LEGS-POMDP: Language and Gesture-Guided Object Search in Partially Observable Environments

Ivy Xiao He
Brown University
Providence, USA
xiao_he@brown.edu

Stefanie Tellex
Brown University
Providence, USA
stefie10@cs.brown.edu

Jason Xinyu Liu
Brown University
Providence, USA
xinyu_liu@brown.edu

Abstract

To assist humans in open-world environments, robots must accurately interpret ambiguous instructions to locate desired objects. Foundation model-based approaches excel at reference expression grounding and multimodal instruction understanding, but lack a principled mechanism to model uncertainty in long-horizon tasks. Conversely, Partially Observable Markov Decision Processes (POMDPs) provide a systematic framework for planning under uncertainty but are typically limited in modalities and environment assumptions. To achieve the best of both worlds, we introduce LanguagE and GeSture-Guided Object Search in Partially Observable Environments (**LEGS-POMDP**), a modular POMDP system that integrates language, gesture, and visual observations for open-world object search. Unlike prior work, LEGS-POMDP explicitly models two sources of partial observability: uncertainty over the target object's identity and its spatial location. Simulation results show that multimodal fusion significantly outperforms unimodal baselines, achieving an average success rate of 89% ± 7% across challenging environments and object categories. Finally, we demonstrate the full system on a quadruped mobile manipulator, where real-world experiments qualitatively validate robust multimodal perception and uncertainty reduction under ambiguous human instructions.

CCS Concepts

• **Human-centered computing** → **Pointing**; **Text input**; **Gestural input**; • **Computer systems organization** → *Robotic components*; **Real-time system architecture**.

Keywords

HRI, POMDP, multimodal fusion, gesture, language

Figure 1: Multimodal fusion and belief updates enable disambiguating human instructions and correct object identification.

1 Introduction

To assist humans in unstructured open-world environments, robots must accurately understand and act upon ambiguous instructions to find target objects. The human-instructed object search problem requires robots to both identify which object is being referred to and determine where it is located, under uncertainty arising from underspecified language, imprecise gestures, and noisy perception. As illustrated in Figure 1, language alone may be vague, gestures may indicate regions containing multiple candidates, and sensor noise further compounds ambiguity. A key observation is that different modalities are often complementary: gestures can disambiguate vague language, while language can clarify imprecise gestures. Although humans naturally combine language and gesture during communication, enabling robots to robustly interpret such multimodal cues in partially observable environments remains challenging. To achieve multimodal referring expression understanding [30] in these settings, robots must jointly reason over uncertainty in language, gesture, and visual perception.

Existing works in human-instructed object search are mostly in two main groups, each with fundamental limitations in open-world scenarios. Methods that use foundation models to ground multimodal input and produce actions [23, 28] lack explicit uncertainty modeling and long-horizon sequential decision-making capabilities and offer little formal guarantees and explainability. It is also difficult to collect large-scale datasets of natural human referral gestures [8, 12, 15, 28] for finetuning. Partially Observable Markov Decision Processes (POMDPs) are a formal sequential decision-making framework that explicitly models uncertainty. However, previous POMDP-based object search works mostly focus on tabletop settings [50], use only language input, or make strict assumptions about the environment [46, 52].

To address these limitations, we introduce LanguagE and GeSture-Guided Object Search in Partially Observable Environments (**LEGS-POMDP**), a modular POMDP framework that integrates language, gesture, and visual observations for open-world object search. LEGS-POMDP explicitly models two sources of partial observability: uncertainty over the human's intent (target object identity) and uncertainty over the environment (target object location). By maintaining joint beliefs over object identity and location, the robot can reason over both instruction-level and environment-level ambiguity and produce explainable decision-making behavior.

Our multimodal observation model leverages state-of-the-art language, gesture, and visual perception modules to represent each modality as a likelihood function over candidate objects, which are fused in log-space to form a joint observation distribution. The modular design enables flexible replacement or upgrading of individual perception components, while preserving principled Bayesian belief updates and interpretability that are difficult to achieve with end-to-end approaches.

We evaluate LEGS-POMDP through both modular reference understanding benchmarks and full-system decision-making experiments to assess multimodal object search under uncertainty. Specifically, our evaluation includes: (i) gesture grounding with five different pointing representations; (ii) visual grounding via Set-of-Marks prompting [47] and Grounding DINO [26]; (iii) visual sensor models with different fan-shaped configurations to test modularity and parameterization; (iv) full-system evaluation in simulated environments with varying levels of complexity and instruction ambiguity; and (v) real-robot evaluation on a quadruped mobile manipulator.

This paper makes three key contributions: (1) We formulate human-instructed object search as a POMDP with two sources of partial observability, explicitly modeling uncertainty over target object identity and spatial location. (2) We propose a modular multimodal observation model that integrates language, gesture, and visual perception as probabilistic likelihoods within a principled Bayesian belief update. (3) We evaluate the proposed framework through extensive simulation experiments under varying levels of instruction ambiguity, and qualitatively validate uncertainty reduction on a real quadruped mobile manipulator.

2 Related Work

Human-instructed object search is challenging because the robot must handle state uncertainty, perceptual noise, and reference ambiguity.[49] Prior research has followed two main paradigms: end-to-end learning based methods and modular approaches. End-to-end methods map multimodal sensor inputs directly to actions, learning semantic priors that support goal-directed exploration and generalization [4, 12, 22, 36]. While some approaches introduce intermediate structure, such as visual state abstractions or topological representations, end-to-end learning remains highly data-intensive and often requires large-scale training[10, 14, 36].

Modular approaches, on the other hand, decompose the task into perception, semantic grounding, and planning components.[11, 45] This structure facilitates engineering, preserves explainability, and enables changeable modules. Recent modular learning approaches further replace hand-designed components with learned ones while retaining the overall pipeline, thereby combining the data efficiency and sim-to-real transfer benefits of modularity with the representational power of learning.[6, 9, 11] Building on this paradigm, researchers have extensively explored gesture grounding, language grounding, and multimodal fusion, as well as decision-theoretic frameworks for planning under uncertainty.

Language Grounding: Grounding natural language commands has long been a central challenge for robots [7, 42]. Previous works have grounded human instructions to a formal representation and latent space for planning and control [1–3, 13, 16, 21, 24, 25, 35, 36]. Interactive approaches such as INGRESS [39] and attribute-guided POMDP frameworks [48] show that asking clarification questions can mitigate linguistic ambiguity. Other work has embedded language directly into observation models [32], or leveraged social feedback to reduce misinterpretations in object fetching tasks [46]. Spatial language understanding in large-scale environments further highlights how ambiguity grows when many candidate objects are present [51]. Our framework extends this literature by jointly modeling language ambiguity alongside gesture uncertainty in a unified probabilistic planning framework.

