HELIOS: Hierarchical Exploration for Language-grounded Interaction in Open Scenes

Katrina Ashton[1], Chahyon Ku[2], Shrey Shah[2], Wen Jiang[1], Kostas Daniilidis[1], Bernadette Bucher[2]
[1] University of Pennsylvania [2] University of Michigan

***Abstract*— Language-specified mobile manipulation tasks in novel environments simultaneously face challenges interacting with a scene which is only partially observed, grounding semantic information from language instructions to the partially observed scene, and actively updating knowledge of the scene with new observations. To address these challenges, we propose HELIOS, a hierarchical scene representation and associated search objective to perform language specified pick and place mobile manipulation tasks. We construct 2D maps containing the relevant semantic and occupancy information for navigation while simultaneously actively constructing 3D Gaussian representations of task-relevant objects. We fuse observations across this multi-layered representation while explicitly modeling the multi-view consistency of the detections of each object. In order to efficiently search for the target object, we formulate an objective function balancing exploration of unobserved or uncertain regions with exploitation of scene semantic information. We evaluate HELIOS on the OVMM benchmark in the Habitat simulator, a pick and place benchmark in which perception is challenging due to large and complex scenes with comparatively small target objects. HELIOS achieves state-of-the-art results on OVMM. As our approach is zero-shot, HELIOS can also transfer to the real world without requiring additional data, as we illustrate by demonstrating it in a real world office environment on a Spot robot.**

I. Introduction

Consider an autonomous robot tasked with bringing a mug from a coffee table to the kitchen counter in a home. If that robot sees a coffee table but cannot currently detect a mug on it, should it go closer to investigate if the mug is actually present? Or should it look in new parts of the home? An autonomous robot should be able to efficiently reason through this question using environment cues. In addition, the robot should be able to successfully perform this task of language-specified pick and place for mobile manipulation using the observations it accumulates during this search process.

Methods for embodied physical intelligence can accumulate information about a novel scene and act on it though observation history with no explicit scene representation [1], [2], [3], only 2D maps [4], [5] or 3D scene graphs [6], [7], [8]. However, these methods all assume dense associations between language, observation, and action. Very different representations for long horizon spatio-temporal reasoning have been developed in problems for semantic search where language grounding is sparse [9], [10], [11]. In order to perform mobile manipulation which includes semantic search, reasoning over vision, language, and action must occur simultaneously in both long and short horizons. Low success rates on new benchmarks targeting open vocabulary pick and place tasks in novel environments have demonstrated that combining this long and short horizon reasoning is still an open challenge [7], [4].

Reasoning jointly over short and long spatio-temporal contexts requires very different policy objectives in addition to the differences in scene representations. Prior work in object search explicitly manages local and global search problems distinctly [12], [13], [14]. Search policies must figure out when to switch between local and global reasoning by deciding the likelihood of being close to the target object. In addition to exploring unobserved regions, efficient search policies also exploit semantic information about the scene in order to search more likely locations of the target object first [15], [16], [17], [18], [9], [19], [10]. This exploration-exploitation tradeoff adds additional complexity to the task of performing object search as a component of mobile manipulation.

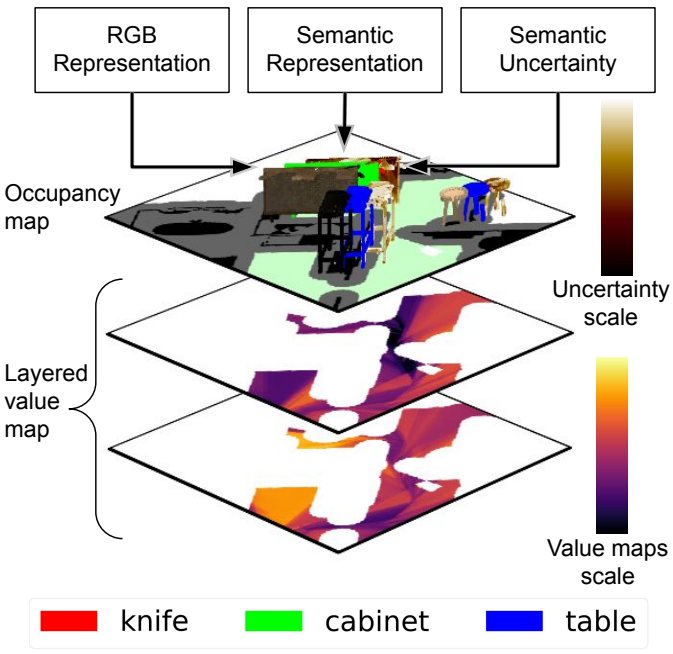

Fig. 1: **Hierarchical scene representation.**

Contributions. We present HELIOS, a hierarchical scene representation and search objective for language specified mobile pick and place tasks in novel environments. We create a hierarchical scene representation using layered 2D value and occupancy maps to efficiently navigate and explore, and sparse collections of 3D Gaussians to represent objects of

interest (see fig. 1). We then formulate an objective function on our hierarchical scene representation that balances exploring the scene to find regions which might contain the target object with exploiting observed semantic information. We introduce an uncertainty-weighted object score to take into account the multi-view consistency of the detections of an object before interacting with it. We conduct an ablation study to verify that each of these components increases our method's performance. Through our experiments, we show the contribution of uncertainty-based reasoning over our novel visual representation in improving robust perception in mobile manipulation. We evaluate HELIOS on the HomeRobot Open-vocabulary Mobile Manipulation benchmark [4], [20] in the Habitat simulator [21], achieving state-of-the-art results. We use HELIOS in semantic navigation as a stop decision, improving overall search success on the Habitat-Matterport 3D [22] object search benchmark. The zero-shot nature of our approach means it can transfer to the real world without requiring additional data, as we show by demonstrating HELIOS in a real world office environment on a Spot robot.

II. Related work

Language-grounded open world pick and place. Recent advancements in vision and language have opened up challenges in natural language instruction following for robots in novel environments. Many methods focus on parsing complex or ambiguous language and accurately grounding this language to observations made during task execution [23], [6], [24], [25], [8]. Others focus on improving execution of language specified pick and place skills [26], [27], [25]. However, benchmarks for targeted instantiations of this problem have identified that a major cause of failure in this task is correctly finding and identifying objects for performing pick and place [4], [7], [5]. Our work addresses this challenge by modeling the multi-view consistency of object detections, allowing us to only interact with objects once we have obtained enough views that we are confident in the results of the object detection.

Object search and detection. To find an object with an RGB camera, that camera needs to record sufficient observations in the environment to correctly identify the object. Active object detection methods obtain additional views of a scene in order to capture an image from which a target object can be correctly identified [28], [29], [30]. When these observations are accumulated in a map of the environment, it enables a larger scale search problem in which the camera is systematically moved to possible locations in the map. Hierarchical object search methods explicitly perform global and local object search to ensure sensor coverage of the scene [12], [13], [14]. To perform object search efficiently, semantic information can be used as a prior about where objects are more likely to be [31]. This semantic prior naturally yields an exploration and exploitation tradeoff [15], [16], [17], [18], [9], [19], [10]. In our work, we perform object search and detection as part of pick and place mobile manipulation tasks. Therefore, we construct an objective for switching between global object search and local object detection while simultaneously trading off exploration of the scene and exploitation of semantic information.

