainty constrained to lie along the ground plane with zero covariance (isotropic). The model parameters θ are learned across environments and activities.

C. Loss Function

We can directly optimize Equation 1 by minimizing the difference in distributions. To ensure the predicted regions $\widetilde{N}$

Kitchen	Activity	Time	Tasks
FAIR	Noodles	9 min	34
GTech	Noodles	20 min	59
IIIT-H-A	Omelette	3 min	47
IIIT-H-B	Tomato Salad	2 min	19
IndianaU	Asian Salad	13 min	52
UMN-A	Scrambled Eggs	10 min	33
UMN-B	Scrambled Eggs	6 min	22
SFU-A	Scrambled Eggs	7 min	16
SFU-B	Coffee Latte	4 min	14
UAndes	Omelette	15 min	23
UPenn	Tomato Salad	8 min	37
UTokyo	Omelette	14 min	66
12 Unique	6 Unique	110 min	422

TABLE I: Kitchen Activities and Tasks

are metrically meaningful, we use the Fréchet Distance, d_F, between the predicted distribution and the canonical target task distribution. Because affordance predictions happen in an egocentric frame, the target task region must be rectified before the distance loss function 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:

$$\mathcal{L}_R = \sum_{\mathbf{x},\mathbf{I}\in\mathbf{V}} \sum_{\mathbf{t}\in\mathcal{T}} d_F(R_{\mathbf{x}}(\mathcal{N}_{\mathbf{t}}), \widetilde{\mathcal{N}}), \tag{6}$$

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

IV. Experimental Setup

A. Training

We curated a dataset consisting of egocentric videos of people accomplishing cooking tasks from the EgoExo4D dataset [3], 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." The resulting dataset contains nearly two hours of localized video recordings gathered from across 12 unique kitchens for a total of 422 different instances of task/environment combinations (Table I). An LLM (GPT-4 [42]) 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 [8] 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 all models on a single V100 GPU and 10 CPU cores. Our base model was trained for 150 epochs in 7 hours of training. Our fine-tuned models are trained on all image frames of the single scene for 25 epochs, taking less than half an hour of training. The resulting models are tested in these environments on the unseen tasks.

B. Baseline: Whole Scene VLMs

Similar to our proposed approach, closely related work such as CLIP-Fields [4], VLMaps [31], and 3D-LLM [33], all build on CLIP encodings to represent semantic image information. However, unlike our proposed approach, these prior works require access to the entire 3D model of the scene at inference time. As a proxy for these types of whole-scene affordance prediction techniques, we introduce a baseline nearest-neighbor based approach which leverages pretrained CLIP as a task-similarity measure that can be applied over all images captured per scene in the dataset (no test/train split). 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 scene $\mathbf{V}$. We can predict the best fitting image as the frame c for which the cosine encoding similarity between the egocentric image and the task description text is maximized. The task position prediction is then $\mathbf{x}_c$, the corresponding position of the viewer 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 $\mathbf{V}$.

V. Results

A. Affordance Grounding

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. 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. Because this task does not involve any spatial prediction or out-of-view tasks, it focuses just on the model's understanding of the connection between task descriptions and an image's affordance.

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 3. 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. Retraining with a frozen E_V (Original CLIP), the model still outperforms the baseline, but by a less significant margin.

Kitchen	Baseline $d_F \downarrow$	Ours ($E_{V\text{frozen}}$) $d_F \downarrow$	Ours $d_F \downarrow$
FAIR	0.44 ± 0.4	0.34 ± 0.1	**0.27 ± 0.1**
GTech	1.56 ± 0.9	0.52 ± 0.4	**0.41 ± 0.3**
IIIT-H-A	0.44 ± 0.2	0.25 ± 0.1	**0.21 ± 0.1**
IIIT-H-B	1.11 ± 1.2	0.54 ± 0.7	**0.52 ± 0.6**
IndianaU	**0.44 ± 0.3**	0.56 ± 0.4	0.46 ± 0.3
UMN-A	0.55 ± 0.4	0.39 ± 0.2	**0.38 ± 0.2**
UMN-B	0.51 ± 0.5	0.48 ± 0.4	**0.40 ± 0.4**
SFU-A	0.59 ± 0.3	0.34 ± 0.2	**0.25 ± 0.2**
SFU-B	0.84 ± 0.2	**0.47 ± 0.2**	0.52 ± 0.4
UAndes	0.70 ± 0.5	0.68 ± 0.4	**0.67 ± 0.5**
UPenn	0.44 ± 0.2	0.35 ± 0.3	**0.28 ± 0.3**
UTokyo	0.44 ± 0.5	0.24 ± 0.2	**0.18 ± 0.1**
Avg Err:	0.68 ± 0.6	0.42 ± 0.3	**0.36 ± 0.3**

