MIGU: Multimodal Instruction Grounding under Uncertainty for Manipulation Planning

Mingke Lu*, Anxing Xiao*, David Hsu

***Abstract*— Understanding natural human instructions is crucial for deploying robots in human-centric environments. We study multimodal instruction grounding, where language and gesture provide complementary but uncertain cues. We present MIGU, a modular framework that combines semantic and geometric evidence into a unified grounding belief and connects it to manipulation planning. MIGU constructs a 3D geometric likelihood by propagating viewing-direction and depth uncertainty through eye–finger geometry while accounting for hand-direction estimation error. A vision–language model (VLM) provides semantic priors over candidate objects and regions, which are combined with the geometric likelihood through Bayes-inspired fusion. The resulting belief supports behavior planning to either proceed directly to downstream planning or request clarification. Grounded targets then define goals for mobile manipulation and tabletop task-and-motion planning. On a real-world benchmark, MIGU outperforms all evaluated baselines, while ablations support the benefit of explicit multimodal uncertainty modeling. Project website: multimodal-instruction.github.io**

I. INTRODUCTION

As robots move from structured environments into human-centric spaces such as homes, hospitals, and offices, their role shifts toward assisting and collaborating with people. This requires robots to understand human intent from natural multimodal instructions, including verbal cues and non-verbal gestures such as pointing [1]. However, pointing in the real world is inherently imprecise due to sensing noise and natural variation in human behavior, meaning that the intended object may not lie exactly along the pointing direction [2], [3]. Humans can still infer the intended target by combining pointing with language and visual context and seek clarification when uncertainty remains. How can we equip robots with the same capability? In this work, we study *uncertainty-aware multimodal instruction grounding* by fusing semantic priors from vision and language with a geometric likelihood derived from pointing geometry and integrating the resulting belief with manipulation planning.

Multimodal instruction grounding remains challenging in open-world deployments. Early learning-based methods jointly learn from labeled multimodal data [4–7], but often generalize poorly out of distribution. With advances in perception models, recent work has adopted more explicit decompositions of language and gesture [8–11], yet often relies on ad hoc pointing models, struggles with open-ended expressions and continuous-region grounding, and rarely connects seamlessly to manipulation planning. Recent VLMs show strong visual grounding through direct prediction [12], [13] or visual prompting [14–16], but when directly applied to multimodal grounding, they still exhibit limited spatial reasoning and can be overconfident.

Fig. 1. MIGU fuses language cues with pointing geometry for robot manipulation under uncertainty, asking clarification questions when needed.

In this work, we explore an explicit and modular approach that exploits the complementary strengths of semantic and geometric cues. We introduce MIGU (**M**ultimodal **I**nstruction **G**rounding under **U**ncertainty), a framework that combines the geometric information provided by pointing gestures with semantic priors derived from foundation models to obtain a unified belief over objects and regions. For pointing, we derive a principled 3D geometric likelihood that captures observation, model-fitting, and eye–pointing mismatch uncertainty, yielding a closed-form uncertainty estimate. For semantic grounding, we study several VLM-based grounding strategies and develop a simple yet effective semantic belief estimator. The geometric and semantic beliefs are then fused through Bayes-inspired fusion, producing a unified belief that can be directly used by planning modules. We implement an attribute-based behavior planner to balance clarification and direct execution. After clarification, the updated belief is passed to downstream planning, as shown in Fig. 1.

We evaluate our approach extensively in real-world multimodal instruction grounding scenarios with diverse environments, objects, referring expressions, and pointing behaviors. Our method outperforms existing baselines, while analysis and ablation studies show that explicitly separating semantic and geometric beliefs provides a more robust basis for multimodal grounding. We further demonstrate how the resulting grounding belief can be integrated into real-world robotic

* denotes equal contribution, alphabetical order

Mingke Lu, Anxing Xiao, and David Hsu are with the School of Computing and Smart Systems Institute, National University of Singapore, Singapore. Correspond to anxingxiao@u.nus.edu.

Mingke Lu is also with the Department of Electrical and Computer Engineering, University of California, Los Angeles, Los Angeles, CA, USA.

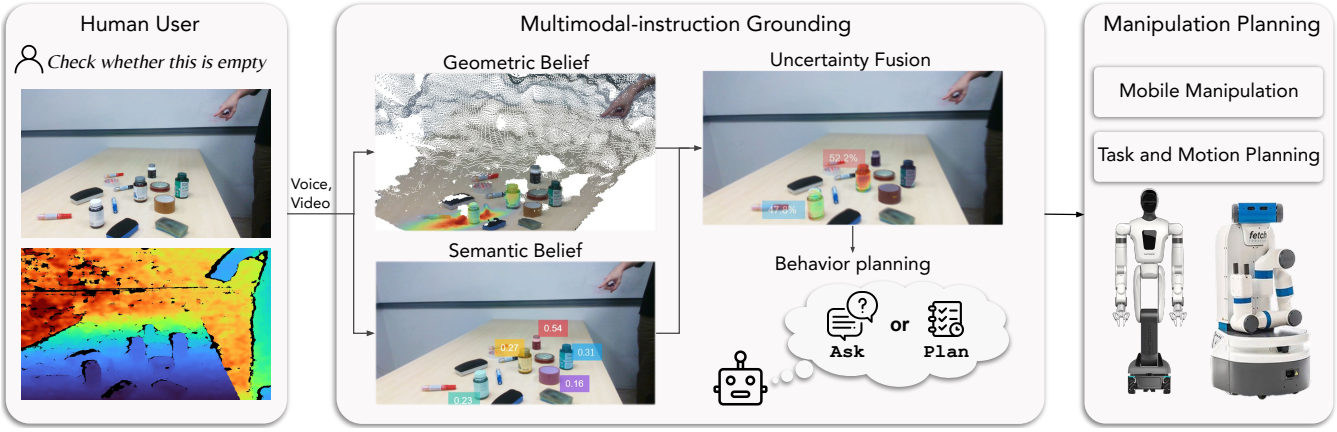

Fig. 2. **Framework Overview.** MIGU integrates geometric and semantic evidence into a grounding belief, which is then used for downstream planning.

systems, including mobile manipulation and tabletop task-and-motion planning.

