Collaborative Instance Object Navigation: Leveraging Uncertainty-Awareness to Minimize Human-Agent Dialogues

Collaborative Instance Object Navigation: Leveraging Uncertainty-Awareness to Minimize Human-Agent Dialogues

Francesco Taioli[1,2], Edoardo Zorzi[2], Gianni Franchi[3], Alberto Castellini[2], Alessandro Farinelli[2], Marco Cristani[2,5], Yiming Wang[4]

[1]Polytechnic of Turin, [2]University of Verona, [3]U2IS, ENSTA Paris, [4] Fondazione Bruno Kessler, [5] University of Reykjavik

francesco.taioli@polito.it, {name.surname}@univr.it, gianni.franchi@ensta-paris.fr, ywang@fbk.eu

https://intelligolabs.github.io/CoIN

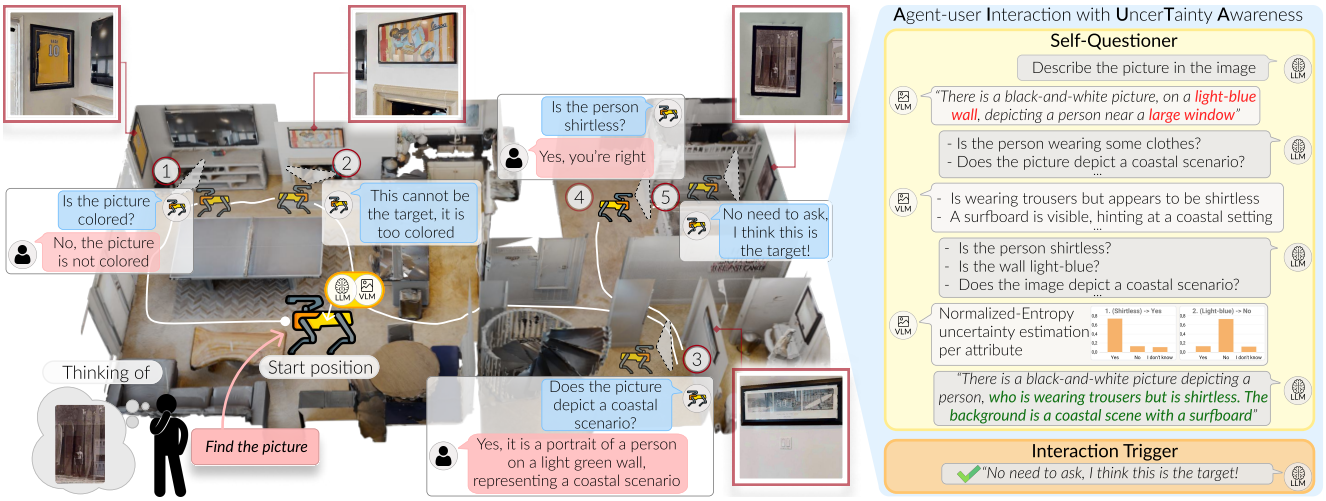

Figure 1. CoIN task illustration. A human provides a request (*"Find the picture"*) in *natural language*. The agent has to locate the object in a *completely unknown* environment *without any target image as input*, interacting with the user only when needed via *template-free*, *open-ended natural-language dialogues*. Our method, **A**gent-user **I**nteraction with **U**ncer**T**ainty **A**wareness (**AIUTA**), minimizes user interactions by equipping the agent with two modules: a **Self-Questioner** and an **Interaction Trigger**. The **Self-Questioner** leverages a LLM and VLM in a self-dialogue to describe the agent's observation and then extract additional relevant details, with a novel entropy-based technique to reduce **hallucinations and inaccuracies**, producing a refined **detection description**. The **Interaction Trigger** uses this refined description to decide whether to pose a question to the user (①,③,④), continue the navigation (②) or halt the exploration (⑤).