Gesture Grounding: Pointing is a natural and frequent modality in human–robot interaction, often used to resolve referential ambiguity. Early work formalized pointing with geometric models such as the pointing cone [18, 33, 34], while later studies analyzed human pointing behaviors in household settings [8, 19], highlighting the prevalence of ambiguity. More recent approaches incorporate skeletal vectors (eye–wrist, shoulder–wrist) to probabilistically model gesture likelihoods [37], and integrate pointing into situated language understanding [38]. With the rise of large models, systems like GIRAF [23] and GestLLM [17] emphasize the semantic and contextual nature of gesture interpretation, while visual prompting methods leverage pointing for downstream VQA [41]. Despite this progress, gesture interpretation remains inherently uncertain due to human variability and sensor noise. Our work builds on this line by explicitly modeling gesture as a probabilistic observation within a POMDP, rather than as a deterministic cue, allowing the robot to reason probabilistically about human intent.

Multimodal Fusion: Many systems integrate gesture and language at the perception level to improve disambiguation. while visual prompting methods use pointing for VQA [29, 41]. These efforts demonstrate that multimodal cues can significantly reduce referential ambiguity; however, fusion is typically confined to the perceptual level and is not connected to downstream decision-making. Systems such as GIRAF [23] and This&That [44] show promising integration of multimodal instructions with robot execution, but their reliance on tabletop domains highlights the need for frameworks that extend to unstructured, large-scale environments. Complementary work has introduced benchmarks targeting perception challenges, such as open-vocabulary segmentation [20, 53], or multimodal disambiguation datasets that probe gesture and language integration [5, 15, 28, 31]. These directions highlight the importance of multimodal fusion for instruction following, while also pointing to the need for frameworks that connect multimodal fusion with downstream tasks.

POMDP & Uncertainty in HRI: POMDPs provide a principled framework for decision-making under uncertainty, with online solvers such as POMCP [40] enabling scalability. They have been
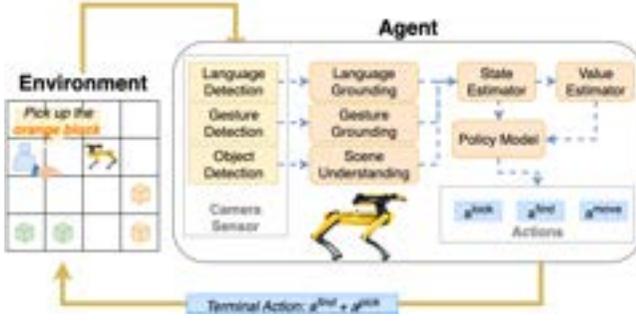
Figure 2: system diagram.

applied to multi-object search [43], where language serves as a prior over candidate states, and extended in systems such as Gen-MOS [52] and INVIGORATE [50] that integrate visual grounding and interactive dialogue. Yet prior POMDP-based work has largely focused on language cues and has rarely incorporated gesture as a probabilistic observation[43, 46, 50]. Our work advances this trajectory by explicitly modeling two layers of uncertainty, i.e., human intent and environment state, and showing that multimodal fusion of gesture and language improves efficiency in object search.

3 Technical Approach

Our system aims to address the language- and gesture-conditioned object search problem, where the robot must interpret uncertain human instructions while exploring a partially observed environment. This requires solving the subproblems of: (i) representing robot's uncertainty and the hidden state of the world, (ii) integrating multimodal human instructions, and (iii) planning efficiently under ambiguous instruction and robot uncertainty. This problem can be naturally modeled as a Partially Observable Markov Decision Process (POMDP), since the target object's location is hidden, human inputs are noisy and ambiguous, and the robot must plan sequences of actions under uncertainty. Figure 2 illustrates the overall architecture of our framework, showing how language, gesture, and vision modules are integrated with the POMDP-based planner.

3.1 POMDP Formulation

In order to represent both the robot's knowledge of the world and its uncertainty about the hidden target, we define the task as a POMDP tuple $(S, A, T, O, Z, R, \gamma)$ as follows:

State Space: Each state $s \in S$ is defined as $s = (s_r, s_o)$, where $s_r = (x, y, \theta)$ is the robot pose, and s_o denotes the latent target location. The obstacle map is known and fixed. We use an object-independent state representation, where objects are labeled as target or distractor based on human intent, rather than category, allowing the framework to focus on uncertainty reasoning rather than object taxonomy.

Action Space: The action space is discrete. The agent choose from three classes of actions: movement actions a_{move}, an observation-gathering action a_{look}, and a termination action a_{find}. Move actions

use four deterministic motion primitives (forward, backward, turn-left, turn-right) defined in the robot's relative frame. This abstraction allows flexible combinations of primitives without changing the POMDP formulation.

Transition Model: The transition model $T(s' \mid s, a)$ updates the robot pose deterministically. For movement actions, $T(s' \mid s, a_{move}) = 1$, where robot pose is updated based on the executed primitive. The look action is designed solely to acquire additional multimodal observations for belief update. The target-object location s_o is static and remains unchanged across transitions.

Observation Space: Observations $o = (o_v, o_g, o_l)$ contain multimodal signals from vision, gesture and language instructions.

Observation Model: $Z(o \mid s)$ defines the likelihood of multimodal signals conditioned on the hidden state, with vision, gesture, and language terms fused by a weighted log-likelihood.

Reward Model: Reward $R(s, a)$ assigns a positive reward for a correct a_{find}, a small negative cost for a_{move} and a_{look} to encourages efficient exploration. This sparse structure ensures that the planner prioritizes correct target finding, while step costs discourage exhaustive exploration. Discount factor $\gamma \in (0, 1)$ balances immediate and future rewards.