TABLE II: Task Region Localization Error, $d_F(m)$

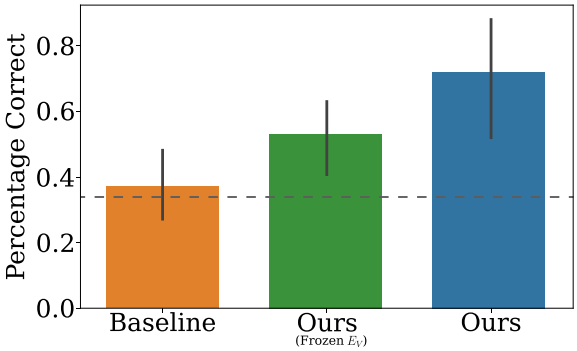

Fig. 3: **Affordance grounding.** When predicting from which in a set of three tasks is the most likely for a given image, the baseline (orange) performs similarly to random guessing (dashed line). Our models with a frozen language encoder (green) significantly outperforms the baseline, and our model with an unfrozen encoder (blue) nearly doubles the baseline.

B. Task Localization

When compared to the baseline, our approach is also significantly more accurate at predicting where a given task will take place relative to an arbitrary egocentric viewpoint (Table II). Task localization allows the model to interface with egocentric navigation techniques such as PointNav [21] and other egocentric robotic works [23], allowing the robot to accomplish tasks such as moving to where you need to "heat the food." Our approach shows statistically significant gain over the baseline [t(82) = 4.683, p <0.001] (Figure 4 left of dashed line) even when testing on both unseen tasks from held-out viewpoints.

The right side Figure 4 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. Taken together, these two results demonstrate the quality of our model's ability to localize within a scene.

C. Rephrasing Stability

Because queries to our model arrive as natural language, the model must be able to make valid predictions across different phrasings of the same task (e.g., "heat the skillet" and "warm up the pan" should have the same predicted position). As our model was trained on a variety of rephrasings, we can expect it to handle this language variation at test time as well. To examine the stability under rephrasing, we generated new synonymous phrases for each task in our testing set, and measure the stability of our prediction over these phrases as the average standard deviation of the predicted position for each rephrasing. We test rephrasing both with the language model used in training, and two other LLMs not seen in training (LLAMA-3 8B [43], and GEMMA-1.1 2B [44]). In all cases, our model was more stable than the baseline (Table III), with only a small amount of variation in predicted positions for different phrasings.

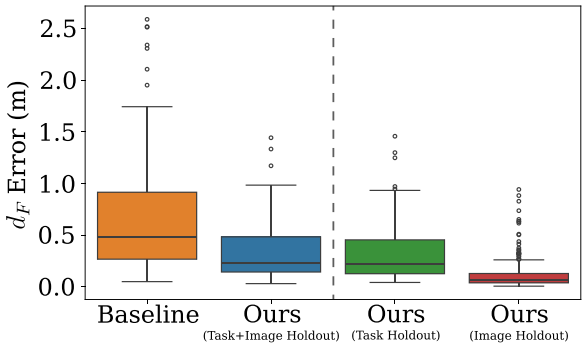

Fig. 4: **Task localization error.** When predicting the region of a given task, the baseline approach does well in some cases (median error of 0.48m) but has many cases with high error or significant outliers. Our approach has significantly lower error than the baseline when testing on unseen images (red), unseen tasks (green), or both unseen images and tasks (blue).

D. Per Scene Fine-tuning

When our model is applied on new environments substantially different from those seen in the training data the quality of the results falls to below that of the baseline. This is expected in that our approach only has access to a single image with a limited field-of-view, meaning that it is forced to guess much of the scene context based only on what is a typical kitchen layout whereas the baseline has access to ground-truth labeled data for every task of every frame in the scene. While we expect training on larger collections of scenes similar to those in testing would somewhat improve generalizing to new scenes, in practice there is too large a degree of variety in environments to reliably produce high-quality predictions of the entire scene from a single image.