3D Gaussians in robot perception. 3D Gaussians [32] have been used in a variety of robotics tasks including SLAM [33], [34], active mapping [35], [36], [37], and table-top manipulation [38], [39]. These methods all build a dense 3D representation of the entire scene. Many methods also incorporate open-vocabulary semantic features in 3D Gaussian representations [40], [41], [42], [43]. In contrast to previous robot perception approaches, we only model target objects of interest with 3D Gaussians, building a sparse 3D map. We adapt wilson2024modeling to perform semantic classification and estimate the associated uncertainty in our sparse 3D Gaussian object map, which forms one layer of our scene representation.

Language-grounded scene representations. Language-grounded scene representations can be dense or sparse. Dense open-vocabulary 3D scene representations map vision-language features which can be dynamically queried with language [44], [45], [46], [47], [41], [48]. However, these dense 3D representations are not necessarily effective or efficient for performing planning and control. For semantic navigation tasks, dense 2D language-grounded scene representations are more efficient and have been shown to be effective [49], [10], [9], [50]. For language specified manipulation tasks, instance level information about objects is important [51], [52], [53], [54]. To enable mobile manipulation, 3D scene graphs build globally consistent maps of object centric representations needed for manipulation [55], [8], [56], [57], [58], [59]. Our work builds on this direction in mobile manipulation by using object instance information to construct a sparse map of 3D Gaussians. In our work, we combine this information for manipulation in a hierarchical map with 2D value maps for semantic navigation.

III. Method

We address the problem of language specified pick and place mobile manipulation tasks in novel environments. To carry out this task, the robot first needs to solve a search problem to find the target object, including correctly identifying the target object. It must then navigate to a suitable grasp position and grasp the object. Finally, it needs to solve another search problem in order to find the place location, and then place the object there in a stable orientation. Note that all of these stages need to be successful, and the robot must also avoid collisions with the environment when navigating and interacting with the objects, so this task is subject to compounding error rates. However the robot can also use information collected in previous stages of the task to aid it later. For example, the search to find the place location can be made more efficient by utilizing information collected when the robot was searching for the target object. In order to collate this information into a useful and efficient format, we propose constructing a hierarchical task-driven map (see Section III-A) with 2D map layers suitable for the search problems and 3D Gaussians to represent objects in the

scene relevant to manipulation. We detail how we explicitly reason over this map to solve a language specified pick and place task in Section III-B.

A. Hierarchical Task-driven Map

We construct a hierarchical map with three layers, where each layer corresponds to the three primary tasks that the robot needs to complete. First, to navigate around obstacles to a specified goal location, the robot requires an occupancy map to perform collision free path planning. Second, to efficiently search for objects, the robot can use semantic information in the environment to prioritize exploring unobserved regions which are similar to target locations. Finally, in order to effectively manipulate and perform robust detection of the objects of interest, we model the components of the scene where we expect to perform pick and place with a sparse 3D representation using 3D Gaussians assigned to instances of classes referenced in the instruction.

1) 2D Occupancy Maps: We construct a 2D bird's-eye view (BEV) occupancy map by ground projecting depth measurements. We use this map to perform collision-free path planning to navigate around obstacles to goal locations. We also identify frontiers on the occupancy map, defined as center-points of boundaries between explored and unexplored areas, which will enable us to search unknown map regions.

2) 2D Semantic Value Map: To choose between frontier points, we leverage semantic information about the scene in order to search efficiently by going to areas more likely to contain the target of interest first. We construct a layered semantic value map to enable this frontier-based approach by extending prior work constructing semantic value maps [10] to incorporate multiple search targets. Each layer in our map is a 2D BEV value map constructed by using BLIP-2 [60] to score the similarity of each observed RGB image to the prompt `Seems like there is a (object) ahead` and fusing the results using a confidence based on the field-of-view cone for each observation. We construct one map layer for the pick location and one for the place location. Since our method is open vocabulary, we can specify the pick location by either referencing the target object directly or by referencing components of the scene where the target object is expected to be.

3) 3D Gaussian representation for modeling objects: In order to enable reasoning about the multi-view consistency of semantic classifications, we represent the objects of interest in the scene using 3D Gaussian Splatting (3DGS) [32]. To increase efficiency over prior applications of 3DGS to robotic tasks [38], [39], instead of modeling the entire scene with 3D Gaussians we only use them to model parts of the scene which have been detected as objects of interest. We assign Gaussians to object instances, allowing us to reason over objects in the scene instead of individual Gaussians. Our sparse 3DGS representation supports tracking the semantic class probability and semantic class uncertainty for each Gaussian which we use to create a novel uncertainty-weighted object score for each instance.

Preliminaries – 3D Gaussian representation rendering. A 3D Gaussian $x(\mu, \Sigma; c, \alpha)$ is defined by its mean position μ, covariance Σ, color c and opacity α, these characteristics can be learned via a rendering loss. A scene is rendered with many of these 3D Gaussians, the final number determined by the task specific conditions in which Gaussians are added and removed. When an image is rendered using 3DGS, the 3D Gaussians comprising the scene representation are first transformed from the world frame to the camera frame and then projected into 2D Gaussians (splats) in the image plane, $x(\mu, \Sigma; c, \alpha) \mapsto \tilde{x}(\tilde{\mu}, \tilde{\Sigma}; c, \alpha)$. Each pixel i's color Q_i is then calculated from the 2D Gaussians using α-blending for the N ordered points on the 2D splats that overlap the pixel. For a pixel with position p_i and a 3D Gaussian x_n, we first find the opacity $\tilde{\alpha}_n(p_i)$ of the corresponding 2D Gaussian at that pixel position by weighting based on the pixel's distance to the center of the 2D Gaussian with $\tilde{\alpha}_n(p_i) = \alpha_n \cdot k(p_i, \tilde{x}_n)$, where $k(p_i, \tilde{x}_n) = \exp\left(\frac{1}{2}(p_i - \tilde{\mu}_n)\tilde{\Sigma}_n^{-1}(p_i - \tilde{\mu}_n)\right)$. Next, the N Gaussians are ordered based on depth, with $\tilde{x}_1$ being the closest to the camera, and the final contribution for each Gaussian is calculated with α-blending to get the final pixel color $Q_i = \sum_{n=1}^{N} c_n \kappa(p_i, \tilde{x}_n; \{\tilde{x}_j\}_{j\in\{1,\dots,N\}})$ where

$$\kappa(p_i, \tilde{x}_n; \{\tilde{x}_j\}_{j\in\{1,\dots,N\}}) := \tilde{\alpha}_n(p_i) \prod_{j=1}^{n-1} (1 - \tilde{\alpha}_j(p_i)). \quad (1)$$

Preliminaries – Semantic classes for 3D Gaussian representation. We represent the semantic class scores with our 3DGS model in addition to color. Following wilson2024modeling, we explicitly model the distribution of semantic estimates of each Gaussian using the categorical distribution. This distribution is then updated using its conjugate prior, the Dirichlet distribution. Note that this method requires specifying number of object classes at the start of the episode. However, any amount of classes can be specified, so this approach supports open-vocabulary mobile manipulation. The probability density function (PDF) of the Dirichlet distribution is given by

$$f(\theta_n | \boldsymbol{\gamma}_n) = \frac{1}{B(\boldsymbol{\gamma}_n)} \prod_{c=1}^{C} \theta_{n,c}^{\gamma_n^c - 1}. \quad (2)$$

where B is the multivariate beta function and C is the number of classes. In our case, θ_n is the categorical distribution for the Gaussian x_n. The concentration parameters, $\boldsymbol{\gamma}_n = (\gamma_n^1, ..., \gamma_n^C)$, of the Dirichlet distribution can be updated after each measurement using Bayesian Kernel Inference as follows [43]

$$\gamma_n^c \leftarrow \gamma_n^c + \sum_{i=1}^{N} y_i^c \kappa(p_i, \tilde{x}_n; \{\tilde{x}_j\}_{j\in\{1,\dots,N\}}), \quad (3)$$

where y_i^c is 1 if p_i is of class c and 0 otherwise and $\kappa(\cdot)$ is defined in eq. (1).