3.2 Multimodal Observation Model

In order to integrate human instructions with perceptual signals, we design an observation model that fuses three modalities: vision, language, and gesture. Unlike end-to-end models, this modular observation formulation is less data-hungry and provides interpretable likelihoods for each modality, enabling explicit reasoning about uncertainty and explainability in downstream planning. Each modality is modeled by its likelihood $P(o_m \mid s)$, representing the probability of observing signal o_m given a hypothesized state s. The contribution of each modality can be controlled by modality-specific weights (w_v, w_g, w_l). The modality-specific likelihoods are combined by the fusion model:

$$\log Z(o \mid s) = w_v \log P_v(o_v \mid s) + w_g \log P_g(o_g \mid s) + w_l \log P_l(o_l \mid s), \tag{1}$$

Our observation model can be formulated as:

$$Z(o \mid s) \propto \prod_{m \in \{v, g, l\}} P_m(o_m \mid s)^{w_m}. \tag{2}$$

This formulation naturally integrates with Bayesian belief updates as shown in Eq.3, where multimodal likelihoods provide evidence to reweight the posterior over hidden states.

$$b'(s') \propto Z(o \mid s') \sum_{s \in S} T(s' \mid s, a) \, b(s). \tag{3}$$

Visual Observation. Camera sensors provide incomplete and noisy detections of objects due to limited field of view and distance-dependent accuracy. To approximate this uncertainty, segmentation outputs are treated as candidate object detections, and the camera is modeled as a decaying fan-shaped sensor, similar to [43]. The likelihood of correctly detecting a target at location (x, y) is defined by Gaussian decay in both angular deviation and range:

$$P_v(o_v = 1 \mid s) \propto \exp\left(-\frac{\theta_{diff}^2}{2\sigma_\theta^2}\right) \cdot \exp\left(-\frac{(r - r_0)^2}{2\sigma_r^2}\right), \tag{4}$$

where θ_{diff} is the angular difference between the camera's central axis and the object direction, r is the distance from the robot to the object, and r_0 is the nominal detection range. This formulation captures the intuition that detections are most reliable when the object is centered in view and within a favorable distance, and less reliable otherwise.

Language Observation. Natural language instructions describe the target object but can be ambiguous and error-prone after automatic speech recognition. For example, the same object may be referred to as "cup," or "mug," and transcription errors may further increase uncertainty. Thus, the challenge is to map an utterance u_l into a probabilistic signal that reflects how well each candidate state location s_o matches the instruction. We model this through a similarity function $\kappa(s_o; u_l) \in [0, 1]$, which measures how well the hypothesized target at s_o aligns with the given instruction. This score is then converted into a likelihood by interpolating between modality-specific false- and true-positive rates:

$$P_l(o_l \mid s) = \epsilon_l^- + (\epsilon_l^+ - \epsilon_l^-)\,\kappa(s_o; u_l), \qquad 0 < \epsilon_l^- < \epsilon_l^+ < 1. \quad (5)$$

ϵ_l^- is the minimum likelihood assigned to irrelevant objects (false positives), and ϵ_l^+ is the maximum likelihood for a perfectly matching description (true positives). This formulation captures graded confidence rather than a binary match, allowing the belief update to weigh language input proportionally to its semantic specificity.

Gesture Observation. Pointing gestures provide a strong cue about the intended target, but they are inherently uncertain due to human variability and perceptual noise. Humans adopt different pointing strategies: sometimes extending the whole arm, sometimes aligning the gaze with the hand, and sometimes raising only the forearm casually.

To account for this variability, we define the pointing direction dynamically as the mean vector of multiple anatomical cues: eye-to-wrist, shoulder-to-wrist, and elbow-to-wrist. The gesture is then represented as a spatial cone with the wrist as the origin and this averaged vector as the central axis. The opening angle of the cone captures the spread of the three vectors. The likelihood of the target being at location (x, y) is then defined as

$$P_g(o_g \mid s) = \exp\left(-\frac{\theta_{\mathrm{diff}}^2}{2\sigma_g^2}\right), \qquad (6)$$

where θ_{diff} is the angular deviation between the central pointing vector and the vector from the wrist to (x, y), and σ_g determines the spread of the cone. This formulation captures the intuition that states closer to the pointing direction are more likely, while off-cone states receive exponentially lower likelihood.

For planning, we employ Partially Observable UCT (PO-UCT) as the solver in both simulation and real-world tests. PO-UCT is a Monte Carlo tree search algorithm that balances exploration and exploitation by simulating trajectories from the current belief. Although not novel in itself, PO-UCT provides a strong, well-established baseline that integrates naturally with our multimodal observation models and supports deployment on the Boston Dynamics Spot robot. This unified choice allows us to attribute performance differences to perception and grounding quality, rather than planning artifacts.

To enable systematic evaluation, we first implement explicit probabilistic observation models in simulation, including a fan-shaped vision sensor, a gesture cone model, and a language similarity mapping. In real-robot experiments, however, the robot directly consumes outputs from perception pipelines: skeleton tracking for gesture estimation (via MediaPipe), a Set-of-Marks (SoM) grounding module combining SAM2 segmentation with GPT-4o reasoning for language input, and onboard object detection from the Spot camera system. This design ensures that the POMDP framework accommodates both analytic likelihoods in simulation and perceptual modules in deployment.

4 Evaluation of LEGS-POMDP

We evaluate whether the LEGS-POMDP framework enables robust and efficient multimodal object search under uncertainty. The evaluation proceeds in three stages: (i) modular tests of gesture and language grounding, (ii) gridworld simulations comparing solvers, modalities, and environment complexity, and (iii) real-robot deployment on the Boston Dynamics Spot. Across these settings, results show that multimodal grounding improves robustness, PO-UCT enhances planning reliability, and the integrated system achieves strong performance in both simulation and real-world.

4.1 Modular Evaluation

We evaluate gesture and language grounding to examine how each modality resolves referential ambiguity. Using the YouRefIt dataset [5], which contains 4,221 annotated pointing and language instances, we benchmark each modality in isolation. The results highlight complementary strengths and limitations, motivating the integration in LEGS-POMDP for robust multimodal grounding.

*4.1.1 **Gesture Grounding.*** Formalized gesture representations provides structured likelihoods that can be directly incorporated into a POMDP observation model, yielding both probabilistic reasoning and interpretability for downstream belief updates. We evaluate gesture grounding to test whether different pointing representations can robustly capture human intent and identify which representation provides the most reliable basis for integration into a decision-making framework. We hypothesize that using a probabilistic cone representation with cues from multiple skeletal landmarks yields improved robustness and accuracy compared to single-vector baselines, especially in the presence of pose estimation noise.

We compare four body landmark vectors (eye-to-wrist, nose-to-wrist, shoulder-to-wrist, and elbow-to-wrist) with a gesture cone representation that merges vector cues to form a probabilistic region of reference. All pointing vectors are anchored at the wrist, which provides a stable and consistently detectable landmark in dynamic scenes. While finer hand landmarks could in principle yield more precise estimates, hand detection is often less reliable under occlusion or motion, making the wrist a more robust endpoint for downstream analysis. MediaPipe [27] is used for skeleton detection, achieving a 92.6% human detection rate on the YouRefIt dataset; all evaluation is conditioned on detected frames. Fig. 3 shows an example frame with different vector- and cone-based pointing representations, with the target object highlighted in green.