A more practical approach is to fine-tune our trained model based on short demonstrations in the new environment. Surprisingly, only a single demonstration is needed to signifi-

LLM	Baseline	Ours
GPT-4	0.19m	**0.12m**
LLAMA 3	0.18m	**0.13m**
Gemma 1.1	0.23m	**0.16m**

TABLE III: Rephrasing Stability

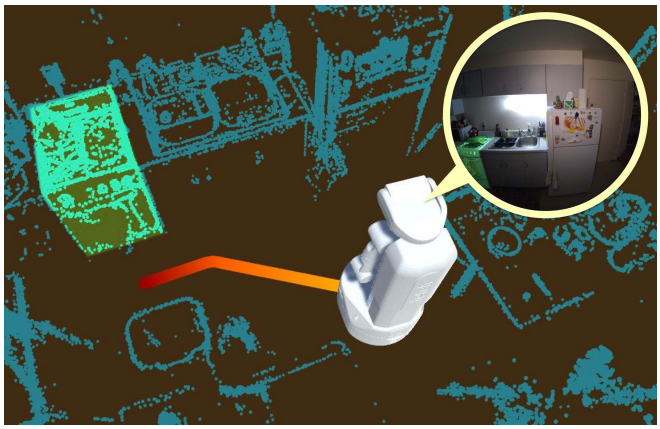

Fig. 5: Trajectory to "Heat the Food" (stove highlighted).

cantly outperform the baseline. In fact, across three different kitchens unseen in training, adding a single demonstration of several tasks from one activity halves the error on unseen tasks within the same activity as shown in Table IV.

Importantly, we find that fine-tuning only the affordance head A results in nearly equal performance gain compared to fine tuning both A and E_V. This allows fine-tuning of the deployed model to happen on commodity GPUs, as optimizing A requires significantly less memory than E_V.

	UMN-C $d_F \downarrow$	UMN-D $d_F \downarrow$	SFU-C $d_F \downarrow$
Baseline	0.28 ± 0.3m	0.34 ± 0.5m	0.69 ± 0.4m
Ours (base)	0.55 ± 0.3m	0.70 ± 0.3m	1.05 ± 0.5m
Ours (FT A)	0.20 ± 0.2m	0.31 ± 0.4m	0.45 ± 0.3m
Ours (FT all)	**0.18 ± 0.2m**	**0.26 ± 0.3m**	**0.39 ± 0.3m**

TABLE IV: Fine-tuning on demonstrations in new scenes.

VI. Navigation Applications

A. EgoCentric Robot Navigation

To characterize the ability of our system to support task-based robot navigation, we collected a new dataset of images from one of the physical environments seen in training. We then used a custom simulator to allow a robot to navigate based on these newly collected images to positions appropriate for new tasks unseen in training. We collected these images using an Aria camera [45] as in training, and based the simulation on the Fetch robot [46] as it has similar physical affordance to humans.

An example navigation is shown in Figure 5. Here a robot is given a new view (shown in the inset bubble) and asked to navigate to the task "*Heat the Food*". Given this single egocentric robot view, the robot is able to predict the tasks' location. A navigation mesh of estimated free space is used to avoid collision during motion.

B. Task Obstacles

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 network 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. Figure 6 shows examples of task obstacles.

Fig. 6: **Task Obstacles** predicted across two task-sets. (left) "Preheat the oil, and wash the dishes" and (right) "Get dishes from the cabinet, and serve dinner."

Algorithm 1: Task Obstacle Generation

Load $\mathbf{A} := E_V,\ E_L,\ A_\theta$,
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($\mathcal{D}$, σ_{bound}) for $\mathcal{D}$ in distributions]
points $=$ [discretize($\mathbf{r}$) for $\mathbf{r}$ in regions]
task_obstacle $=$ convex_hull(points)

VII. Discussion

We have introduced a new framework to predict spatial affordances of where people perform tasks within a robot's environment. Our system is trained on egocentric video demonstrations and shows generalizability to new tasks described in natural language

Limitations & Future Work 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. While fine-tuning on demonstrations in the new environments helps, it still requires new training cycles which could be inconvenient in a deployed system. 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-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.

REFERENCES

[1] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G. M. Farinella, D. Damen, and T. Tommasi, "An outlook into the future of egocentric vision," *International Journal of Computer Vision*, pp. 1–57, 2024.

[2] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. González, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolář, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. Ruiz, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbeláez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. 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)*, June 2022, pp. 18 995–19 012.