Then, for a 3D Gaussian x_n and class c, the expected probability of x_n being of category c and its variance is

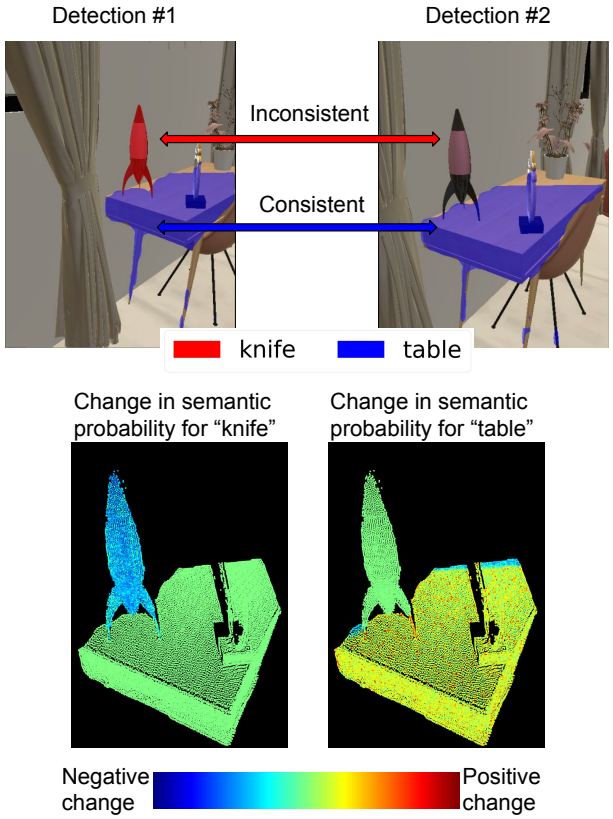

Fig. 2: **Example of multi-view fusion.** We show two observations, in the first a toy rocket is incorrectly identified as a knife and the table is correctly identified, in the second the table is again correctly identified. Below this we show the change in the semantic probability for each class in the 3DGS part of our scene representation when it is updated with the second detection. We can see that the incorrect detection of the object on the table as a knife is not multi-view consistent and so the probability of this object being a knife goes down when we include the second detection. The table is correctly detected across multiple frames so the probability goes up after fusion.

given by

$$\mathbb{E}[\theta_n^c] = \frac{\gamma_n^c}{\sum_{j=1}^{C} \gamma_n^j}, \quad \text{Var}[\theta_n^c] = \frac{\mathbb{E}[\theta_n^c](1 - \mathbb{E}[\theta_n^c])}{1 + \sum_{j=1}^{C} \gamma_n^j}. \tag{4}$$

The variance can be considered a measure of the pixel-wise uncertainty of that class score based on the multi-view consistency. During rendering we use $\mathbb{E}[\theta_n^c]$ and $\sqrt{\text{Var}[\theta_n^c]}$ in place of the color parameter for rendering the semantic class scores and uncertainty, respectively. Figure 2 shows an example of how the semantic class score is updated when we obtain a new measurement.

Preliminaries – Information gain. Using the Dirichlet distribution to model the semantic state of the Gaussians allows us to find the entropy of the concentration parameters [61]

$$H(\theta_n) = \log B(\boldsymbol{\gamma}_n) + (T(\gamma_n) - C)\psi(T(\gamma_n)) - \sum_{c=1}^{C}(\gamma_n^c - 1)\psi(\gamma_n^c), \tag{5}$$

where $T(\gamma_n) := \sum_{c=1}^{C} \gamma_n^c$ and ψ is the digamma function.

If we obtain a set of new observations, $Y = \{y_1, ..., y_m\}$ at poses $P = \{p_1, .., p_m\}$ then the information gain is

$$\text{IG}(\theta_n, Y|P) = H(\theta_n) - H(\theta_n|P, Y). \tag{6}$$

Given P and Y, $H(\theta_n|P, Y)$ can be found by updating θ_n and then calculating the updated entropy.

Instances for object-level reasoning. We assign 3D Gaussians to instances so we can reason about objects. Because the objects are not always perfectly segmented this assignment is done by clustering in 3D within Gaussians which have the same most likely semantic class. To prevent the time requirements becoming intractable for large scenes, we detect which Gaussians are updated for a new observation and only perform the clustering with these Gaussians and any other Gaussians within the same instance.

Using these instances we can reason over the set of objects our representation is modeling, let us call this set $\mathcal{O}$. Each object in $\mathcal{O}$ consists of 3D Gaussians belonging to the same instance, and the class of this object is given by the most common highest-probable class among the 3D Gaussians belonging to that instance, i.e. for $o_i \in \mathcal{O}$, its class is given by $\text{mode}_{\theta \in o_i}\left(\text{argmax}_{c \in \{\text{classes}\}} \mathbb{E}[\theta_n^o]\right)$.

For each object $o_i \in \mathcal{O}$ we also define the class score $S_c := \frac{1}{|o_i|}\sum_{\theta_n \in o_i} \mathbb{E}[\theta_n^c]$, that is, the mean probability of the 3D Gaussians which make up the instance o_i being of class c. Likewise, we define the uncertainty $U_c := \frac{1}{|o_i|}\sum_{\theta_n \in o_i} \sqrt{\text{Var}[\theta_n^c]}$.

a) Uncertainty-weighted object score: To determine whether we are confident in our estimate of an object's class we define our uncertainty-weighted object score, which takes into account both the class score and uncertainty (balanced by a hyper-parameter α_{cs}) for an object $o_i \in \mathcal{O}$ for class c:

$$\Psi_c(o_i) := S_c(o_i) - \alpha_{cs} U_c(o_i). \tag{7}$$

That is, the lower bound of the α_{cs}-sigma estimate of o_i.

B. Hierarchical Search

We plan over our hierarchical scene representation in a zero-shot manner, searching for the pick location using our global search objective to balance between exploring new frontiers and exploiting semantic information. Once we detect a target object we use our uncertainty-weighted object score to decide whether we are confident enough in the classification to attempt to grasp it. Once the target object has been picked up we perform a similar search procedure until we are confident we have found the place location. Figure 3 shows the logical flow of our method.