II. Related Work

A. Visual Grounding with Language

Visual grounding connects language instructions to entities in the observed scene [17]. Probabilistic approaches explicitly model uncertainty over candidate referents [18], [19], while language-aligned visual representations and open-vocabulary detectors broaden the range of recognizable concepts [20–22]. Vision-language models (VLMs) [12], [13] further introduce commonsense reasoning into visual grounding, often through direct prediction or visual prompting techniques [14–16]. However, 3D gesture-grounding data are scarce, and learned models may yield unreliable beliefs at inference time. For robotic applications, interactive visual grounding has also been explored [23–25], enabling robots to actively resolve ambiguity through clarification questions or take additional actions [26] when needed. Inspired by visual grounding, we adopt a similar belief-based formulation that combines semantic and visual evidence from a VLM with geometric pointing likelihoods in a modular framework. Our method maintains a grounding belief over object and surface references and requests clarification when necessary.

B. Gesture-Informed Grounding

Pointing helps resolve spatial references, although its interpretation varies with gesture style [3] and context [27]. Geometric methods estimate pointing directions [28], [29] and associate them with candidate objects using rays or cones [2], [30], [31]. Probabilistic models further capture directional uncertainty from body landmarks in 3D [32] or 2D [11]. However, such landmarks may be partially or entirely outside the robot camera's field of view. End-to-end methods [4–7] directly infer multimodal grounding, but often generalize poorly out of distribution and offer limited explicit 3D reasoning. Another line of work focuses on modular fusion of language and pointing, for example by combining object-category evidence with geometric pointing cues for relatively simple instructions [10], [33]. LEGS-POMDP [11] combines language-target semantic similarity with a probabilistic 2D pointing-cone model. With the rise of foundation models, recent systems use LLMs and VLMs to interpret multimodal instructions and generate executable code [1], [8]. iChores [9] uses an LLM for instruction interpretation and fuses object detections with a ray-distance-based pointing model [34]. Our approach more fully exploits VLM commonsense and semantic priors while deriving a more principled model of pointing uncertainty, yielding a more accurate grounding belief that benefits behavior planning and downstream manipulation.

III. Problem Formulation

We consider a robot receiving a multimodal instruction $\mathcal{I} = (L, O)$, where L is a language instruction and O is an RGB-D observation containing a pointing gesture. The grounding problem can be formulated as estimating the spatial belief $p(x \mid L, O)$ over candidate target locations $x \in \mathbb{R}^3$. We assume that the intended target is represented in the estimated scene. For a set of candidate object and region hypotheses $\mathcal{H}$, the corresponding belief over hypotheses is $b(h) = p(h \mid L, O)$, with $h \in \mathcal{H}$. The robot is equipped with action primitives $\mathcal{A} = \{a_1, \ldots, a_K\}$ and clarification queries $\mathcal{Q}$, each of which elicits a binary response. Given the grounding belief, it must decide whether to plan and execute the actions or request clarification from the user. The objective is to maximize expected task reward while accounting for clarification costs, subject to the symbolic and geometric constraints required for physical execution.

IV. Methods

A. Overview

MIGU converts a language instruction L and an RGB-D observation O into a grounded task goal for robot execution. As shown in Fig. 2, we first model geometric uncertainty to estimate the likelihood of the observed pointing direction for each candidate 3D target. In parallel, a vision-language model identifies candidate objects and regions and assigns semantic confidence scores. Bayes-inspired fusion combines these cues into a belief over hypotheses and locations. A belief-based rule decides whether to execute or request clarification, balancing task reward and interaction cost. The selected target then defines the downstream planning goal.

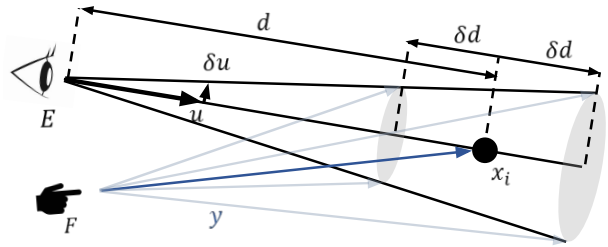

Fig. 3. **Gesture Uncertainty Modeling.** Eye–finger geometry propagates viewing-direction and depth uncertainty into the pointing model.

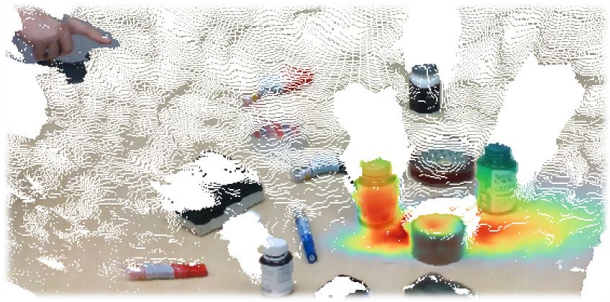

Fig. 4. **Illustration of Geometric Likelihood over 3D Scene Points.**

B. Geometric Uncertainty Modeling

We model gesture uncertainty as arising from the mismatch between viewing and pointing directions, as illustrated in Fig. 3. We assume an average constant eye–hand offset and model how target-perception uncertainty in the viewing direction propagates to the pointing direction. Let $E \in \mathbb{R}^3$ denote the eye position, and let $x_i \in \mathbb{R}^3$ denote the i-th 3D point in the scene. We define the nominal viewing distance and unit viewing direction as

$$d_i := \|x_i - E\|_2, \qquad u_i := \frac{x_i - E}{d_i} \in \mathbb{S}^2, \tag{1}$$

where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^3$. We model uncertainty in the viewing direction locally on the tangent plane $T_{u_i}\mathbb{S}^2 := \{v \in \mathbb{R}^3 : u_i^\top v = 0\}$ and depth uncertainty along the corresponding viewing ray:

$$u \approx u_i + \delta u,\ \delta u \sim \mathcal{N}\left(0, \sigma_u^2 P_i\right),\ P_i := I_3 - u_i u_i^\top \tag{2}$$

$$d = d_i + \delta d, \quad \delta d \sim \mathcal{N}\left(0, \sigma_{d,i}^2\right), \quad \sigma_{d,i} := \alpha_d d_i, \tag{3}$$

where I_3 is the 3×3 identity matrix, $P_i \in \mathbb{R}^{3\times 3}$ is the orthogonal projector onto $T_{u_i}\mathbb{S}^2$, $\sigma_u \geq 0$ controls the angular uncertainty, and $\alpha_d \geq 0$ is the relative depth-noise coefficient. Consequently, $u_i^\top \delta u = 0$ almost surely, and the covariance $\sigma_u^2 P_i$ represents a rank-two Gaussian distribution supported on the tangent plane. We assume $\delta u \perp\!\!\!\perp \delta d$. Based on the visual-localization standard deviations reported by Odegaard et al. [35] and the relative standard deviation of visual distance judgments reported by Dukes et al. [36], we set $\sigma_u = 0.044$ rad and $\alpha_d = 0.04$, respectively.