Abstract

Language-driven instance object navigation assumes that a human initiates the task by providing a detailed description of the target to the embodied agent. While this description is crucial for distinguishing the target from other visually similar instances, providing it prior to navigation can be demanding for humans. We thus introduce Collaborative Instance object Navigation (CoIN), a new task setting where the agent actively resolves uncertainties about the target instance during navigation in natural, template-free and open-ended dialogues with the human, minimizing user input. We propose a novel training-free method, Agent-user Interaction with UncerTainty Awareness (AIUTA), which operates independently from the navigation policy, and focuses on the human-agent interaction reasoning using Vision-Language Models (VLMs) and Large Language Models (LLMs). First, upon object detection, a Self-Questioner model initiates internal self-dialogues within the agent to obtain a complete and accurate observation with a novel uncertainty estimation technique. Then, an Interaction Trigger module determines whether to ask a question to the human, continue, or halt navigation. For evaluation, we introduce CoIN-Bench, with a curated dataset designed for challenging multi-instance scenarios. CoIN-Bench supports both online evaluation with humans and reproducible experiments with simulated user-agent interactions. On CoIN-Bench, we show that AIUTA serves as a competitive baseline, whereas existing language-driven instance navigation methods struggle in multi-instance scenes.

1. Introduction

Recent advances in Large Language Models (LLMs) and Vision-Language Models (VLMs) have significantly reinvigorated research on *language-driven* navigation tasks [1, 3, 19, 23, 56], where a human engages with embodied agents via natural language only, the most intuitive human-agent interaction among other forms (*e.g.*, visual reference [73]). In this paper, we focus on the *language-driven* Instance Object Navigation (InstanceObjectNav) task [19, 23], a practical task where the agent aims to locate a *specific* instance within an unknown 3D scene, based on a detailed instance description (differently from ObjectNav [6] where *any* object of a category can be located). The instance description typically contains nuanced details about the intrinsic (*e.g.*, color, material) and extrinsic (*e.g.*, context, spatial relations) attributes of the searched object instance, which are essential to *uniquely identifying* the target amid visual ambiguity. However, the standard language-driven InstanceObjectNav task assumes that the detailed instance description is provided upfront, before navigation begins. This assumption can be demanding and impractical in real world, as users may unable or unwilling to supply all details in advance.

We introduce the **Co**llaborative **I**nstance object **N**avigation (**CoIN**) task, which engages a human user via natural-language dialogues to resolve instance visual ambiguity during navigation. CoIN enables human users to initiate the InstanceNav task *without* providing extensive instance description. For instance, the user can just specify the instance category, *e.g.*, "*Find the picture*", a challenging minimal-guidance scenario. Notably, CoIN introduces, for the first time, *template-free*, *open-ended* human-agent dialogues, a significant departure from the templated question-answer pairs used in prior work [12]. Instead, our agent engages in dialogue solely based on the understanding gained during navigation. Within CoIN, two key research questions arise: *1) When* and *2) How* should agent-user interaction occur? To address the "*When*", the agent must develop an internal model of its perceived environment to determine the optimal moments for seeking assistance from the user, resolving ambiguities effectively. To address the "*How*", the agent must formulate *the most informative questions* to maximize its chances of locating the target.

We introduce a novel *zero-shot* approach called **A**gent-user **I**nteraction with **U**ncer**T**ainty **A**wareness (**AIUTA**). AIUTA equips the agent with two onboard modules, the *Self-Questioner* and the *Interaction Trigger*, leveraging pre-trained VLMs and LLMs without additional training. The *Self-Questioner* enables the agent to autonomously generate *self-dialogues* to inquire additional *target-relevant* details, and verify essential details with a *novel technique for uncertainty estimation*. As shown in Fig. 1, upon detection, the LLM first prompts the VLM to obtain an initial detection description which can be *incomplete* and *inaccurate*. To enrich with target-relevant details, the LLM further generates questions for the VLM, whose responses complement the initial description. However, since VLMs cannot guarantee accurate responses grounded in the visual counterpart [28, 41, 58], we further prompt the LLM to generate sanity-check questions about all relevant details (e.g., "*Is the wall light-blue?*"). We instruct the VLM's response to be either `Yes`, `No` or `I don't know`, proposing a novel *Normalized-Entropy*-based technique to quantify the VLM uncertainty. Finally, the *Interaction Trigger* module leverages the LLM to predict an *alignment score* between the refined detection description and the known target's *facts* acquired from previous agent-human dialogues, if any. With the score, the module decides whether to continue navigation, terminate it, or ask human clarifying questions.