[3] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, E. Byrne, Z. Chavis, J. Chen, F. Cheng, F.-J. Chu, S. Crane, A. Dasgupta, J. Dong, M. Escobar, C. Forigua, A. Gebreselasie, S. Haresh, J. Huang, M. M. Islam, S. Jain, R. Khirodkar, D. Kukreja, K. J. Liang, J.-W. Liu, S. Majumder, Y. Mao, M. Martin, E. Mavroudi, T. Nagarajan, F. Ragusa, S. K. Ramakrishnan, L. Seminara, A. Somayazulu, Y. Song, S. Su, Z. Xue, E. Zhang, J. Zhang, A. Castillo, C. Chen, X. Fu, R. Furuta, C. Gonzalez, P. Gupta, J. Hu, Y. Huang, Y. Huang, W. Khoo, A. Kumar, R. Kuo, S. Lakhavani, M. Liu, M. Luo, Z. Luo, B. Meredith, A. Miller, O. Oguntola, X. Pan, P. Peng, S. Pramanick, M. Ramazanova, F. Ryan, W. Shan, K. Somasundaram, C. Song, A. Southerland, M. Tateno, H. Wang, Y. Wang, T. Yagi, M. Yan, X. Yang, Z. Yu, S. C. Zha, C. Zhao, Z. Zhao, Z. Zhu, J. Zhuo, P. Arbelaez, G. Bertasius, D. Crandall, D. Damen, J. Engel, G. M. Farinella, A. Furnari, B. Ghanem, J. Hoffman, C. V. Jawahar, R. Newcombe, H. S. Park, J. M. Rehg, Y. Sato, M. Savva, J. Shi, M. Z. Shou, and M. Wray, "Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives," 2023.

[4] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, "Clip-fields: Weakly supervised semantic fields for robotic memory," 2023.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, p. 99–106, dec 2021. [Online]. Available: https://doi.org/10.1145/3503250

[6] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai, "PaliGemma: A versatile 3B VLM for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[7] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2024.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[9] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[10] Z. Cao, H. Gao, K. Mangalam, Q. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," in *European Conference on Computer Vision (ECCV)*, 2020.

[11] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] J. Qiu, L. Chen, X. Gu, F. P.-W. Lo, Y.-Y. Tsai, J. Sun, J. Liu, and B. P. L. Lo, "Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion," *IEEE Robotics and Automation Letters*, vol. 7, pp. 8799–8806, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:240353763

[13] B. Pan, B. Shen, D. Rempe, D. Paschalidou, K. Mo, Y. Yang, and L. J. Guibas, "Copilot: Human collision prediction and localization from multi-view egocentric videos," *ArXiv*, vol. abs/2210.01781, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252692922

[14] J. Wang, D. Luvizon, W. Xu, L. Liu, K. Sarkar, and C. Theobalt, "Scene-aware egocentric 3d human pose estimation," *CVPR*, 2023.

[15] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022.

[16] K. Q. Lin, J. Wang, M. Soldan, M. Wray, R. Yan, E. Z. XU, D. Gao, R.-C. Tu, W. Zhao, W. Kong, C. Cai, W. HongFa, D. Damen, B. Ghanem, W. Liu, and M. Z. Shou, "Egocentric video-language pretraining," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 7575–7586. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/31fb284a0aaaad837d2930a610cd5e50-Paper-Conference.pdf

[17] M. El Banani, A. Raj, K.-K. Maninis, A. Kar, Y. Li, M. Rubinstein, D. Sun, L. Guibas, J. Johnson, and V. Jampani, "Probing the 3D Awareness of Visual Foundation Models," in *CVPR*, 2024.

[18] M. Liu, L. Ma, K. Somasundaram, Y. Li, K. Grauman, J. M. Rehg, and C. Li, "Egocentric activity recognition and localization on a 3d map," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 621–638.