Performance is evaluated using two metrics: (i) *Coverage Accuracy @25%* is defined as the percentage of samples where the
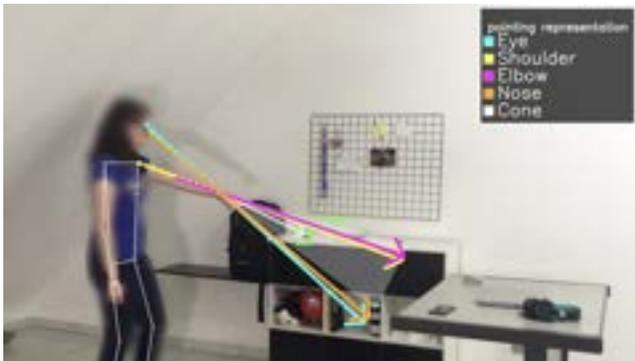
Figure 3: Example frame showing different vector- and cone-based models of the pointing direction, with the target marked in green.

Table 1: Comparison of gesture representation for pointing estimation for pointing estimation. Metrics are reported as mean ± 95% confidence interval (CI).

Pointing Representation	Cov. @25%	$\theta_{\mathbf{diff}}$ (°)
Eye-to-Wrist	0.718 ± 0.014	24.4 ± 0.8
Nose-to-Wrist	0.746 ± 0.014	23.2 ± 0.8
Shoulder-to-Wrist	0.865 ± 0.011	17.0 ± 0.7
Elbow-to-Wrist	0.772 ± 0.013	20.2 ± 0.8
Gesture Cone	**0.890 ± 0.010**	**14.4 ± 0.4**

predicted cone overlaps more than 25% of the ground-truth bounding box area. For single-vector representations, we used a 15° fixed opening angle, while for the gesture cone the opening angle is dynamically determined from the spread of included vectors. (ii) *Angular Error* (θ_{diff}), the deviation in degrees between the predicted pointing direction and the ground-truth reference.

Results in Table 1 show that the gesture cone achieves the lowest angular error (14.4°) and the highest coverage accuracy (0.89). Compared to the best single-vector baseline (shoulder-to-wrist, 0.865), it improves coverage by 2.5% and reduces angular error by 2.6°. In contrast, single-vector representations are more sensitive to target height and arm posture. We also observed qualitative differences across pointing representations. For low-lying targets, the elbow-to-wrist vector was often the most accurate, but it tended to overshoot for elevated targets. Wrist flexion or extension also altered pointing reliability. In cluttered scenes, the nose-to-wrist vector sometimes provided better disambiguation. Moreover, under frontal-facing condition, gaze-based vectors occasionally diverged from arm-based vectors, making generalization with single vector representation difficult. By averaging cues, the gesture cone stabilizes performance across varied conditions, making it a more reliable representation for downstream POMDP belief updates.

4.1.2 ***Visual Grounding***. Robust language grounding is essential for downstream decision-making in the POMDP framework since grounding failure leads to corrupted belief state, which in turn biases future planning toward incorrect targets. Thus, we evaluate visual grounding to test how well different grounding strategies resolve referential expressions. Our hypothesis is that decoupling

Figure 4: Visual grounding comparison between SoM prompting and a detector baseline (GroundingDINO).

Table 2: Grounding Success rate under different language conditions. LLM grounding (SoM: SAM2 + GPT-4o) vs. detector grounding (GroundingDINO).

Grounding Acc	LLM (SoM)	Detector (DINO)
None	0.793	0.603
Spatial	0.957	0.577
Attribute	0.944	0.665
Spatial + Attribute	0.818	0.956

perception and reasoning via a two-stage SoM pipeline (segmentation + LLM classification) yields more accurate and interpretable grounding than end-to-end detector-based methods.

We conduct a comparative analysis between two distinct visual grounding paradigms: a detector-based baseline using GroundingDINO and an LLM-based Set-of-Marks (SoM) approach that integrates SAM2 segmentation with GPT-4o reasoning capabilities as shown in Fig.4. Queries may involve object attributes and/or spatial relations, allowing us to evaluate grounding robustness across different linguistic conditions. Performance is evaluated using three complementary metrics: (i) *Detection Accuracy* (det. Acc) measures whether the system successfully localizes any candidate region for the queried object, providing a recall-oriented view of localization; (ii) *IoU@25%*; (iii) *Grounding Accuracy* (Grounding Acc) evaluates whether the predicted region correctly overlaps the GT target, reflecting semantic correctness of the grounding.

The Set-of-Marks (SoM) approach achieved higher detection success (92.3% vs. 87.8%) and grounding accuracy (91.4% vs. 62.4%) compared to the detector-based baseline. However, its IoU@25% was lower (0.219 vs. 0.501), largely because SAM2 produced fine-grained masks smaller than the annotated bounding boxes, leading to underestimated overlap. This highlights that SoM is limited by its dependence on segmentation quality. In addition, SoM incurs significantly higher inference time due to its two-stage pipeline, trading efficiency for robustness in resolving referring expressions.

Results in Table 2 report grounding accuracy conditioned on successful detection. We observe that the detector-based baseline struggles in most cases: when only spatial (0.577) or attribute (0.665) references are provided, accuracy drops sharply in cluttered scenes. As illustrated in Fig. 4, the SoM approach (a) correctly grounds the query "the sign on the wall" by isolating the intended region. In contrast, the detector baseline (b) detects a sign but misinterprets the spatial qualifier "on the wall," incorrectly selecting a lower

sign. The detector performs relatively well when both spatial and attribute cues are combined (0.956), suggesting that it leverages explicit learned patterns from training data when richer descriptions are available. In contrast, the SoM approach maintains consistently high performance across single-reference conditions, achieving 0.957 on spatial and 0.944 on attribute queries. However, SoM accuracy decreases with compounded descriptions (0.818), likely due to segmentation ambiguities and language model parsing errors. The results indicate that LLM-based SoM grounding generalizes more robustly to underrepresented linguistic conditions, while the detector benefits more from detailed but less natural multi-cue descriptions. Importantly, SoM's robustness in handling single but ambiguous references provides more stable observation likelihoods, reducing the risk of belief corruption in downstream POMDP planning.

4.2 System Evaluation

We evaluate our system in a gridworld simulation environment designed to capture the challenges of multimodal instruction following. The environment consists of grid cells populated with target objects, distractors, and static obstacles, requiring the agent to actively explore while maintaining a belief state over possible target locations. We prepared three grid environments of increasing spatial complexity (5×5, 10×10, 20×20). Human inputs (gesture, language, or both) are injected as observations that directly influence belief updates, while distractors and obstacles introduce ambiguity and navigation cost. This setup allows us to systematically vary environment size and ambiguity, and to test how different modalities and solvers affect success, efficiency, and belief convergence.