To evaluate CoIN, we propose the first benchmark, **CoIN-Bench**, with a curated dataset that specifically focuses on the visual ambiguity challenge in CoIN. The dataset is created on top of the recent large-scale GOAT-Bench [19], where we only consider episodes involving multiple instances in a scene, with high-quality visual observations on the target. In total, CoIN-Bench consists of $1,649$ evaluation episodes, with on average five distractors (non-target instances of the same category) per episode. Our benchmark supports *on-line* evaluation with humans and reproducible evaluation via simulated user-agent interactions. We empirically show that the simulated user-agent interaction yields results that are in line with human evaluation. Using CoIN-Bench, AIUTA, *while being training-free*, outperforms state-of-the-art InstanceObjectNav methods trained on the dataset in the *zero-shot* setting, in terms of success rate and path efficiency. Finally, to evaluate VLM uncertainty estimation, we introduce the "I Don't Know Visual Question Answering" (*IDKVQA*). On IDKVQA, our proposed *Normalized-Entropy*-based technique outperforms recent competitors [29], being a more reliable uncertainty measure.

Paper Contributions are summarized as follows:

- We introduce *CoIN*, a practical InstanceObjectNav setting that minimizes human input through agent-human dialogues during navigation.
- We propose *AIUTA*, a training-free method addressing CoIN, using self-dialogues within the agent to reduce perception uncertainty and minimize agent-user interactions.
- We propose a novel *Normalized-Entropy-based* technique to quantify VLM perception uncertainty. Using IDKVQA dataset, we show improved reliability over baselines.
- We present *CoIN-Bench*, a benchmark for CoIN with challenging multi-instance scenarios, enabling reproducible evaluation via VLM-simulated user and real human.

2. Related Works

Instance Object Navigation. InstanceObjectNav extends the Object-Goal navigation (ObjectNav) [2, 5] task. Un-

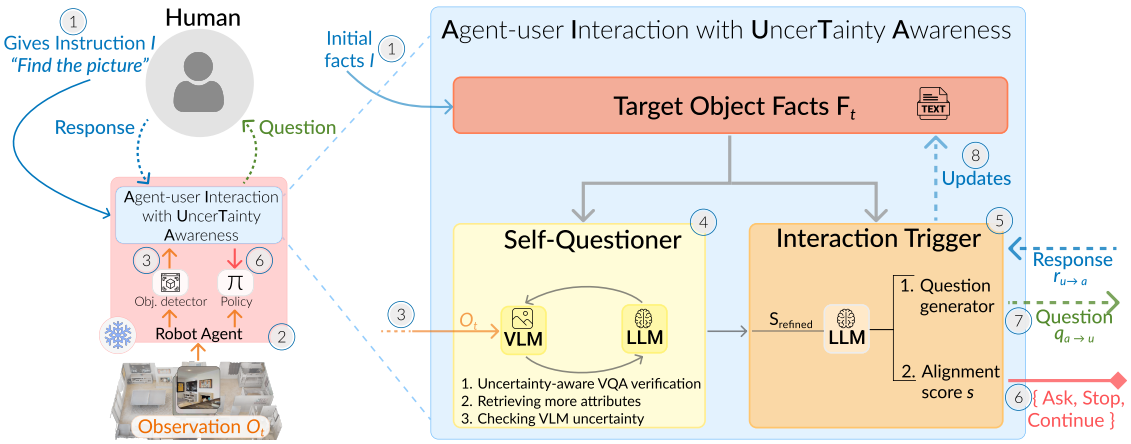

Figure 2. Graphical depiction of **AIUTA**: left shows its interaction cycle with the user, and right provides an exploded view of our method. ① The agent receives an initial instruction I: "Find a c =<object category>". ② At each timestep t, a zero-shot policy π [63], comprising a frozen object detection module [27], selects the optimal action a_t. ③ Upon detection, the agent performs the proposed AIUTA. Specifically, ④ the agent first obtains an initial scene description of observation O_t from a VLM. Then, a ***Self-Questioner*** module leverages an LLM to automatically generate attribute-specific questions to the VLM, acquiring more information and refining the scene description with reduced attribute-level uncertainty, producing $S_{refined}$. ⑤ The ***Interaction Trigger*** module then evaluates $S_{refined}$ against the "*facts*" related to the target, to determine whether to terminate the navigation (if the agent believes it has located the target object ⑥), or to pose *template-free, natural-language* questions to a human ⑦, updating the "*facts*" based on the response ⑧.

like ObjectNav, which seeks *any* instance of a given category, InstanceObjectNav locates a *specific* instance. While the instance can be given via an image [20], we focus on users describing the target only in natural language. Recent policies can be divided into two categories: training-based [10, 19, 31, 44, 52, 64] and zero-shot policies [11, 21, 54, 63, 65, 68, 71]. Trained policies rely exclusively on RL [19, 31, 52] or in conjunction with behavioral cloning [44]. Vision-language-aligned embeddings offer a promising alternative by enabling policies to incorporate detailed natural language descriptions as input. For instance, GOAT-Bench [19] employ CLIP embeddings as the goal modality, while [31, 52] train on image-goal navigation [73] and evaluate on the ObjectNav task. Among zero-shot policies, several methods extend the frontier-based exploration [62], by incorporating LLM reasoning [21, 65, 68, 71], CLIP-based localization [11] or vision-language maps [53] for frontier selections [63]. The recent [4] provides agents with multimodal instance references, *i.e.*, a set of images and textual descriptions. Differently, we enable human-agent interactions during navigation *without* any access to a target image.