[19] T. Nagarajan and K. Grauman, "Shaping embodied agent behavior with activity-context priors from egocentric video," *ArXiv*, vol. abs/2110.07692, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239009498

[20] S. Datta, O. Maksymets, J. Hoffman, S. Lee, D. Batra, and D. Parikh, "Integrating egocentric localization for more realistic point-goal navigation agents," *ArXiv*, vol. abs/2009.03231, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221516690

[21] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *ICLR*, 2020. [Online]. Available: https://arxiv.org/pdf/1911.00357

[22] O. Maksymets, V. Cartillier, A. Gokaslan, E. Wijmans, W. Galuba, S. Lee, and D. Batra, "Thda: Treasure hunt data augmentation for semantic navigation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15 354–15 363, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244100709

[23] D. Hoeller, L. Wellhausen, F. Farshidian, and M. Hutter, "Learning a state representation and navigation in cluttered and dynamic environments," *IEEE Robotics and Automation Letters*, vol. 6, pp. 5081–5088, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232147867

[24] R. Partsey, E. Wijmans, N. Yokoyama, O. Dobosevych, D. Batra, and O. Maksymets, "Is mapping necessary for realistic pointgoal navigation?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17 232–17 241.

[25] T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Science Robotics*, vol. 8, no. 79, p. eadf6991, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.adf6991

[26] A. Kumar, S. Gupta, and J. Malik, "Learning navigation subroutines from egocentric videos," in *Proceedings of the Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, L. P. Kaelbling, D. Kragic, and K. Sugiura, Eds., vol. 100. PMLR, 30 Oct–01 Nov 2020, pp. 617–626. [Online]. Available: https://proceedings.mlr.press/v100/kumar20a.html

[27] D. Shah, B. Osinski, B. Ichter, and S. Levine, "LM-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=UW5A3SweAH

[28] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. 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.

[29] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, "Zson: Zero-shot object-goal navigation using multimodal goal embeddings," in *Neural Information Processing Systems (NeurIPS)*, 2022.

[30] V. S. Dorbala, G. A. Sigurdsson, R. Piramuthu, J. Thomason, and G. S. Sukhatme, "Clip-nav: Using clip for zero-shot vision-and-language navigation," *ArXiv*, vol. abs/2211.16649, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254095893

[31] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[32] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, "Open-vocabulary queryable scene representations for real world planning," in *arXiv preprint arXiv:2209.09874*, 2022.

[33] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 20 482–20 494. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/413885e70482b95dcbeeddc1daf39177-Paper-Conference.pdf

[34] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *The international journal of robotics research*, vol. 30, no. 7, pp. 846–894, 2011.

[35] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *International Conference on Robotics and Automation (ICRA)*, 2018.

[36] B. Wen, W. Lian, K. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," *ICRA 2022*, 2022.

[37] A. Makhal and A. K. Goins, "Reuleaux: Robot base placement by reachability analysis," *2018 Second IEEE International Conference on Robotic Computing (IRC)*, pp. 137–142, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:4596903

[38] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," *arXiv preprint arXiv: Arxiv-2310.12931*, 2023.

[39] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 778–13 790.

[40] T. Yoshida, S. Kurita, T. Nishimura, and S. Mori, "Text-driven affordance learning from egocentric vision," 2024.

[41] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[42] OpenAI, "Gpt-4 technical report," 2023.

[43] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[44] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy, "Gemma: Open models based on gemini research and technology," 2024.

[45] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, C. Peng, C. Sweeney, C. Wilson, D. Barnes, D. DeTone, D. Caruso, D. Valleroy, D. Ginjupalli, D. Frost, E. Miller, E. Mueggler, E. Oleinik, F. Zhang, G. Somasundaram, G. Solaira, H. Lanaras, H. Howard-Jenkins, H. Tang, H. J. Kim, J. Rivera, J. Luo, J. Dong, J. Straub, K. Bailey, K. Eckenhoff, L. Ma, L. Pesqueira, M. Schwesinger, M. Monge, N. Yang, N. Charron, N. Raina, O. Parkhi, P. Borschowa, P. Moulon, P. Gupta, R. Mur-Artal, R. Pennington, S. Kulkarni, S. Miglani, S. Gondi, S. Solanki, S. Diener, S. Cheng, S. Green, S. Saarinen, S. Patra, T. Mourikis, T. Whelan, T. Singh, V. Balntas, V. Baiyya, W. Dreewes, X. Pan, Y. Lou, Y. Zhao, Y. Mansour, Y. Zou, Z. Lv, Z. Wang, M. Yan, C. Ren, R. D. Nardi, and R. Newcombe, "Project aria: A new tool for egocentric multi-modal ai research," 2023.

[46] M. Wise, M. Ferguson, D. King, E. Diehr, and D. Dymesich, "Fetch and freight: Standard platforms for service robot applications," in *Workshop on autonomous mobile service robots*, 2016, pp. 1–6. [Online]. Available: https://api.semanticscholar.org/CorpusID:42886148