4.2.1 Solver Comparison. Solver comparison test determines whether sophisticated planning algorithms with principled belief representations enhance robustness in ambiguous visual grounding tasks compared to simple approaches. Four solvers are compared under a consistent observation model and reward function. The *Greedy* baseline always executes the Find action as soon as any object is observed, ignoring uncertainty and planning. The *Belief Heuristic* policy moves toward the grid cell with the highest current belief, considering only the most likely target location at each step. *POMCP* is a Monte Carlo Tree Search–based solver that leverages a particle belief representation for scalable online planning. Finally, *PO-UCT* extends the UCT algorithm with deeper lookahead, balancing exploration and exploitation to improve planning under uncertainty. We hypothesize that principled POMDP solvers (POMCP and PO-UCT) will demonstrate more robust performance compared to heuristic baselines (Greedy and Belief Heuristic).

We conducted controlled experiments with no human input to isolate planning performance from perceptual challenges. Approximately 100 independent trials per solver-belief representation configuration is executed to ensure statistical reliability, with random initialization of object locations and agent starting positions. Performance was evaluated using three metrics: (i) *Success rate*, the fraction of trials in which the agent correctly identified and executed a Find action on the target object; (ii) *Total steps*, mean number of actions required until task completion, measuring exploration effectiveness; and (iii) *Total time*, total execution time including both planning overhead and action execution.

Table 3: Solver performance under histogram vs. particle belief. Metrics are mean ± 95% CI over all trials.

Belief	Solver	Success	Steps	Time [s]
Histogram	Heuristic	0.68 ± 0.11	111.3 ± 24.0	12.0 ± 2.6
	Greedy	0.63 ± 0.11	227.7 ± 20.4	24.6 ± 2.2
	PO-UCT	0.96 ± 0.06	124.9 ± 14.7	32.2 ± 8.3
Particles	Heuristic	0.21 ± 0.10	42.3 ± 7.9	4.7 ± 0.9
	Greedy	0.27 ± 0.10	183.6 ± 18.4	20.5 ± 2.1
	POMCP	0.24 ± 0.09	183.4 ± 20.2	36.8 ± 4.5
	PO-UCT	0.45 ± 0.11	121.4 ± 10.3	30.6 ± 6.3

Table 4: Modality comparison. Success rate, steps, and time are reported with 95% confidence intervals.

Modality	Success Rate	Steps	Time [s]
multimodal conflicted	0.024 ± 0.007	59.0 ± 17.3	21.1 ± 12.8
wrong language	0.093 ± 0.064	95.9 ± 35.2	24.5 ± 10.8
wrong gesture	0.170 ± 0.049	91.2 ± 33.1	22.8 ± 11.1
No Input	0.482 ± 0.130	162.3 ± 35.2	37.0 ± 10.4
Gesture	0.618 ± 0.045	122.5 ± 25.5	23.2 ± 4.7
Language	0.710 ± 0.057	95.8 ± 34.4	20.4 ± 5.4
Multimodal	0.888 ± 0.073	76.8 ± 27.4	16.7 ± 5.6

Table 3 demonstrates that PO-UCT achieves optimal performance under histogram belief representation, achieving 96% success rate while maintaining competitive step counts and reasonable computational overhead. The Greedy baseline exhibits poor success rates despite minimal planning time, while POMCP shows inconsistent performance with higher variance in both success and efficiency metrics. Under particle belief representation, all solvers experience degraded performance due to increased representational noise and approximation errors in belief updates. However, PO-UCT maintains the smallest performance degradation, suggesting greater robustness to belief representation quality. The results validate our hypothesis that planning depth and stable belief representation are essential for reliable performance in ambiguous visual grounding tasks. While heuristic approaches offer computational efficiency, they sacrifice reliability. PO-UCT emerges as the optimal balance between robustness and efficiency.

4.2.2 Modality Evaluation. Human inputs directly influence the POMDP observation model; grounding failures or missing modalities can therefore corrupt belief updates. We evaluate how individual modalities guide exploration and how multimodal fusion improves robustness under ambiguity. We fix the solver to PO-UCT and evaluate across five environments (small, small-ambiguous, medium, large, and large-ambiguous) under seven instruction conditions: no input, gesture-only, language-only, multimodal, wrong gesture, wrong language, and conflicted multimodal input. Performance is measured by success rate, steps to completion, and total time, with belief dynamics analyzed via max-belief and target-belief convergence over 10 trials per condition.

As shown in Table 4, multimodal input achieves the highest success rate (0.888 ± 0.073) with the fewest steps and fastest completion time, validating the complementarity of gesture and language.
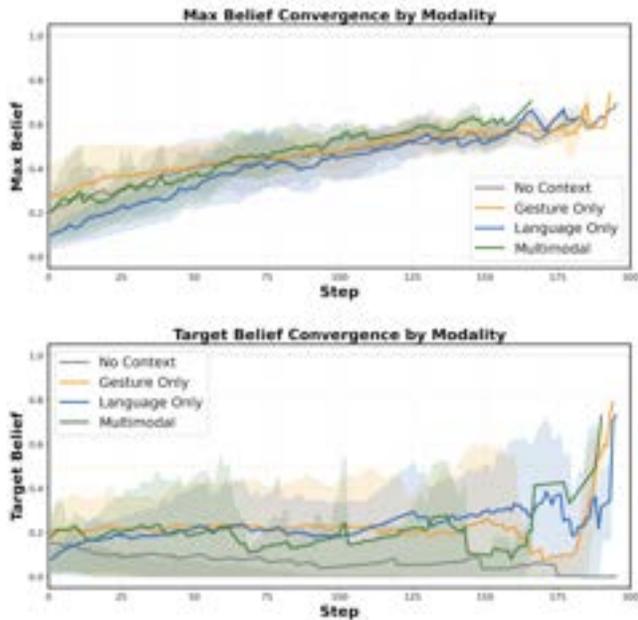
Figure 5: Belief convergence in the **large environment**. (Top) Max-belief traces show how certainty in the most likely state evolves over time. (Bottom) Target-belief traces show probability mass assigned to the true target.

While this result is not surprising, it demonstrates that our POMDP-based approach and solver is capable of successfully fusing information across multiple modalities to achieve improved performance. Language-only (0.710 ± 0.057) and gesture-only (0.618 ± 0.045) perform moderately, while no-instruction drops sharply to 0.482 ± 0.130. Time usage shows a similar trend, with multimodal trials completing in 16.7 seconds on average, nearly half of the no-instruction condition. These results highlight the complementary effect of gesture and language, showing that combining modalities not only improves task success but also improves efficiency. In contrast, wrong gesture (0.170 ± 0.049), wrong language (0.093 ± 0.064), and especially conflicted multimodal input (0.024 ± 0.007) nearly always fail, highlighting how erroneous inputs corrupt the belief state.