Let $F \in \mathbb{R}^3$ denote the fingertip position, and $y := E + du - F$ denote the unnormalized finger-to-target vector. A first-order approximation around (d_i, u_i) gives

$$y \mid x_i \dot{\sim} \mathcal{N}(\bar{y}_i, \Sigma_{y,i}), \qquad \bar{y}_i = x_i - F, \tag{4}$$

$$\Sigma_{y,i} = \sigma_{d,i}^2 u_i u_i^\top + d_i^2 \sigma_u^2 \left(I_3 - u_i u_i^\top\right), \tag{5}$$

Fig. 5. **Generated semantic beliefs.** Left: an ideal prior. Right: a biased prior generated by the VLM.

where $\dot{\sim}$ denotes equality in distribution under the first-order approximation. For $x_i \neq F$, let $r_i := \|\bar{y}_i\|_2$, $\mu_i := \bar{y}_i / r_i \in \mathbb{S}^2$ denote the nominal finger-to-target distance and direction, respectively. The Jacobian of the normalization map $h(y) = y/\|y\|_2$, evaluated at $\bar{y}_i$, is J_i. The corresponding first-order approximation of $h(y)$ around $\bar{y}_i$ is

$$h(y) \approx \mu_i + J_i(y - \bar{y}_i), \quad J_i = \frac{1}{r_i}\left(I_3 - \mu_i \mu_i^\top\right). \tag{6}$$

Let $B_i = [b_{i,1}, b_{i,2}] \in \mathbb{R}^{3\times 2}$ be an orthonormal basis of the tangent plane $T_{\mu_i}\mathbb{S}^2$, satisfying $B_i^\top B_i = I_2$, and $B_i^\top \mu_i = 0$. The geometrically induced covariance C_i^{geo} and the overall covariance C_i, both expressed in tangent coordinates, are

$$C_i^{\text{geo}} = B_i^\top J_i \Sigma_{y,i} J_i^\top B_i = \frac{1}{r_i^2} B_i^\top \Sigma_{y,i} B_i, \tag{7}$$

$$C_i = C_i^{\text{geo}} + B_i^\top \Sigma^{\text{fit}} B_i, \tag{8}$$

where $\Sigma^{\text{fit}} \in \mathbb{R}^{3\times 3}$ is the covariance of the hand-direction estimation error, assumed independent of the geometric uncertainty, and is projected onto $T_{\mu_i}\mathbb{S}^2$ and combined with the geometric covariance.

The discrepancy between the observed direction and the candidate-induced direction is evaluated using the logarithmic map on $\mathbb{S}^2$. Let $g^{\text{obs}} \in \mathbb{S}^2$ denote the observed unit pointing direction estimated from the detected hand landmarks. For $\theta_i = \arccos\left(\mu_i^\top g^{\text{obs}}\right)$, the logarithmic map is

$$\text{Log}_{\mu_i}\left(g^{\text{obs}}\right) = \frac{\theta_i}{\sin\theta_i}\left(g^{\text{obs}} - \cos\theta_i\, \mu_i\right). \tag{9}$$

The corresponding two-dimensional tangent residual is $z_i = B_i^\top \text{Log}_{\mu_i}\left(g^{\text{obs}}\right)$. The gesture likelihood is then approximated by a Gaussian distribution in the tangent plane:

$$p\left(g^{\text{obs}} \mid x_i\right) \approx \mathcal{N}(z_i; 0, C_i) = \frac{\exp\left(-\frac{1}{2} z_i^\top C_i^{-1} z_i\right)}{2\pi |C_i|^{1/2}}. \tag{10}$$

C. Semantic Uncertainty Modeling and Uncertainty Fusion

We decompose the multimodal observation O into the visual scene observation O^{vis} and the observed pointing direction g^{obs}. Given L and O^{vis}, a VLM is prompted to identify N candidate bounding boxes, corresponding to the hypothesis set $\mathcal{H} = \{h_1, \ldots, h_N\}$. For each hypothesis $h_i \in \mathcal{H}$, the VLM provides a bounding box $\mathcal{B}_i \subset \mathbb{R}^2$ and a confidence score $c_i \in [0, 1]$. We make a practical design choice to avoid explicit calibration of such confidence, as it requires a huge calibration dataset. The VLM jointly reasons over the instruction and scene to infer referents, attributes, affordances, and object roles. We use GPT-5.6 Sol for its state-of-the-art visual reasoning capabilities. Each

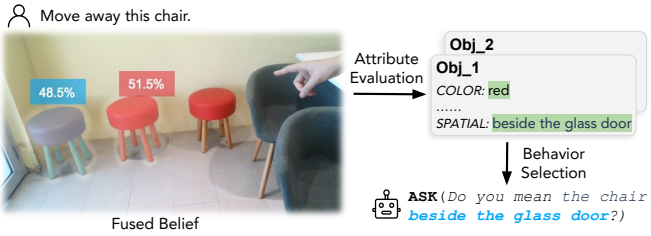

Fig. 6. **Behavior Planning.** Attribute-based clarification distinguishes competing targets when the grounding belief remains ambiguous.