Interactive Embodied AI. Common approaches involve agents asking users for help [55], with responses typically consisting of shortest-path actions to target [7, 51] or simple natural language sub-goals guiding navigation [30, 35, 36, 40, 45]. In [34, 47] authors proposed a framework to measure the uncertainty of an LLM-based planner, enabling the agent to determine the next action or ask for help. [16] show that LLMs can generate inner monologues when given environmental feedback, improving planning in robotic control scenarios. Both [12, 57] include a dialog-guided task completion benchmark using human-annotated question-answer pairs collected via Amazon Mechanical Turk. In [37], agent requests are limited to 3 fixed types. Similarly, [13] asks templated questions focused on appearance, location, direction, while [74] asks "Should I go [dir] to the [obj]?" and [26] adopts fixed-format templated multiple-choice Q&A. ELBA [50] generates both oracle-based and model-based templated questions from oracle answers. FindThis [32] only lets the agent respond with candidate object images, without the ability to ask questions or use free-form natural language. In [9], the ZIPON task requires agents to find personalized objects. However, personalized goals are manually annotated, and the user, simulated by an LLM, can only reply with this ground-truth data. Both [9, 32] rely on a pre-built top-down semantic/occupancy map to locate the objects of interest. In [34], requests are allowed only at predefined location; the simulator then provides a natural language navigation subtask and an image of the target. [69] relies on manually defined disambiguation intents from the dataset to decide for clarification, limiting real world applicability. In contrast, our agent identifies the target instance only through *open-ended, template-free natural language dialogues* with the user.

Vision-Language Models Uncertainty. Hallucinations, biases, reasoning failures and the generation of unfaithful text by LLMs are well-known issues [18]. [38] shows that truthful information clusters in specific tokens, which can be leveraged to enhance error detection performance. How-

ever, these error detectors fail to generalize across datasets. Similarly, recent studies highlight systematic limitations in the visual capabilities of VLM [28, 58], leading them to hallucinate or giving inaccurate answers to unanswerable or misleading questions [41]. To this end, PAI [28] amplifies attention weights on image tokens to prioritize visual input. In [70], a linear probing on the logits distribution of the first tokens determines whether visual questions are answerable/unanswerable. CLARA [39] estimates LLMs uncertainty via context sampling to distinguish between certain and uncertain commands. [72] introduces a VLM-LLM dialogue for image captioning. In contrast, AIUTA uses the self-dialogue for an embodied task, combining both observation and target facts for question generation, reducing hallucinations with a novel uncertainty estimation technique.

3. Collaborative Instance Object Navigation

Collaborative Instance Object Navigation (CoIN) introduces a novel setting for the InstanceObjectNav task, where an agent navigates in an unknown environment to locate a specific target instance in collaboration with a human user via *template-free*, *open-ended* and *natural-language* interactions. The agent decides whether an interaction is needed to gather necessary target information from the user during the navigation. The objective of CoIN is to successfully locate the target instance with *minimal user input*, reducing the effort for the user in providing a detailed description.

Initially, the agent is positioned randomly in an *unknown* 3D environment [43]. The navigation starts upon receiving a user request I in natural language, which can be as minimal as by only specifying an open-set category c, *e.g.*, "*Find the <category>*". The agent does not have access to any visual reference of the target instance. We assume that the user is: *(i)* aware of the full details about the target instance, and *(ii) collaborative* to provide the true response when being asked by the agent. At each time step t, the agent perceives a visual observation O_t of the scene, allowing it to guide a policy π to pick an action $a_t \in A = \{$`Forward 0.25m, Turn Right 15°, Turn Left 15°, Stop, Ask`$\}$, where `Ask` is the novel action that comes with our CoIN task. When invoked, the agent asks the user a *template-free open-ended* question $q_{a \to u}$ in natural language to gather more information about the target. With the user response $r_{u \to a}$, the agent updates the set of *facts* (set of attributes and characteristics) F_t, representing information derived exclusively from the interaction. Formally, the updated set of facts is represented as $F_t = F_{t-1} \cup r_{u \to a}$. The navigation terminates when certain criteria are met, *e.g.*, the agent selects the `Stop` action or exceeds the maximum number of allowed actions. Notably, the agent can move anywhere in the continuous environment [48]. CoIN is particularly relevant in *challenging* scenarios where many visually ambiguous instances co-exist.