In the large environment, belief convergence further illustrates the advantages of multimodal input. As shown in Fig. 5, max-belief curves across modalities grow at similar rates, but multimodal trials terminate in fewer steps, reflecting faster convergence and decision-making efficiency. Target-belief curves reveal a sharper contrast: without instruction, belief quickly collapses toward distractors, while any valid human input yields sustained growth in target belief. These findings confirm that multimodal guidance mitigates the challenges of large, ambiguous environments by stabilizing belief updates and accelerating task completion.

Figure 6 shows how environment complexity directly impacts grounding performance. As the environment becomes larger and more ambiguous, performance of single-modality instructions degrades sharply. The no-instruction baseline collapses almost entirely in these cases. In contrast, multimodal input maintains relatively high success even under severe ambiguity, demonstrating
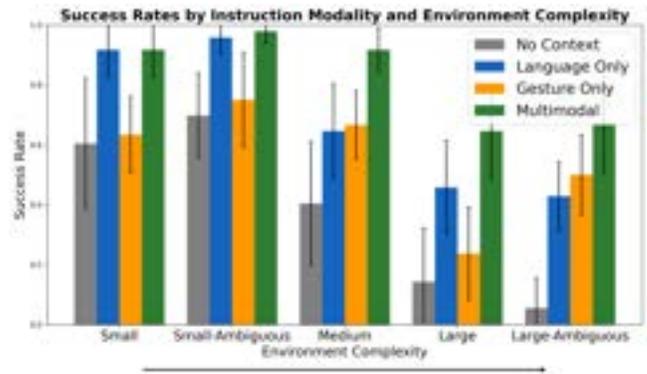

Figure 6: Success rates by instruction modality across environments of increasing complexity.

that combining gesture and language provides robustness against increasing environmental complexity. By jointly analyzing solvers, modalities, and environment complexity, we establish that multimodal POMDP planning is both more robust and more efficient in ambiguous search settings. This also demonstrates where an end-to-end approach trained on datasets would fall short in the large ambiguous environment because it would not be able to systematically search. In the future we plan to investigate end-to-end models with memory that can generalize in this way.

4.3 Robot Testing

The goal of this evaluation is to test whether the LEGS-POMDP framework transfers from simulation to a real-world platform, and how different modalities influence belief uncertainty under realistic conditions. One advantage of our modular approach is the ability to easily transfer between different robot hardware and different environments without collecting additional data or retraining, since each module is already trained on internet-scale data.

To evaluate whether multimodal input accelerates disambiguation, we conducted an ablation study in a 10×10 grid world with five objects, including three identical red cups placed in different locations to induce ambiguity. Without allowing the robot to execute move actions, we measured how belief uncertainty changed over 10 observation steps(A_{LOOK}). As shown in Fig. 8, the multimodal (G+L) condition achieves the steepest entropy reduction rate, reducing entropy by 60.8%. Gesture alone also contributes strongly, achieving a 40.6% reduction, while unimodal visual and language conditions reduce entropy by 30.1% and 34.2%, respectively. These results indicate that both gesture and multimodal inputs more effectively narrow the prior belief compared to language and vision baselines. We further validated this trend by demonstrating successful execution of the object search task on the robot.

5 Conclusion and Future Work

We presented LEGS-POMDP, a multimodal POMDP framework for gesture- and language-conditioned object search under uncertainty. Simulation and modular evaluations show that multimodal fusion consistently outperforms single-modality baselines with an average success rate of 89% in challenging simulated environments. On a real quadruped mobile manipulator, we demonstrated the
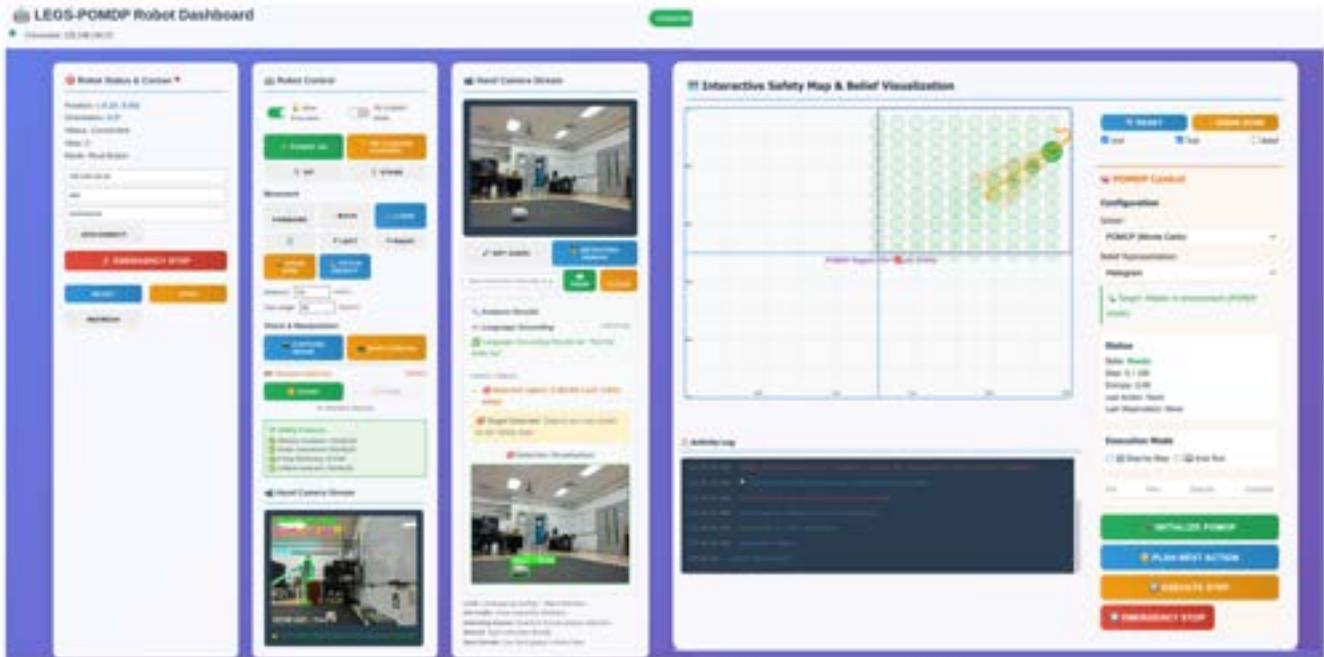
Figure 7: LEGS-POMDP robot testing and dashboard interface. The figure illustrates both the real-robot experimental setup and the integrated UI, which visualizes multimodal grounding, POMDP belief states, and robot control in real time.
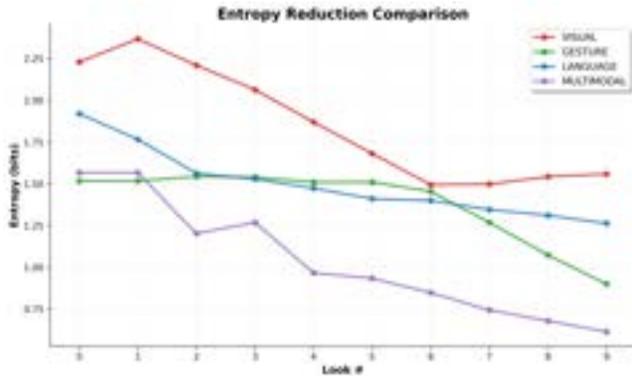