bounding box $\mathcal{B}_h$ is provided as a spatial prompt to a class-agnostic segmentation model, which produces the corresponding pixel-level mask M_h. Using the RGB-D correspondence, M_h defines a set of associated 3D scene points $\mathcal{X}_h := \{x_i \in \mathbb{R}^3 \mid \pi(x_i) \in M_h\}$, where $\pi(\cdot)$ denotes projection into the image plane. The VLM confidence induces the following unnormalized semantic score over candidate–point pairs:

$$s_{\text{sem}}(x_i, h \mid L, O^{\text{vis}}) \propto c_h \, \mathbb{I}[x_i \in \mathcal{X}_h], \tag{11}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Thus, the semantic confidence of a hypothesis is distributed over all 3D points contained in its mask, without prematurely selecting a particular point. We assume that the semantic evidence and the observed gesture are conditionally independent given the intended 3D target point, i.e., $p(g^{\text{obs}} \mid x_i, h, L, O^{\text{vis}}) = p(g^{\text{obs}} \mid x_i)$. Motivated by the Bayesian factorization implied by this conditional- independence assumption, we combine the gesture likelihood and semantic score into the following unnormalized pointwise fused score:

$$\begin{aligned} q_h(x_i) &\propto p(g^{\text{obs}} \mid x_i)\, s_{\text{sem}}(x_i, h \mid L, O^{\text{vis}}) \\ &\propto p(g^{\text{obs}} \mid x_i)\, c_h \, \mathbb{I}[x_i \in \mathcal{X}_h]. \end{aligned} \tag{12}$$

For each hypothesis, we aggregate its pointwise fused scores using max pooling, retaining the point that best agrees with the observed gesture. The resulting hypothesis scores are normalized to obtain an approximate grounding belief:

$$\tilde{b}(h) := \max_{x_i \in \mathcal{X}_h} q_h(x_i), \qquad b(h) := \frac{\tilde{b}(h)}{\sum_{h' \in \mathcal{H}} \tilde{b}(h')}. \tag{13}$$

D. Behavior Planning Under Uncertainty

We formulate the execute-or-clarify choice as a one-step belief-space decision problem, equivalent to a one-step POMDP. To exclude hypotheses with negligible probability, we retain the smallest set of highest-probability hypotheses whose cumulative belief is at least 90%, and renormalize the belief over the retained set. For simplicity, we continue to denote the resulting candidate set and belief by $\mathcal{H}$ and b, respectively. The robot first considers direct execution of the maximum-a-posteriori hypothesis, whose expected value is:

$$V_{\text{exec}}(b) = b(h^*) R_{\text{success}}, \quad h^* = \arg\max_{h_i \in \mathcal{H}} b(h_i), \tag{14}$$

where $R_{\text{success}} > 0$ denotes the reward for executing the task with the correct grounding hypothesis. Alternatively, the robot may request clarification using a binary attribute-based query. We define a fixed attribute vocabulary $\mathcal{C}$ covering visual appearance and spatial relations, such as color, size, shape, and relative position. At runtime, the VLM evaluates these attributes for each candidate hypothesis $h_i \in \mathcal{H}$, as visualized in Fig. 6. Each attribute that partitions the candidate set into two nonempty subsets defines a query $q \in \mathcal{Q}$, together with the binary predicate $\phi_q(h_i) \in \{0, 1\}$, which equals one if hypothesis h_i satisfies the attribute queried by q, and zero otherwise. Attributes already specified in the instruction are excluded to avoid redundancy. Let $y \in \mathcal{Y} = \{\text{yes}, \text{no}\}$ denote the user's response. Assuming that the response is deterministic and consistent with the intended hypothesis, its likelihood is

$$p(y \mid h_i, q) = \begin{cases} \phi_q(h_i), & y = \text{yes}, \\ 1 - \phi_q(h_i), & y = \text{no}. \end{cases} \tag{15}$$

The response likelihood under the current belief and the corresponding posterior belief are

$$p(y \mid b, q) = \sum_{i=1}^{N} p(y \mid h_i, q) b(h_i), \tag{16}$$

$$b^{q,y}(h_i) = \frac{p(y \mid h_i, q) b(h_i)}{\sum_{j=1}^{N} p(y \mid h_j, q) b(h_j)}. \tag{17}$$

Since only one clarification round is allowed, the robot executes immediately after receiving the response. Let $C_{\text{query}}(q) \geq 0$ denote the interaction cost of issuing query q, which may account for communication effort and execution delay. The value of query q is therefore

$$V_{\text{query}}(b, q) = -C_{\text{query}}(q) + \sum_{y \in \mathcal{Y}} p(y \mid b, q) V_{\text{exec}}(b^{q,y}). \tag{18}$$

The optimal query is $q^* = \arg\max_{q \in \mathcal{Q}} V_{\text{query}}(b, q)$, and clarification is requested iff $V_{\text{query}}(b, q^*) > V_{\text{exec}}(b)$. The selected query is phrased in the context of the original instruction. Otherwise, the robot directly executes h^*. This one-step formulation is intentionally lightweight for real-time interaction and excludes multi-round dialogue.

E. System Integration

We extract pointing keyframes from synchronized RGB-D images (timestamp offset $\leq$ 80 ms) using MediaPipe Hands [37]. Index-finger landmarks 5–8 must form projected segments of at least two pixels, with adjacent direction cosine similarity ≥ 0.75. Fingertip speed must remain ≤ 80 pixels/s for three consecutive observations, with continuity reset after a 0.5 s gap or invalid hand detection. Landmarks are back-projected using calibrated intrinsics and median valid depth in a 5×5 pixel neighborhood; observations with invalid depth are discarded. Whisper [38] transcribes the corresponding language instruction. After behavior planning, the grounded object and region become fixed arguments of the downstream planning goal. We describe two integrations here.