Global search objective. Our global search objective balances exploring new frontiers with exploiting detections

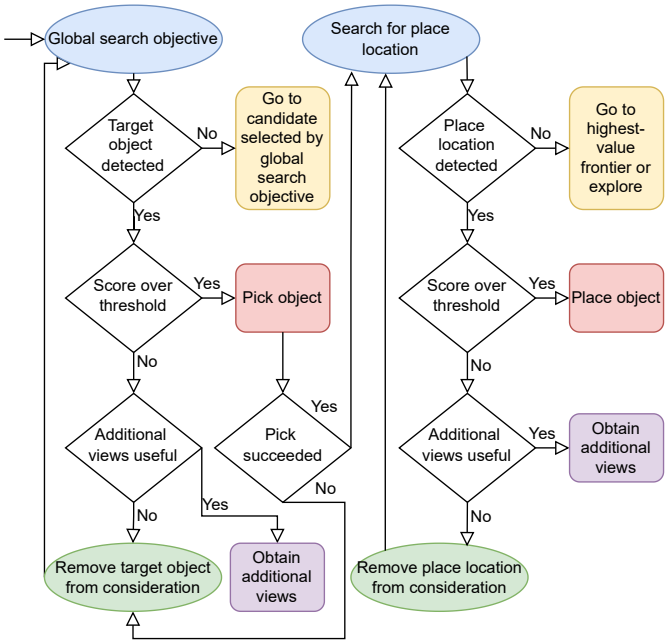

Fig. 3: **Method flow chart for HELIOS.**

of candidate pick locations. First we introduce some new notion, let $\mathcal{A} \subset \mathcal{O}$ be the set of objects whose class is that of the pick location and let $\mathcal{F}$ be the set of frontiers.

First, we will evaluate the benefit of searching for a detected object. We can work out whether obtaining additional views Y from poses P of candidate pick location $a_i \in \mathcal{A}$ is likely to be informative by considering the information gain (IG). We obtain the proposed poses as described in the local search section, but we do not have the observations Y unless we move to these poses. In the case of search, we prioritize avoiding false negatives more than false positives since ultimately an effective search policy should provide coverage of the full search space. Thus, we propose an optimistic approach where we assume the best-case scenario that all the observations in Y classify a_i as the pick location a. Specifically, we define the estimated information gain as $\text{IG}_a(a_i|P, Y^*) := \sum_{\theta_n \in a_i} H(\theta_n) - H(\theta_n|P, Y^*)$, where Y^* classifies a_i as class a. We will drop the condition and just write $\text{IG}_a(a_i)$ for brevity. We can then combine the class score and the IG by multiplying them, i.e. $S_a(a_i)\text{IG}_a(a_i)$, to get a measure of how much we want to search a candidate pick location a_i.

This information gain weighted object score allows us to compare candidate objects to each other, but we also want to be able to compare them to frontiers. When we choose a frontier $f_i \in \mathcal{F}$, we store its location and current score from our value map, denote this $F_0(f_i)$. During global planning, the first time each $a_i \in \mathcal{A}$ is detected we store the initial class score, $S_{a0}(a_i)$, the initial information gain, $\text{IG}_{a0}(a_i)$ as well as its initial center position. Then, we want to find the best candidate object while taking into account the distance to the frontier. Explicitly, let $\mathcal{F}'$ be the set of previously chosen frontiers. Then we can calculate an estimated value for a previously chosen frontier $f'_i \in \mathcal{F}'$ based on its proximity to detected candidate objects as:

$$V_0(f'_i) := \max_{a_i \in \mathcal{A}} \left(S_{a0}(a_i)\text{IG}_{a0}(a_i) - \alpha_d \text{dist}(a_j, f'_i) \right) \tag{8}$$

where α_d is a hyper-parameter which controls the relative importance of candidate object score to distance and $\text{dist}(a_j, f'_i)$ is the Euclidean distance between the stored center of a_j and f'_i.

Given this association between previous frontiers and candidate object scores we can find an association between frontier scores and candidate object scores by averaging the ratio of this new score to the frontier score over all the previous frontiers:

$$F_0 := \frac{1}{|\mathcal{F}'|} \sum_{f'_i \in \mathcal{F}_p} \frac{V_0(f'_i)}{F_0(f'_i)} \tag{9}$$

This allows us to associate a frontier f_i with a candidate object score by multiplying its score $F(f_i)$ by F_0. We also take into account distance to form the following score function for $r_i \in \mathcal{A} \cup \mathcal{F}$:

$$V(r_i) := \begin{cases} S_a(r_i)\text{IG}_a(r_i) - \alpha_d \text{dist}(r_i) & \text{if } r_i \in \mathcal{A} \\ F(r_i)F_0 - \alpha_d \text{dist}(r_i) & \text{if } r_i \in \mathcal{F} \end{cases} \tag{10}$$

where $F(f_i)$ is the current score from our value map for $f_i \in \mathcal{F}$ and $\text{dist}(r_i)$ is the Euclidean distance from the agent to the center point of r_i.

Local search. When local search is performed on an identified object we generate and go to gaze point positions in a contour around the 2D ground-projection of the 3D Gaussians making up our representation of that object. The orientation of a gaze point is set so that the agent will look towards the center of the object in the ground-plane and the highest point on the object. After performing local search on an object we mark it as visited and no longer consider it a candidate for local search.

IV. Experimental Results

A. Open vocabulary mobile pick and place in a novel environment

Dataset and benchmark. We evaluate HELIOS on the validation split of the Home Robot OVMM benchmark [4], [20] which uses scenes from the Habitat Synthetic Scenes Dataset (HSSD) [62] in the Habitat simulator [63] and consists of 1199 episodes. In this benchmark, the robot must carry out an instruction of the form "Move (`object`) from the (`start_receptacle`) to the (`goal_receptacle`)" in an unknown environment. An oracle pick skill is provided, and we use a simple heuristic place skill to drop the object above the `goal_receptacle`.

Metrics. We report the following metrics from the OVMM benchmark [4], [20] indicating the success of each phase of the task: **FindObj** if the robot is ever close enough to the `object`, **Pick** if the robot successfully picks up the `object`, **FindRec** if the robot is ever close enough to a `goal_receptacle` after picking up the `object`. We

TABLE I: **Ablation study for components of our method**, with comparison to using the HomeRobot [4] baseline agents and recent method MoManipVLA [26] on the val split of the OVMM challenge. For HomeRobot the results are included for different configurations of skills for navigation, gaze and place. E.g. R/N/H uses RL for navigation, no skill for gaze and heuristic skill for place.

	Method	FindObj	Pick	FindRec	Place	SR
	HomeRobot H/N/H	28.7	15.2	5.3	-	0.4
	HomeRobot H/R/R	**29.4**	13.2	5.8	-	0.5
	HomeRobot R/N/H	21.9	11.5	6.0	-	0.6
	HomeRobot R/R/R	21.7	10.2	6.2	-	0.4
	MoManipVLA[1]	23.7	12.7	7.1	-	1.7
1 pick	Trusting agent	13.7 ± 1.0	12.3 ± 0.9	6.8 ± 0.7	2.1 ± 0.4	1.3 ± 0.3
1 pick	W/o global search objective	16.8 ± 1.1	12.0 ± 0.9	6.8 ± 0.7	2.6 ± 0.5	1.7 ± 0.4
1 pick	HELIOS	23.8 ± 1.2	**17.2 ± 1.1**	**10.0 ± 0.9**	**3.3 ± 0.5**	**2.5 ± 0.5**
5 picks	Trusting agent	20.4 ± 1.2	18.3 ± 1.1	10.2 ± 0.9	3.2 ± 0.5	1.8 ± 0.4
5 picks	W/o global search objective	27.8 ± 1.3	21.2 ± 1.2	12.8 ± 1.0	4.9 ± 0.6	2.3 ± 0.4
5 picks	HELIOS	**39.2 ± 1.4**	**28.7 ± 1.3**	**17.4 ± 1.1**	**5.8 ± 0.7**	**3.1 ± 0.5**
Unlim.	Trusting agent	21.9 ± 1.2	19.3 ± 1.1	10.8 ± 0.9	3.3 ± 0.5	1.8 ± 0.4
Unlim.	W/o global search objective	29.6 ± 1.3	22.0 ± 1.2	13.2 ± 1.0	5.0 ± 0.6	2.3 ± 0.4
Unlim.	HELIOS	**42.3 ± 1.4**	**30.5 ± 1.3**	**18.6 ± 1.1**	**6.3 ± 0.7**	**3.2 ± 0.5**

additionally report **Place** which indicates if the robot placed the `object` on the `goal_receptacle` and the `object` remained stationary on the `goal_receptacle` after the set wait period. We also report the success rate (**SR**) as defined in the OVMM benchmark – if all of these stages succeeded without collisions, then episode is considered a success.