Figure 8: Entropy loss curve.

feasibility of the proposed framework by qualitatively validating multimodal grounding and uncertainty reduction in physical settings. Multimodal fusion substantially improves disambiguation in human-instructed object search, particularly under high angular ambiguity. Cone-based gesture likelihoods capture spatial intent, while SoM-based language grounding provides semantic specificity. Weighted log-likelihood fusion enables more robust belief updates and faster convergence than unimodal alternatives.

Several limitations remain. Our fusion model assumes conditional independence between modalities, simplifying belief updates but ignoring potential correlations such as alignment between deictic language and pointing gestures. Our system also relies on accurate visual segmentation, where errors may degrade perception and downstream belief updates in cluttered or dynamic environments.

While we demonstrate feasibility in simulation and on a real robot, the scale and diversity of real-world experiments remain limited.

Future work will explore richer multimodal integration, including tactile input and additional gesture types such as iconic gestures, as well as user studies in naturalistic environments to better understand how non-expert users employ multimodal communication during collaborative object search. These directions aim to support more natural, robust, and adaptive human–robot interaction in open-world settings.

Acknowledgments

References

[1] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. 2024. π0: A vision-language-action flow model for general robot control. CoRR, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164* (2024).

[2] Anthony Brohan, Noah Brown, et al. 2023. RT-1: Robotics Transformer for Real-World Control at Scale. In *Robotics: Science and Systems*.

[3] Anthony Brohan, Noah Brown, et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning*.

[4] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. 2020. Object Goal Navigation using Goal-Oriented Semantic Exploration. arXiv:2007.00643 [cs.CV] https://arxiv.org/abs/2007.00643

[5] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. YouRefIt: Embodied Reference Understanding with Language and Gesture. arXiv:2109.03413 [cs.CV] https://arxiv.org/abs/2109.03413

[6] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burch-fiel, and Shuran Song. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*.

[7] Vanya Cohen, Jason Xinyu Liu, Raymond Mooney, Stefanie Tellex, and David Watkins. 2024. A Survey of Robotic Language Grounding: Tradeoffs between Symbols and Embeddings. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[8] Bartłomiej Filipek, Marcin Banach, Olga Franć, Jan Zguda, and Dominik Belter. 2025. An Empirical Study on Pointing Gestures Used in Communication in Household Settings. *Electronics* 14, 12 (2025), 2346.

[9] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. 2024. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. arXiv:2401.02117 [cs.RO] https://arxiv.org/abs/2401.02117

[10] Sourav Garg, Krishan Rana, Mehdi Hosseinzadeh, Lachlan Mares, Niko Sünder-hauf, Feras Dayoub, and Ian Reid. 2024. RoboHop: Segment-based Topological Map Representation for Open-World Visual Navigation. arXiv:2405.05792 [cs.RO] https://arxiv.org/abs/2405.05792

[11] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. 2024. GAPartNet: Cross-Category Domain-Generalizable Object Perception and Manipulation via Generalizable and Actionable Parts. *arXiv preprint arXiv:2303.04137v5* (2024).

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

[13] Eric Hsiung, Hiloni Mehta, Junchi Chu, Xinyu Liu, Roma Patel, Stefanie Tellex, and George Konidaris. 2022. Generalizing to new domains by mapping natural language to lifted LTL. In *IEEE International Conference on Robotics and Automation*.

[14] Nathan Hughes, Yun Chang, Siyi Hu, Rajat Talak, Rumaia Abdulhai, Jared Strader, and Luca Carlone. 2024. Foundations of spatial perception for robotics: Hierarchical representations and real-time systems. *The International Journal of Robotics Research* 43, 10 (2024), 1457–1505.

[15] Md. Mofijul Islam, Reza Mirzaiee, Alexi Gladstone, Haley N. Green, and Tariq Iqbal. 2022. CAESAR: An Embodied Simulator for Generating Multimodal Referring Expression Datasets. *Neural Information Processing Systems* (2022).

[16] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024).

[17] Oleg Kobzarev, Artem Lykov, and Dzmitry Tsetserukou. 2025. GestLLM: Advanced Hand Gesture Interpretation via Large Language Models for Human-Robot Interaction. doi:10.48550/arXiv.2501.07295 arXiv:2501.07295 [cs].

[18] Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2006. Deixis: How to Determine Demonstrated Objects Using a Pointing Cone. In *Gesture in Human-Computer Interaction and Simulation*, Sylvie Gibet, Nicolas Courty, and Jean-François Kamp (Eds.). Vol. 3881. Springer Berlin Heidelberg, Berlin, Heidelberg, 300–311. doi:10.1007/11678816_34 Series Title: Lecture Notes in Computer Science.

[19] Tymon Kukier, Alicja Wróbel, Barbara Sienkiewicz, Julia Klimecka, Antonio Galiza Cerdeira Gonzalez, Paweł Gajewski, and Bipin Indurkhya. 2025. An Empirical Study on Pointing Gestures Used in Communication in Household Settings. *Electronics* 14, 12 (2025). doi:10.3390/electronics14122346

[20] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. 2023. Semantic-SAM: Segment and Recognize Anything at Any Granularity. doi:10.48550/arXiv.2307.04767 arXiv:2307.04767 [cs].

[21] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as Policies: Language Model Programs for Embodied Control. In *IEEE International Conference on Robotics and Automation*.

[22] Hsiu-Chin Lin, Soshi Iba, and Fabio Ramos. 2020. Multi-Level Structure vs. End-to-End-Learning in High-Performance Tactile Robotic Manipulation. In *Conference on Robot Learning (CoRL)*. PMLR, 516.

[23] Li-Heng Lin, Yuchen Cui, Yilun Hao, Fei Xia, and Dorsa Sadigh. 2023. Gesture-Informed Robot Assistance via Foundation Models. doi:10.48550/arXiv.2309.02721 arXiv:2309.02721 [cs].