Mobile Manipulation. For mobile picking and placing, the grounded targets specify the object and destination, and we adopt a similar strategy to [39]. For grasping, we generate 12 heuristic grasp poses and sample candidate base poses from an inverse reachability map [40]. For each base–grasp pair,

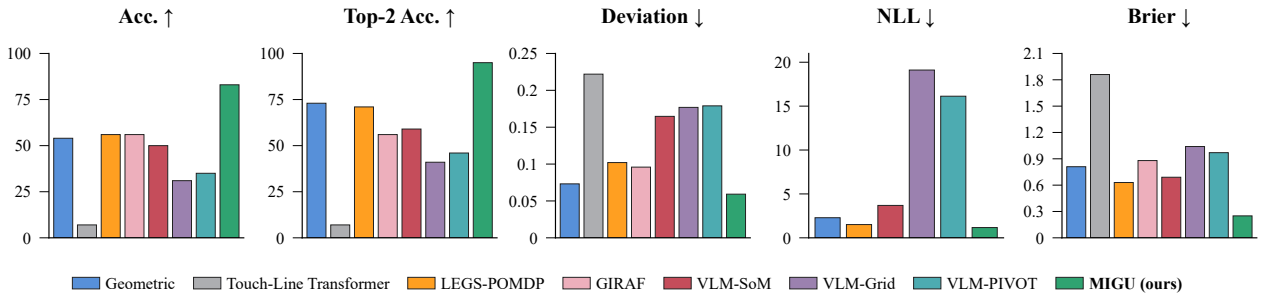

Fig. 7. **Grounding performance against baselines.** MIGU is green; arrows indicate better performance. NLL is omitted for deterministic methods.

cuRoboV2 [41] solves constrained inverse kinematics (IK) in batches. Batched collision checks reject infeasible solutions, and the remaining solutions determine the base-pose ranking. MoveBase navigates to the selected base pose. The robot then observes the target again, generates grasps with Contact-GraspNet [42], and solves IK before planning the arm motion with VAMP [43]. The placement strategy is similar, but uses heuristic sampling of end-effector placements at both near and far distances.

Tabletop Task and Motion Planning. We integrate the grounding results with PDDLStream [44]. We follow a similar system design as in [45], replacing the object detection with OWLv2 [22] and the segmentation model with SAM [46]. A grounded object o instantiates a symbolic goal such as $on(o_g, o)$, where o_g is the referred object. A grounded region is represented by a continuous variable r_g; for example, "place the object near here" translates to $near(o, r_g)$. PDDLStream samples grasps and placements, solves inverse kinematics, and generates collision-free motions to satisfy the goal. The resulting plan is passed to the robot's trajectory-tracking controller for execution.

V. Experiments

To better understand the design principles underlying multimodal instruction grounding, we conduct a series of evaluations across diverse real-world daily-life scenarios.

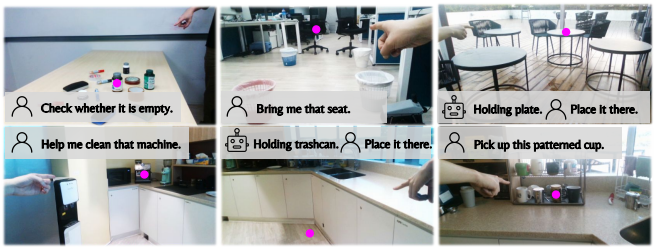

Fig. 8. Benchmark examples spanning object and region references. Magenta dots mark annotated targets.

A. Multimodal Instruction Grounding Evaluation

We first evaluate MIGU's grounding accuracy and probabilistic prediction quality through comparisons with baseline methods and component-wise ablations.

1) Benchmark and Metrics: We collect 100 multimodal instruction-grounding scenarios, each paired with the language instruction and a ground-truth target, as visualized in Fig. 8. The dataset encompasses semantic ambiguity, geometric ambiguity, and their combination, with interaction distances ranging from short-range tabletop settings (0.2–0.5 m) to medium-range settings (0.5–3 m). We report *object-level accuracy (Acc.)*, *top-2 accuracy (Top-2)*, *scaled 2D grounding deviation (Deviation.)*, *negative log-likelihood (NLL)*, and *Brier score*. Higher accuracy and lower deviation, *NLL*, and *Brier score* indicate better performance. *NLL* evaluates the probability assigned to the correct referent, while *Brier score* evaluates the complete candidate distribution.

2) Baselines: We compare against a geometric method [47], Touch-Line Transformer [48], LEGS-POMDP [11], and GIRAF [8]. Three pure VLM variants use Set-of-Mark prompting (VLM-SoM) [14], image grid visual prompting (VLM-Grid) [15], or iterative visual proposal refinement (VLM-PIVOT) [16]. All comparisons concern grounding performance on this benchmark. Touch-Line Transformer and GIRAF are deterministic, so NLL are not applicable.

3) Results and Analysis: As shown in Fig. 7, *MIGU outperforms all baseline methods in object-level grounding accuracy*. MIGU also achieves the lowest scaled 2D deviation, demonstrating more accurate fine-grained spatial grounding. This is particularly important for geometrically sensitive tasks such as placing, wiping, and cleaning. *Touch-Line Transformer* performs the worst, likely due to limited generalization from in-domain training, while the pure VLM baseline struggles in geometrically ambiguous cases because of its limited spatial reasoning capability. *LEGS-POMDP* relies primarily on 2D geometric computation and also does not fully exploit the advanced reasoning capabilities of modern VLMs. In terms of probabilistic prediction, MIGU achieves the lowest reported NLL and Brier score, outperforming LEGS-POMDP by 22.5% and 60.3%, respectively. This demonstrates its stronger capability for downstream decision-making under uncertainty.

4) Ablation Study: Table I compares modality removal and component replacement. In the variant names, $\rightarrow$ denotes replacing the indicated geometric (Geo.) or semantic (Sem.) component with the method on the right. *ray-distance* [34] scores distance to the pointing ray, *ray-angle* [33] uses a Gaussian angular likelihood with $1\,\mathrm{rad}$ standard deviation, and *distance-threshold* assigns uniform likelihood within $0.2\,\mathrm{m}$ of the ray. *VLM+detector* uses VLM-generated queries with an open-vocabulary detector to obtain semantic boxes, with a uniform distribution within each box.