Baselines and ablations. We evaluate the performance of HELIOS compared to the HomeRobot [4] baseline agents and MoManipVLA [26]. HomeRobot provides modular implementations of the skills required to carry out the OVMM task, we compare to the results for their reported configurations. Additionally, to isolate the effects of our hierarchical scene representation and global search objective, we include the following ablations of our method:

- **Trusting agent**: this agent uses the same 2D maps and methods for local navigation and place as our full method, but without the 3D portion of our hierarchical scene representation, our gaze points and global search objective. It goes to the frontier with the highest value for the `start_receptacle` until it detects an `object` (fully trusting the output of the object detector), at which point it picks up the `object`. If the pick succeeds, it then goes to the frontier with the highest value for the `goal_receptacle` until it detects a `goal_receptacle`, at which point it places the object on it.
- **W/o global search objective**: this agent uses everything from our full method except for the global search objective. Instead, it always prioritizes searching candidate objects over going to frontiers.
- **HELIOS**: our full method, which uses our global search objective to balance when to collect views of a detected `start_receptacle` and when to go to a frontier.

Pick Attempts. In the OVMM benchmark, the agent is allowed an unlimited number of pick attempts. We report results for our method and its ablations with limited numbers of pick attempts (1 and 5) as well as unlimited attempts. With limited pick attempts, if the agent exceeds the limit, we set all metrics for that episode to 0. A benefit of our hierarchical objective is the incorporation of retry logic when we move back and forth between global and local reasoning. In contrast, the baselines do not re-attempt picking. In Table I, we see that allowing 5 pick attempts provides a significant improvement over 1 pick attempt for HELIOS in all metrics. However, the further benefit of unlimited pick attempts is marginal.

Note that the physical process of grasping the object is not modeled during pick attempts in the OVMM benchmark. The pick action only fails when the target object is not in frame, revealing ground truth information about the scene. Thus, our method has access to ground truth information not accessed by the baselines (which in the real world only corresponds to our method making additional observations) when attempting greater than 1 pick attempt.

Results. Table I shows the results of our benchmarking and ablation study. Our full method limited to 1 pick outperforms the baselines on all metrics except for FindObj. Adding our hierarchical scene representation and gaze points improves performance compared to our trusting agent, and adding our global search objective results in further improvement for all metrics. This supports our claims that our hierarchical scene representation and global search objective are beneficial for this task.

The place skill is a major cause of failure for our method. We used a simple approach of dropping the `object` above the highest detected point in a region in front of the agent. Because we did not adjust the orientation of the gripper before dropping, we qualitatively observed that the object sometimes rolled off the the `goal_receptacle`. Due to the modularity of HELIOS, we could incorporate other modular solutions to picking without changing our novel contributions.

B. Semantic object search stop decision

Experiment setting. We investigate the ability of our sparse 3DGS scene representation and associated

TABLE II: **Adaptation of our method as a stop decision for semantic object goal navigation.** We compare to the original implementation of VLFM, as well as a variant that removes the filtering of detections of objects which are at the sides of images which VLFM uses.

Method	SPL ↑	SR ↑
VLFM without detection filtering	29.6	50.4
VLFM [10]	**30.4**	52.5
VLFM with our stop decision	28.4	**54.0**

uncertainty-weighted object score to contribute to robust object detection in the stopping decision during semantic object search. We replace the stop decision in VLFM [10], a leading modular semantic object search method, with our approach. Specifically, when a potential target object is identified the robot goes to generated gaze points around the object and then only stops if the uncertainty-weighted object score is high enough. We evaluate on the validation split of the HM3D dataset [22], consisting of 2000 episodes.

Metrics. Following VLFM [10] we report the Success Rate (SR) and Success weighted by inverse Path Length (SPL). The SPL is a measure of the robot's efficiency, for a successful episode it is given by the ratio of the length of the shortest successful path for that episode to the path the robot took. It is zero for unsuccessful episodes.

Results. Table II shows the results. We can see that using our stop decision improves the success rate by 1.5% compared to VLFM and 3.6% compared to VLFM without detection filtering, the proposed approach in [10] to perform robust object detection. SPL decreases in both cases since our method collects additional views of the objects, prioritizing accuracy over efficiency.

C. Hardware demonstrations

We demonstrate HELIOS on a Boston Dynamics Spot robot in a real-world office environment. In these experiments, we utilize the Spot API to perform grasping and to navigate to the waypoints output by our path planner. We also utilize Bochkovskii2024 for monocular depth estimation. Videos of these demonstrations are provided in the supplementary material.

V. Conclusion

We present HELIOS, a hierarchical scene representation and associated search objective, to perform language-specified pick and place mobile manipulation. HELIOS achieves state-of-the-art results on the Open Vocabulary Mobile Manipulation (OVMM) benchmark [20], [4] and improves the success rate for modular approaches to semantic object search when used as a stop decision. We demonstrate HELIOS performing language-specified pick and place in a real-world office environment with a Spot robot.

Limitations. The performance of HELIOS is limited by errors during execution of subskills including collision avoidance and physical placing which can be improved by integrating better component methods for physical subskills in future work. In addition, since we restrict the total time for executing the pick and place tasks in our work, all of our metrics measure success within a restricted period of time. Therefore, another avenue for increasing performance is by optimizing the choice of gaze points during local search. Filtering for informative gaze points or considering the information gain when generating the gaze points could enable us to achieve improved confidence during local search with fewer total gaze points. Reducing the number of gaze points would allow additional time to enable exploration of more regions in the environment.

Acknowledgments: The authors gratefully appreciate support from the Samsung LEAP-U program and through the following grants: NSF FRR 2220868, NSF IIS-RI 2212433, ONR N00014-22-1-2677.

[1]We use the reported result for their method without GT semantics for a fair comparison. They do not specify which split of the dataset they use for their evaluation so we assume they use the val split as is standard.

References

[1] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia *et al.*, "Open-world object manipulation using pre-trained vision-language models," *CoRL*, 2023.

[2] Physical Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "π0. 5: a vision-language-action model with open-world generalization, 2025," *https://www.physicalintelligence.company/download/pi05.pdf*, 2025.

[3] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl *et al.*, "Gemini robotics: Bringing ai into the physical world," *arXiv preprint arXiv:2503.20020*, 2025.