[24] Jason Xinyu Liu, Ankit Shah, George Konidaris, Stefanie Tellex, and David Paulius. 2024. Lang2LTL-2: Grounding Spatiotemporal Navigation Commands Using Large Language and Vision-Language Models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[25] Jason Xinyu Liu, Ziyi Yang, Ifrah Idrees, Sam Liang, Benjamin Schornstein, Stefanie Tellex, and Ankit Shah. 2023. Grounding Complex Natural Language Commands for Temporal Tasks in Unseen Environments. In *Conference on Robot Learning (CoRL)*.

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*. Springer, 38–55.

[27] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. doi:10.48550/arXiv.1906.08172 arXiv:1906.08172 [cs].

[28] Atharv Mahesh Mane, Dulanga Weerakoon, Vigneshwaran Subbaraju, Sougata Sen, Sanjay E. Sarma, and Archan Misra. 2025. Ges3ViG: Incorporating Pointing Gestures into Language-Based 3D Visual Grounding for Embodied Reference Understanding. arXiv:2504.09623 [cs.CV] https://arxiv.org/abs/2504.09623

[29] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. 2022. Point and Ask: Incorporating Pointing into Visual Question Answering. doi:10.48550/arXiv.2011.13681 arXiv:2011.13681 [cs].

[30] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 11–20.

[31] Shu Nakamura, Yasutomo Kawanishi, Shohei Nobuhara, and Ko Nishino. 2023. DeePoint: Visual Pointing Recognition and Direction Estimation. arXiv:2304.06977 [cs.CV] https://arxiv.org/abs/2304.06977

[32] Thao Nguyen, Vladislav Hrosinkov, Eric Rosen, and Stefanie Tellex. 2023. Language-Conditioned Observation Models for Visual Object Search. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Detroit, MI, USA, 10894–10901. doi:10.1109/IROS55552.2023.10341492

[33] Kai Nickel and Rainer Stiefelhagen. 2003. Pointing gesture recognition based on 3D-tracking of face, hands and head orientation. In *Proceedings of the 5th International Conference on Multimodal Interfaces*. ACM, 140–146.

[34] Kai Nickel and Rainer Stiefelhagen. 2007. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing* 25, 12 (2007), 1875–1884.

[35] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. 2024. Octo: An Open-Source Generalist Robot Policy. In *Robotics: Science and Systems*.

[36] Abby O'Neill, Abdul Rehman, et al. 2024. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *IEEE International Conference on Robotics and Automation*.

[37] Madeline Helmer Pelgrim, Ivy Xiao He, Kyle Lee, Falak Pabari, Stefanie Tellex, Thao Nguyen, and Daphna Buchsbaum. 2024. Find it like a dog: Using Gesture to Improve Object Search. *Proceedings of the Annual Meeting of the Cognitive Science Society* 46, 0 (2024). https://escholarship.org/uc/item/0nk6w9fd

[38] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. 2016. Situated Language Understanding with Human-like and Visualization-Based Transparency. In *Robotics: Science and Systems XII*. Robotics: Science and Systems Foundation. doi:10.15607/RSS.2016.XII.040

[39] Mohit Shridhar and David Hsu. 2018. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. doi:10.48550/arXiv.1806.03831 arXiv:1806.03831 [cs].

[40] David Silver and Joel Veness. 2010. Monte-Carlo planning in large POMDPs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 2* (Vancouver, British Columbia, Canada) *(NIPS'10)*. Curran Associates Inc., Red Hook, NY, USA, 2164–2172.

[41] Kosei Tanada, Shigemichi Matsuzaki, Kazuhito Tanaka, Shintaro Nakaoka, Yuki Kondo, and Yuto Mori. 2024. Pointing Gesture Understanding via Visual Prompting and Visual Question Answering for Interactive Robot Navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*. https://openreview.net/forum?id=sJjwtGvK5D

[42] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots That Use Language. *Annual Review of Control, Robotics, and Autonomous Systems* 3 (3 May 2020), 25–55. doi:10.1146/annurev-control-101119-071628 Publisher Copyright: Copyright © 2020 by Annual Reviews. All rights reserve..

[43] Arthur Wandzel, Yoonseon Oh, Michael Fishman, Nishanth Kumar, Lawson L.S. Wong, and Stefanie Tellex. 2019. Multi-Object Search using Object-Oriented POMDPs. In *2019 International Conference on Robotics and Automation (ICRA)*. 7194–7200. doi:10.1109/ICRA.2019.8793888 ISSN: 2577-087X.

[44] Boyang Wang, Nikhil Sridhar, Chao Feng, Mark Van der Merwe, Adam Fishman, Nima Fazeli, and Jeong Joon Park. 2024. This&That: Language-Gesture Controlled Video Generation for Robot Planning. doi:10.48550/arXiv.2407.05530 arXiv:2407.05530 [cs].

[45] Zihan Wang, Xiangyu Yang, Jiahao Liang, Jing Xu, Yuehu Luo, Zhiming Yang, Haojian Zhang, Xiaoyu Hu, and Yandong Wang. 2024. Sim-to-Real Transfer via 3D Feature Fields for Vision-and-Language Navigation. *arXiv preprint arXiv:2406.09798* (2024).

[46] David Whitney, Eric Rosen, James MacGlashan, Lawson L. S. Wong, and Stefanie Tellex. 2017. Reducing errors in object-fetching interactions through social feedback. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 1006–1013. doi:10.1109/ICRA.2017.7989121

[47] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. doi:10.48550/arXiv.2310.11441 arXiv:2310.11441 [cs].

[48] Yang Yang, Xibai Lou, and Changhyun Choi. 2022. Interactive Robotic Grasping with Attribute-Guided Disambiguation. doi:10.48550/arXiv.2203.08037 arXiv:2203.08037 [cs].

[49] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. 2024. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*.

IEEE, 42–48.

[50] Hanbo Zhang, Yunfan Lu, Cunjun Yu, David Hsu, Xuguang Lan, and Nanning Zheng. 2024. INVIGORATE: Interactive Visual Grounding and Grasping in Clutter. doi:10.48550/arXiv.2108.11092 arXiv:2108.11092 [cs].

[51] Kaiyu Zheng, Deniz Bayazit, Rebecca Mathew, Ellie Pavlick, and Stefanie Tellex. 2021. Spatial Language Understanding for Object Search in Partially Observed City-scale Environments. doi:10.48550/arXiv.2012.02705 arXiv:2012.02705 [cs].

[52] Kaiyu Zheng, Anirudha Paul, and Stefanie Tellex. 2023. A System for Generalized 3D Multi-Object Search. doi:10.48550/arXiv.2303.03178 arXiv:2303.03178 [cs].

[53] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2023. Segment Everything Everywhere All at Once. doi:10.48550/arXiv.2304.06718 arXiv:2304.06718 [cs].