Removing semantics or geometry reduces accuracy from

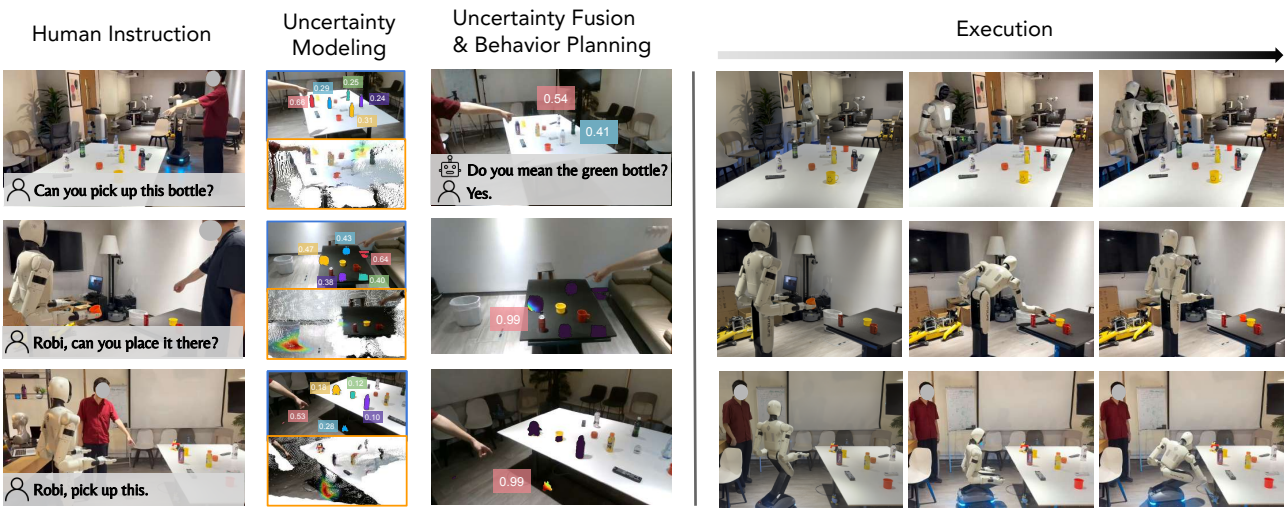

Fig. 9. **Snapshots of Mobile Manipulation Tasks.** Mobile manipulation from language and pointing, with clarification for ambiguous targets.

TABLE I: Ablation Study.

Variant	Acc. ↑	Top-2. ↑	Deviation ↓	NLL ↓	Brier ↓
w/o semantics	51%	76%	0.0869	1.37	0.49
w/o geometry	46%	64%	0.1746	1.76	0.57
Geo. → ray-distance	76%	89%	0.0805	2.91	0.40
Geo. → ray-angle	58%	83%	0.1441	1.63	0.51
Geo. → distance-threshold	74%	90%	0.0851	2.88	0.36
Sem. → VLM+detector	60%	76%	0.0984	5.08	0.47
Sem. → VLM-SoM	77%	85%	0.0907	3.49	0.41
Sem. → VLM-Grid	40%	45%	0.1287	17.66	1.02
Sem. → VLM-PIVOT	42%	49%	0.1237	14.67	0.93
MIGU (full)	**83%**	**95%**	**0.0592**	**1.17**	**0.25**

TABLE II: Real World Results.

	Semantic Only	Geometry Only	Without Clarification	With Clarification
Planning success	45.8%	41.7%	83.3%	91.7%

83% to 51% and 46%, respectively, demonstrating the benefit of combining both modalities. Replacing the proposed geometric model with ray-distance also yields poorer performance in both accuracy and probabilistic metrics, *supporting that the proposed geometric model contributes to better belief modeling and probabilistic prediction*. We further investigate different strategies for generating the semantic prior. Replacing direct VLM-generated grounding regions with an external detector API, SoM, Grid, or PIVOT consistently degrades performance. A possible explanation is that *these approaches introduce detector or prompting bottlenecks that lose fine-grained spatial information, whereas directly querying the VLM preserves richer grounding cues*.

B. Integration with Mobile Manipulation

We integrate MIGU with the mobile manipulation system described in Sec. IV-E on an Autolife S2 wheeled humanoid robot, equipped with an omnidirectional base, a 4-DoF waist, and two 7-DoF arms with grippers. We consider eight real-world tasks, each repeated three times. The scenes follow a setup similar to the grounding benchmark, where language instructions and pointing gestures jointly specify manipulation targets. In the clarification-enabled setting, the robot determines whether and how to ask a clarification question as described in Sec. IV-D, then updates its grounding belief based on the user's response. We compare the planning success rates (*Planning SR.*) of four variants: semantic-only grounding, geometry-only grounding, and MIGU without and with clarification. As shown in Table II, fusing semantic and geometric evidence achieves a success rate of 83.3%, substantially higher than semantic-only (45.8%) and geometry-only (41.7%) grounding. Enabling clarification further improves success to 91.7%, an increase of 8.4 percentage points. These results demonstrate that both multimodal evidence fusion and clarification-based belief refinement improve the reliability of human multimodal instruction grounding.

Figure 9 illustrates three mobile manipulation tasks. In the first task (top row), the instruction remains ambiguous among multiple bottles, while sensor noise affects pointing estimation. After fusion, the two leading candidates retain probabilities of 0.54 and 0.41. The robot requests clarification, uses the response to resolve the ambiguity, and plans a whole-body grasp. This case demonstrates the value of clarification when combined semantic and geometric evidence remains inconclusive. In the second task, the robot holds a cup while the user specifies a placement destination. Geometry alone assigns substantial probability to the floor and locations near the trash can, whereas the semantic prior favors placing the cup on a table. Fusion concentrates the belief on the left-side table region, guiding whole-body placement and illustrating how commonsense reasoning resolves geometric ambiguity. In the final task, fusion identifies the toy underneath the table with high confidence, allowing execution without clarification. The robot plans a constrained whole-body motion to reach and grasp it, demonstrating the connection between multimodal grounding and manipulation in confined spaces.