[4] S. Yenamandra, A. Ramachandran, K. Yadav, A. S. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. Clegg, J. M. Turner, Z. Kira, M. Savva, A. X. Chang, D. S. Chaplot, D. Batra, R. Mottaghi, Y. Bisk, and C. Paxton, "Homerobot: Open-vocabulary mobile manipulation," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=b-cto-fetlz

[5] A. Melnik, M. Büttner, L. Harz, L. Brown, G. C. Nandi, A. PS, G. K. Yadav, R. Kala, and R. Haschke, "Uniteam: Open vocabulary mobile manipulation challenge," *arXiv preprint arXiv:2312.08611*, 2023.

[6] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. D. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable task planning," *CoRR*, 2023.

[7] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint arXiv:2401.12202*, 2024.

[8] D. Honerkamp, M. Büchner, F. Despinoy, T. Welschehold, and A. Valada, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," *IEEE Robotics and Automation Letters*, 2024.

[9] G. Georgakis, B. Bucher, K. Schmeckpeper, S. Singh, and K. Daniilidis, "Learning to map for active semantic goal navigation," *arXiv preprint arXiv:2106.15648*, 2021.

[10] N. H. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfm: Vision-language frontier maps for zero-shot semantic navigation," in *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[11] M. Chang, T. Gervet, M. Khanna, S. Yenamandra, D. Shah, T. Min, K. Shah, C. Paxton, S. Gupta, D. Batra, R. Mottaghi, J. Malik, and D. Singh Chaplot, "GOAT: GO to any thing," 2023.

[12] K. Zheng, A. Paul, and S. Tellex, "Asystem for generalized 3d multi-object search," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1638–1644.

[13] F. Schmalstieg, D. Honerkamp, T. Welschehold, and A. Valada, "Learning hierarchical interactive multi-object search for mobile manipulation," *IEEE Robotics and Automation Letters*, 2023.

[14] Y. Li, Y. Ma, X. Huo, and X. Wu, "Remote object navigation for service robots using hierarchical knowledge graph in human-centered environments," *Intelligent Service Robotics*, vol. 15, no. 4, pp. 459–473, 2022.

[15] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4247–4258, 2020.

[16] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 890–18 900.

[17] J. Ye, D. Batra, A. Das, and E. Wijmans, "Auxiliary tasks and exploration enable objectnav," *ICCV*, 2021.

[18] J. Zhang, L. Dai, F. Meng, Q. Fan, X. Chen, K. Xu, and H. Wang, "3d-aware object goal navigation via simultaneous exploration and identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6672–6682.

[19] B. Yu, H. Kasaei, and M. Cao, "L3mvn: Leveraging large language models for visual target navigation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3554–3560.

[20] S. Yenamandra, A. Ramachandran, M. Khanna, K. Yadav, D. S. Chaplot, G. Chhablani, A. Clegg, T. Gervet, V. Jain, R. Partsey, R. Ramrakhya, A. Szot, T.-Y. Yang, A. Edsinger, C. Kemp, B. Shah, Z. Kira, D. Batra, R. Mottaghi, Y. Bisk, and C. Paxton, "The homerobot open vocab mobile manipulation challenge," in *Thirty-seventh Conference on Neural Information Processing Systems: Competition Track*, 2023. [Online]. Available: https://aihabitat.org/challenge/2023_homerobot_ovmm/

[21] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[22] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," *arXiv preprint arXiv:2109.08238*, 2021.

[23] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on robot learning*. PMLR, 2023, pp. 287–318.

[24] N. Yokoyama, A. Clegg, J. Truong, E. Undersander, T.-Y. Yang, S. Arnaud, S. Ha, D. Batra, and A. Rai, "Asc: Adaptive skill coordination for robotic mobile manipulation," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 779–786, 2023.

[25] R. Shah, A. Yu, Y. Zhu, Y. Zhu, and R. Martín-Martín, "Bumble: Unifying reasoning and acting with vision-language models for building-wide mobile manipulation," *arXiv preprint arXiv:2410.06237*, 2024.

[26] Z. Wu, Y. Zhou, X. Xu, Z. Wang, and H. Yan, "Momanipvla: Transferring vision-language-action models for general mobile manipulation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[27] Q. Wu, Z. Fu, X. Cheng, X. Wang, and C. Finn, "Helpful doggybot: Open-world object fetching using legged robots and vision-language models," *arXiv preprint arXiv:2410.00231*, 2024.

[28] N. Atanasov, B. Sankaran, J. Le Ny, T. Koletschka, G. J. Pappas, and K. Daniilidis, "Hypothesis testing framework for active object detection," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 4216–4222.

[29] W. Ding, N. Majcherczyk, M. Deshpande, X. Qi, D. Zhao, R. Madhivanan, and A. Sen, "Learning to view: Decision transformers for active object detection," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7140–7146.

[30] X. Han, H. Liu, F. Sun, and X. Zhang, "Active object detection with multistep action prediction using deep q-network," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3723–3731, 2019.

[31] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.

[32] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[33] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian splatting slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 039–18 048.

[34] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, "Splatam: Splat track & map 3d gaussians for dense rgb-d slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 357–21 366.

[35] R. Jin, Y. Gao, Y. Wang, Y. Wu, H. Lu, C. Xu, and F. Gao, "Gs-planner: A gaussian-splatting-based planning framework for active high-fidelity reconstruction," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11 202–11 209.

[36] L. Jin, X. Zhong, Y. Pan, J. Behley, C. Stachniss, and M. Popović, "Activegs: Active scene reconstruction using gaussian splatting," *IEEE Robotics and Automation Letters*, 2025.

[37] W. Jiang, B. Lei, K. Ashton, and K. Daniilidis, "Multimodal llm guided exploration and active mapping using fisher information," *ICCV*, 2025.

[38] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang, "Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation," in *European Conference on Computer Vision*. Springer, 2024, pp. 349–366.

[39] Y. Zheng, X. Chen, Y. Zheng, S. Gu, R. Yang, B. Jin, P. Li, C. Zhong, Z. Wang, L. Liu *et al.*, "Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping," *IEEE Robotics and Automation Letters*, 2024.

[40] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi, "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 676–21 685.

[41] J.-C. Shi, M. Wang, H.-B. Duan, and S.-H. Guan, "Language embedded 3d gaussians for open-vocabulary scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5333–5343.

[42] C. Zhang and G. H. Lee, "econsg: Efficient and multi-view consistent open-vocabulary 3d semantic gaussians," *International Conference on Learning Representations*, 2025.

[43] J. Wilson, M. Almeida, M. Sun, S. Mahajan, M. Ghaffari, P. Ewen, O. Ghasemalizadeh, C.-H. Kuo, and A. Sen, "Modeling uncertainty in 3d gaussian splatting through continuous semantic splatting," *arXiv preprint arXiv:2411.02547*, 2024.

[44] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language embedded radiance fields for zero-shot task-oriented grasping," in *7th Annual Conference on Robot Learning*, 2023.

[45] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 815–824.

[46] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 729–19 739.

[47] K. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J. Tenenbaum, C. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba, "Conceptfusion: Open-set multimodal 3d mapping," *Robotics: Science and Systems (RSS)*, 2023.

[48] S. Kobayashi, E. Matsumoto, and V. Sitzmann, "Decomposing nerf for editing via feature field distillation," *Advances in neural information processing systems*, vol. 35, pp. 23 311–23 330, 2022.

[49] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 10 608–10 615.

[50] G. Georgakis, K. Schmeckpeper, K. Wanchoo, S. Dan, E. Miltsakaki, D. Roth, and K. Daniilidis, "Cross-modal map learning for vision and language navigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 460–15 470.