C. Integration with Task and Motion Planning

We demonstrate integration with the TAMP system described in Sec. IV-E on a Fetch robot [49] in cluttered table-

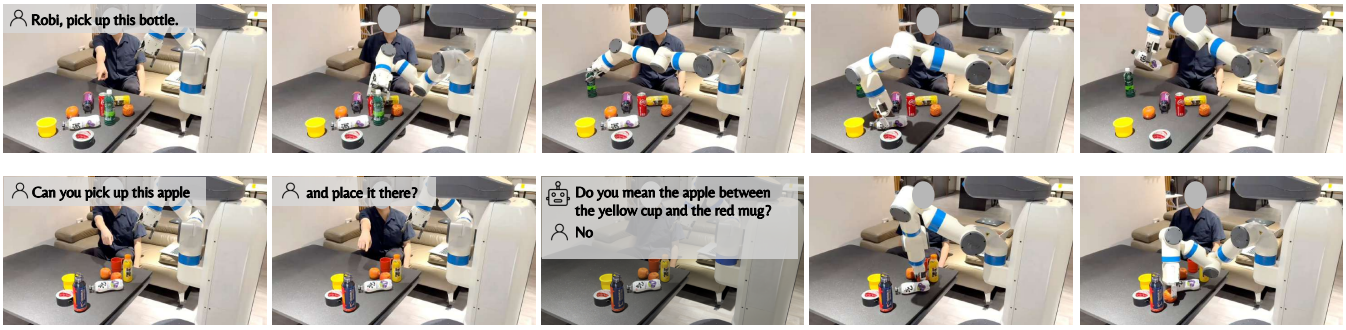

Fig. 10. **Snapshots of Tabletop Manipulation.** Top: obstacle rearrangement and bottle retrieval. Bottom: clarification followed by apple pick-and-place.

top scenes. Language and pointing jointly specify the target object and, for placement tasks, the destination region. The LLM first preprocesses the raw instruction and determines whether the grounding model should be queried twice when a single instruction contains two distinct pointing references. Figure 10 illustrates two tasks. In the bottle-retrieval task, the robot receives a multimodal instruction and proceeds with planning based on the fused belief without requesting clarification. The solver determines that a blocking bottle must first be removed before retrieving the target, illustrating how a grounded goal can require intermediate manipulation actions. In the apple pick-and-place task, the robot chooses to ask a spatial clarification question to distinguish between candidate apples. The user's negative response further refines the grounding belief, after which the robot generates a constrained motion plan to complete the task. These examples demonstrate the connection between multimodal reference resolution and task and motion planning, covering both object selection and region specification.

VI. Discussion and Future Work

Our current formulation reflects several practical design compromises. First, it relies on the commonsense reasoning capability of the VLM, assuming that semantic cues do not severely bias the geometric pointing evidence. Second, the semantic confidence provided by the VLM is used as a relative prior rather than a calibrated probability, since reliable calibration is difficult without a huge in-domain calibration dataset. Third, clarification is restricted to a single binary interaction round to meet real-time requirements, rather than being propagated into the manipulation planning process for more tightly coupled decision-making. These choices motivate future work on ensemble-based VLM grounding, calibrated semantic uncertainty, and more tightly integrated uncertainty-aware manipulation planning [50], [51].

VII. Conclusions

We presented MIGU, a modular framework that connects multimodal instruction grounding under uncertainty with manipulation planning. MIGU combines a 3D pointing likelihood with a VLM-derived semantic prior to maintain a belief over candidate objects and regions, enabling the robot to choose between direct execution and clarification. On the real-world benchmark, MIGU achieves 83% grounding accuracy, with ablations confirming the contribution of both modalities. Clarification further improves mobile manipulation planning success from 83.3% to 91.7%. Mobile manipulation and tabletop TAMP demonstrations show how grounded references support complex object rearrangement. These results highlight explicit uncertainty modeling as a practical basis for resolving human intent.

Acknowledgment

The authors used OpenAI's ChatGPT to assist with code writing and manuscript text editing. The authors take full responsibility for the final manuscript and its conclusions.

References

[1] A. Xiao *et al.*, "Robi butler: Multimodal remote interaction with a household robot assistant," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 4337–4344.

[2] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human–robot interaction," *Image and vision computing*, vol. 25, no. 12, pp. 1875–1884, 2007.

[3] T. Kukier *et al.*, "An empirical study on pointing gestures used in communication in household settings," *Electronics*, vol. 14, no. 12, p. 2346, 2025.

[4] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning from unscripted deictic gesture and language for human-robot interactions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, 2014.

[5] Y. Chen *et al.*, "Yourefit: Embodied reference understanding with language and gesture," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 1365–1375.

[6] M. M. Islam, A. Gladstone, and T. Iqbal, "Patron: Perspective-aware multitask model for referring expression grounding using embodied multimodal cues," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 971–979.

[7] A. M. Mane, D. Weerakoon, V. Subbaraju, S. Sen, S. E. Sarma, and A. Misra, "Ges3vig: Incorporating pointing gestures into language-based 3d visual grounding for embodied reference understanding," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 9017–9026.

[8] L.-H. Lin, Y. Cui, Y. Hao, F. Xia, and D. Sadigh, "Gesture-informed robot assistance via foundation models," in *7th Annual Conference on Robot Learning*, 2023.

[9] A. Galiza Cerdeira Gonzalez *et al.*, "iChores: Intuitive collaboration with household robots in everyday settings," *IEEE Robotics and Automation Practice*, vol. 1, pp. 60–65, 2026.

[10] P. Vanc, R. Skoviera, and K. Stepanova, "Tell and show: Combining multiple modalities to communicate manipulation tasks to a robot," *arXiv preprint arXiv:2404.01702*, 2024. [Online]. Available: https://arxiv.org/abs/2404.01702

[11] I. X. He, S. Tellex, and J. X. Liu, "Legs-pomdp: Language and gesture-guided object search in partially observable environments," in *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, 2026, pp. 227–236.

[12] S. Bai *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[13] G. R. Team *et al.*, "Gemini robotics: Bringing ai into the physical world," *arXiv preprint arXiv:2503.20020*, 2025.

[14] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V," *arXiv preprint arXiv:2310.11441*, 2023. [Online]. Available: https://arxiv.org/abs/2310.11441

[15] F. Liu, K. Fang, P. Abbeel, and S. Levine, "Moka: Open-world robotic manipulation through mark-based visual prompting," in *Robotics: Science and Systems (RSS)*, 2024.