[51] J. Qian, Y. Li, B. Bucher, and D. Jayaraman, "Task-oriented hierarchical object decomposition for visuomotor control," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=hV97HJm7Ag

[52] J. Qian, A. Panagopoulos, and D. Jayaraman, "Recasting generic pretrained vision transformers as object-centric scene encoders for manipulation policies," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 17 544–17 552.

[53] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," in *7th*

Annual Conference on Robot Learning, 2023. [Online]. Available: https://openreview.net/forum?id=9SM6l0HyY_

[54] J. Shi, J. Qian, Y. J. Ma, and D. Jayaraman, "Plug-and-play object-centric representations from "what" and "where" foundation models," in *ICRA*, 2024.

[55] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5021–5028.

[56] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1689–1696.

[57] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," *Robotics: Science and Systems (RSS)*, 2022.

[58] D. Maggio, Y. Chang, N. Hughes, M. Trang, D. Griffith, C. Dougherty, E. Cristofalo, L. Schmid, and L. Carlone, "Clio: Real-time task-driven open-set 3d scene graphs," *IEEE Robotics and Automation Letters*, 2024.

[59] Y. Chang, L. Fermoselle, D. Ta, B. Bucher, L. Carlone, and J. Wang, "Ashita: Automatic scene-grounded hierarchical task analysis," *CVPR*, 2025.

[60] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[61] J. Lin, "On the dirichlet distribution," *Department of Mathematics and Statistics, Queens University*, vol. 40, 2016.

[62] M. Khanna, Y. Mao, H. Jiang, S. Haresh, B. Shacklett, D. Batra, A. Clegg, E. Undersander, A. X. Chang, and M. Savva, "Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 384–16 393.

[63] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[64] S. Garrido, L. Moreno, D. Blanco, and F. Martin, "Fm2: A real-time fast marching sensor-based motion planner," in *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, 2007.

[65] D. S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2020.

[66] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[67] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[68] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European conference on computer vision*. Springer, 2022, pp. 350–368.

[69] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[70] X. Puig, E. Undersander, A. Szot, M. D. Cote, R. Partsey, J. Yang, R. Desai, A. W. Clegg, M. Hlavac, T. Min, T. Gervet, V. Vondrus, V.-P. Berges, J. Turner, O. Maksymets, Z. Kira, M. Kalakrishnan, J. Malik, D. S. Chaplot, U. Jain, D. Batra, A. Rai, and R. Mottaghi, "Habitat 3.0: A co-habitat for humans, avatars and robots," 2023.

[71] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, J. Ye, J. Hu, M. Tancik, and A. Kanazawa, "gsplat: An open-source library for gaussian splatting," *Journal of Machine Learning Research*, vol. 26, no. 34, pp. 1–17, 2025.

VI. Appendix

We provide implementation details of our method including our hyper-parameter choices (Section VI-A), compute details (Section VI-B), an additional ablation using ground truth semantics (Section VI-C) and details including licenses of existing assets we use in this work (Section VI-D).

A. Implementation details

a) Hyperparameters: The value of the hyper-parameters used in our experiments are given in Table III.

b) Once the target object has been detected: First we introduce some new notion, let $\mathcal{G} \subset \mathcal{O}$ be the set of objects whose class is that of the target object and $\mathcal{A} \subset \mathcal{O}$ be the set of objects whose class is that of the place location.

Once a candidate target $g_i \in \mathcal{G}$ has been detected we check if it's uncertainty-weighted object score (given by eq. (7)) is over some threshold τ_g, and if $\Psi_g(g_i) \geq \tau_g$ we will treat g_i as the target object.

If $\Psi_g(g_i) < \tau_g$ we can calculate the class score, $S'(g_i)$, and the uncertainty, $U'(g_i)$, if we take m observations Y from poses P and again assume the best-case scenario that each classified g_i as class g. Then we can obtain the class score this would give us as

$$\Psi'_g(g_i) := S'(g_i) - \alpha_{cs} U'(g_i). \tag{11}$$

When deciding where to obtain additional views, we consider both if obtaining these views could increase the uncertainty-weighted object score to above the threshold and if the increase in uncertainty-weighted object score is larger than a threshold τ_{inc}. This second condition is so that the agent can obtain views of objects which have not been observed much, and so will have a lower uncertainty-weighted object score due to higher uncertainty. If $\Psi'_g(g_i) \geq \min(\tau_g, \Psi_g(g_i) + \tau_{inc})$ then we will obtain the additional observations of g_i, otherwise we return to global search.

After the target object has been grasped, we use the same formulation to decide whether something is the correct class for a place location as we do for deciding whether to grasp a target object but potentially with a different threshold. That is, if we have seen a candidate place location $b_i \in \mathcal{B}$ we first check if $\Psi_b(b_i) \geq \tau_b$ and if so we go there to place the target object, otherwise we check if obtaining additional views satisfies $\Psi'_b(b_i) \geq \min(\tau_b, \Psi_b(b_i) + \tau_{inc})$ and if so we obtain them. If not or if there is no candidate b_i we go to the frontier with the highest value for the place location.

c) Path Planner: We modify the fast marching squared [64] motion planner from Home Robot OVMM's baseline [20] to generate navigation actions from the map and the goal pose. Similar to the baseline, our planner also builds the arrival-time map with velocity directly proportional to the distance from the closest obstacle, which balances the efficiency and safety of the motion plan. However, to account for the fine navigation actions required for mobile manipulation, we make 3 modifications to the baseline: 1. Our planner doubles the resolution of the map at 2000 x 2000 cells of 2.5cm x 2.5cm, as the map is directly derived from the depth observations instead of being predicted through

TABLE III: **Hyperparameters.** We provide a list of the hyper-parameters of our method with a description and the value used in our experiments. Some hyperparameters are only referenced in the supplementary material and not in the main paper.

Name	Description	Value
α_{cs}	Weighting of uncertainty for uncertainty-weighted object score	1
α_d	Weighting of distance term for global search objective	0.001
τ_g	Threshold for uncertainty-weighted object score to pick up an `object`	0.5
τ_b	Threshold for uncertainty-weighted object score to place on a `goal_receptacle`	0.5
τ_{inc}	Minimum change in uncertainty-weighted object score that would cause us to look at an `object` or `goal_receptacle`	0.05
od_a	Threshold for object detector confidence for `start_receptacle` class	0.35
od_g	Threshold for object detector confidence for `object` class	0.25
od_b	Threshold for object detector confidence for `goal_receptacle` class	0.45
cs_a	Class score for an object to be considered a candidate `start_receptacle` for the global search objective	0.3
cs_g	Class score for an object to be considered a candidate `object` for deciding whether to obtain additional views	0.3
cs_b	Class score for an object to be considered a candidate `goal_receptacle` for deciding whether to obtain additional views	0.3
α_{cpa}	Absolute concentration parameter update scaling	3

a neural network as in the baseline [65]. 2. Our planner supports continuous actions of moving forward $[0.1m, 1.0m]$ or rotating $[5°, 30°]$, as opposed to fixed actions of moving forward $0.3m$ or rotating $30°$ from the baseline. 3. Our planner explicitly verifies that all intermediate positions for a forward move are collision-free, greatly improving safety around tighter choke-points common in home environments.

d) Modifications to 3DGS semantic update: We apply a scaling α_{cpa} directly to the concentration parameter update to control the speed of this update, which corresponds to each observation being repeated α_{cpa} times.

e) Additional details of 3DGS instance creation: We spatially cluster gaussians into instances by putting the gaussians in a voxel grid based on the gaussian's center, clustering them by connected components of neighboring voxels, and assigning instance labels to the clusters based on previous assignments. First, we put gaussians of the same semantic label into a grid of 0.5m x 0.5m x 0.5m (adequate due to the spatial sparsity of relevant objects) voxels aligned with the odometry coordinate frame. Then, we take the connected components on the graph of 26-connected voxels containing gaussians. Finally, we assign instance labels to each cluster by taking the minimum of previous instance labels over all gaussians in the cluster. If no gaussian in a cluster previously had an instance label, we assign (maximum instance label over all gaussians) + 1. In practice, this is implemented as a sequence of $\frac{max\_object\_size=10m}{voxel\_size=0.5m} = 20$ min pooling operations on a voxel grid neighborhood graph [66] using the pytorch geometric library [67]. Note we perform the above procedure with only the Gaussians which were updated by the last measurement or which were assigned to the same instance as any of these updated Gaussians.

f) Gaussian creation: We detect when a new observation represents data which is not already part of our scene representation using the depth error. When an observation is taken, we first make a mask of the pixels which have been detected as an object of interest. Within this mask, we calculate the absolute difference between the measured depth and the rendered depth. We then mask this difference again to keep only the parts where the measured depth is over 0. We find the parts of this difference which are over 1m or over 0.001m and remain after an erosion operation, and create a new Gaussian for each of them. Each Gaussian's position is initialized using the measured depth and camera pose to obtain it's 3D location.

g) Re-observing previously detected parts of the scene: As we only model parts of the scene with 3D Gaussians we need to detect when we are re-observing an area which is modeled with 3D Gaussians versus looking towards such an area which is occluded. If we did not do this and only updated the representation when an object is detected then we would not include any negative results (i.e. an object not being detected) and thus we would become over-confident in the classes of objects. One possibility would be to just update if there are any 3D Gaussians in the viewing direction as if they are occluded the new Gaussians should be placed on the occluding object not on the original object, however this is inefficient. Thus we render the depth of our 3D Gaussian scene representation in the viewing direction and then find the pixels in the measured depth image with less than 0.5m of difference to this rendering and finally perform a morphological transformation to close small holes. We then only update the 3D Gaussians using the rendering which lies within this mask.

h) Expanded explanation of how we calculate information gain: When updating the global objective score we use

$$\text{IG}_o(o_i|P, Y^*) := \sum_{\theta_n \in o_i} H(\theta_n) - H(\theta_n|P, Y^*). \quad (12)$$

To obtain Y^*, for each $\theta_n \in o_i$ we create a copy of the associated 3D Gaussian but with the semantic class probabilities set to 1 for the class o and 0 for all other classes, then render using these parameters at pose P – this rendered image is used as Y^*. Then using Y^* we update a copy of the concentration parameters using Eq. 3 and re-calculate the entropy using the updated concentration parameters with Eq. 5 to obtain $H(\theta_n|P, Y^*)$.

TABLE IV: **Ablation study for including ground-truth semantics.** We show the performance increase from using ground-truth semantics (with gt) for both our trusting agent, which does not reason about the uncertainty of object detections, and our full method HELIOS, which does. We show the results for our methods with unlimited picks. We also include results of the recent method MoManipVLA [26] for additional comparison. The standard error of the mean is indicated.

Method	FindObj	Pick	FindRec	Place	SR
MoManipVLA	23.7	12.7	7.1	-	1.7
MoManipVLA with gt	66.1	**62.6**	53.1	-	15.8
Trusting agent	21.9 ± 1.2	19.3 ± 1.1	10.8 ± 0.9	3.3 ± 0.5	1.8 ± 0.4
Trusting agent with gt	57.5 ± 1.4	56.5 ± 1.4	44.7 ± 1.4	20.9 ± 1.2	12.8 ± 1.0
HELIOS	42.3 ± 1.4	30.5 ± 1.3	18.6 ± 1.1	6.3 ± 0.7	3.2 ± 0.5
HELIOS with gt	**66.3 ± 1.4**	58.3 ± 1.4	**53.4 ± 1.4**	**29.8 ± 1.3**	**21.0 ± 1.2**

i) Object detector: We use the DETIC [68] object detector as implemented in the HomeRobot codebase. We set separate thresholds for the detections for each class, with the thresholds for the `object` and `start_receptacle` a bit lower than the default used by HomeRobot (0.45) as our method is designed to filter out false positives but does not address false negatives as shown in Table III.

B. Compute resources

The experiments presented in this paper ran on 8 nodes in a cluster, each with a 2080ti GPU with 16GB of VRAM and 32GB of RAM. Each full run of our method or its ablations on the val split took around 288 hours for 1199 episodes.

C. Ablation using Ground Truth Semantics

We perform an ablation study to show the effect of using ground-truth semantics on performance, the results are shown in Table IV. We can see that our full method our-performs our trusting agent when both use ground truth semantics, this may be due to fact that HELIOS performs local search of detected pick locations whereas our trusting agent doesn't. The gap between the pick success of our trusting agent and our full method is much smaller with ground truth semantics (11.2% without ground truth semantics and 1.8% with ground truth semantics). Likewise, the gap in pick success with and without semantics is much higher for both MoManipVLA and our trusting agent than for HELIOS (49.9% for MoManipVLA, 37.2% for our trusting agent and 27.8% for HELIOS). These results indicate that our full method is less of an improvement when ground truth semantics are used. This makes sense because alleviating issues from imperfect object detections is the main focus of the components of HELIOS which are included in the full method but not in our trusting agent. Addressing this challenge is not necessary when ground truth semantics are provided.

The relatively low overall success rates with ground truth semantics for both MoManipVLA and our method indicate there is still more work required to increase search efficiency and the success rate of physical subskills such as collision-free navigation and place. However the large gap between the results with and without ground truth semantics for MoManipVLA and our trusting agent, especially for the pick skill, still shows that robust object detection is a key bottleneck for this task. While HELIOS still has a performance gap when not using ground truth semantics it takes a step towards addressing this issue.

D. Details of existing assets used

Directly-used assets:

- Home Robot OVMM benchmark and code [4], [20]: MIT License, commit `ede6a67a` (main branch as of submission). https://github.com/facebookresearch/home-robot
- Habitat Synthetic Scenes Dataset (HSSD) [62]: cc-by-nc-4.0, obtained using Home Robot's download script https://huggingface.co/datasets/hssd/hssd-hab
- Habitat [69], [63], [70]: MIT License, for habitat-lab we used HomeRobot's modified code, for habitat-sim we use `v0.2.5`. https://github.com/facebookresearch/habitat-lab https://github.com/facebookresearch/habitat-sim
- VLFM [10]: MIT License https://github.com/bdaiinstitute/vlfm
- gsplat [71]: Apache License 2.0 https://github.com/nerfstudio-project/gsplat
- SplaTAM [34]: BSD 3-Clause License, some code used with modifications rather than directly importing https://github.com/spla-tam/SplaTAM/

Key assets used in above works that we also use:

- BLIP2 [60]: BSD 3-Clause License, v1.0.2 https://github.com/salesforce/LAVIS
- DETIC [68]: Apache License 2.0, installed via Home-Robot https://github.com/facebookresearch/Detic