[16] S. Nasiriany *et al.*, "PIVOT: Iterative visual prompting elicits actionable knowledge for VLMs," *arXiv preprint arXiv:2402.07872*, 2024. [Online]. Available: https://arxiv.org/abs/2402.07872

[17] V. Cohen, J. X. Liu, R. Mooney, S. Tellex, and D. Watkins, "A survey of robotic language grounding: tradeoffs between symbols and embeddings," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI '24, 2024. [Online]. Available: https://doi.org/10.24963/ijcai.2024/885

[18] S. Tellex *et al.*, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 25, no. 1, 2011, pp. 1507–1514.

[19] Z. Zhao, W. S. Lee, and D. Hsu, "Differentiable parsing and visual grounding of natural language instructions for object placement," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 546–11 553.

[20] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[21] S. Liu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*. Springer, 2024, pp. 38–55.

[22] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," in *Advances in Neural Information Processing Systems*, vol. 36, 2023. [Online]. Available: https://papers.neurips.cc/paper_files/paper/2023/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html

[23] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 217–232, 2020.

[24] H. Zhang, Y. Lu, C. Yu, D. Hsu, X. Lan, and N. Zheng, "Invigorate: Interactive visual grounding and grasping in clutter," in *Robotics: Science and Systems (RSS)*, 2021.

[25] Y. Yang, X. Lou, and C. Choi, "Interactive robotic grasping with attribute-guided disambiguation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8914–8920.

[26] A. Xiao, H. Zhang, T. Hu, and D. Hsu, "Hypothesis-driven model expansion under uncertainty for open-world robot planning," in *Robotics: Science and Systems (RSS)*, 2026.

[27] A. Sauppé and B. Mutlu, "Robot deictics: How gesture and context shape referential communication," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 2014, pp. 342–349.

[28] N. Dhingra, E. Valli, and A. Kunz, "Recognition and localisation of pointing gestures using a rgb-d camera," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 205–212.

[29] S. Nakamura, Y. Kawanishi, S. Nobuhara, and K. Nishino, "Deeppoint: Visual pointing recognition and direction estimation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 20 520–20 530.

[30] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth, "Deixis: How to determine demonstrated objects using a pointing cone," in *International Gesture Workshop*. Springer, 2005, pp. 300–311.

[31] M. Arslanoglu, K. Yilmaz, C. K. Özaltan, T. Linder, and B. Leibe, "Pointing3d: A benchmark for 3d object referral via pointing gestures," in *9th Annual Conference on Robot Learning*, 2025.

[32] M. H. Pelgrim *et al.*, "Find it like a dog: Using gesture to improve object search," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, 2024.

[33] D. Whitney, M. Eldon, J. Oberlin, and S. Tellex, "Interpreting multimodal referring expressions in real time," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 3331–3338. [Online]. Available: https://h2r.cs.brown.edu/wp-content/uploads/2016/02/whitney16.pdf

[34] M. Prochazka, S. Dratva, P. Vanc, R. Skoviera, K. Stepanova, and M. Vavrecka, "Probabilistic reasoner integrating multiple gestures and language instructions to control a sequential action execution in a humanoid robot," in *2025 IEEE International Conference on Advanced Robotics (ICAR)*. IEEE, 2025, pp. 696–701.

[35] B. Odegaard, D. R. Wozny, and L. Shams, "Biases in visual, auditory, and audiovisual perception of space," *PLOS Computational Biology*, vol. 11, no. 12, p. e1004649, 2015.

[36] J. M. Dukes, J. F. Norman, and C. D. Shartzer, "Visual distance perception indoors, outdoors, and in the dark," *Vision Research*, vol. 194, p. 107992, 2022.

[37] F. Zhang *et al.*, "MediaPipe Hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020. [Online]. Available: https://arxiv.org/abs/2006.10214

[38] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[39] T. Hu, A. Xiao, D. Hsu, and H. Zhang, "Visibility-aware mobile grasping in dynamic environments," *arXiv preprint arXiv:2605.02487*, 2026.

[40] N. Vahrenkamp, T. Asfour, and R. Dillmann, "Robot placement based on reachability inversion," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 1970–1975.

[41] B. Sundaralingam, A. Murali, and S. Birchfield, "cuRoboV2: Dynamics-aware motion generation with depth-fused distance fields for high-DoF robots," *arXiv preprint arXiv:2603.05493*, 2026. [Online]. Available: https://arxiv.org/abs/2603.05493

[42] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 438–13 444. [Online]. Available: https://arxiv.org/abs/2103.14127

[43] W. Thomason, Z. Kingston, and L. E. Kavraki, "Motions in microseconds via vectorized sampling-based planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 8749–8756.

[44] C. R. Garrett, T. Lozano-Pérez, and L. P. Kaelbling, "PDDLStream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 30, no. 1, 2020, pp. 440–448.

[45] A. Curtis, X. Fang, L. P. Kaelbling, T. Lozano-Pérez, and C. R. Garrett, "Long-horizon manipulation of unknown objects via task and motion planning with estimated affordances," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1940–1946.

[46] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html

[47] P. Kondaxakis, K. Gulzar, and V. Kyrki, "Temporal arm tracking and probabilistic pointed object selection for robot to robot interaction using deictic gestures," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 186–193.

[48] Y. Li *et al.*, "Understanding embodied reference with touchline transformer," in *International Conference on Learning Representations*, 2023. [Online]. Available: https://arxiv.org/abs/2210.05668

[49] M. Wise, M. Ferguson, D. King, E. Diehr, and D. Dymesich, "Fetch & Freight: Standard platforms for service robot applications," in *IJCAI Workshop on Autonomous Mobile Service Robots*, 2016. [Online]. Available: https://www.robotandchisel.com/assets/papers/Fetch-and-Freight-Workshop-Paper.pdf

[50] C. R. Garrett, C. Paxton, T. Lozano-Pérez, L. P. Kaelbling, and D. Fox, "Online replanning in belief space for partially observable task and motion problems," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5678–5684.

[51] A. Curtis *et al.*, "Partially observable task and motion planning with uncertainty and risk awareness," in *Robotics: Science and Systems (RSS)*, 2024.