4. Proposed Method

Our proposed **A**gent-user **I**nteraction with **U**ncer**T**ainty **A**wareness (**AIUTA**), a module that enriches the agent, is illustrated in Fig. 2. Upon receiving an initial user request I with minimal guidance that only specifies the category, *e.g.*, *"Find the picture"* (① in Fig. 2), AIUTA updates the known facts regarding the target instance, *i.e.*, $F_{t=0} = \{I\}$. Then, it activates a zero-shot navigation method, VLFM [63], perceiving the scene observation O_t and providing the navigation policy (② in Fig. 2). VLFM constructs an occupancy map to identify frontiers in the explored space, and a value map that quantifies the semantic relevance of these frontiers for target object localization using the pre-trained BLIP-2 [22] model. Object detection is performed by Grounding-DINO, an open-set object detector [27]. More details about VLFM [63] in the *Supp. Mat.* (Sec. C.1).

AIUTA is triggered upon the detection of an object belonging to the target class (③ in Fig. 2), executing two key components sequentially. First, the *Self-Questioner* (Sec. 4.1) leverages a VLM and a LLM to obtain an accurate and detailed understanding of the observed object via self-questioning, enabling reliable verification of the detection against the target (④ in Fig. 2). Next, the *Interaction Trigger* (Sec. 4.2), determines whether an agent-user interaction is necessary (in such case, triggering the action `Ask`), based on the observed object and known target facts F_t, and whether the agent should halt (*i.e.*, `Stop`) or proceed with the navigation (⑤ in Fig. 2). In the case of `Ask` (⑦ in Fig. 2), AIUTA updates the target facts F_t with the user's response (⑧ in Fig. 2). The agent end the navigation task once the target instance is deemed to be found. The complete algorithm can be found in *Supp. Mat.* (Sec. J). In the following, *Self-Questioner* and *Interaction Trigger* are fully detailed.

4.1. Self-Questioner

Upon detection, the Self-Questioner component aims to obtain a thorough and accurate description of the detected object. As suggested by previous studies [28, 41, 58], generative VLMs may produce descriptions that are not fully grounded on the visual content, leading to inaccuracy or hallucination. To mitigate this issue, we leverage an LLM to automatically generate attribute-specific questions for the VLM. In particular, we propose a novel technique for estimating uncertainty in VLM perception, enabling the refinement of detection descriptions. The technique has three steps: *(i)* generating an initial detection description with detailed information relevant to target identification; *(ii)* estimating VLM perception uncertainty to validate object detection; and *(iii)* refining the detection description by filtering out uncertain attributes. Each step is detailed below.

Generation of the initial detection description. The agent initially prompts the VLM for an initial description S_{init} of the observation O_t by providing the prompt P_{init} = *"De-*

scribe the <target object> in the provided image." Formally, $S_{init} = \text{VLM}(O_t, P_{init})$. The description S_{init} returned by the VLM could miss essential details for locating the specific instance, *e.g.*, when looking for a picture, the content of the picture itself may not be specified in the description. To mitigate this issue, we prompt the LLM to create a list of questions $Q^{details}_{a\to a} = \{q_j\}$ given S_{init} and F_t (the symbol $Q_{a\to a}$ is used to represent the self-dialogue performed by the *agent*). Formally, $Q^{details}_{a\to a} = \text{LLM}(P_{details}, S_{init}, F_t)$, where $P_{details}$ is the prompt guiding the question generation to obtain more details (*Supp. Mat.* Sec. I.2). The questions of $Q^{details}_{a\to a}$ are subsequently answered by the VLM. Specifically, it answers each question $q_j \in Q^{details}_{a\to a}$ with a response $r_j = \text{VLM}(O_t, q_j)$ given the observation O_t. Finally, we concatenate all responses to the initial detection S_{init}, obtaining an enriched detection description $S_{enriched}$.

Perception uncertainty estimation. VLMs can generate hallucinated or inaccurate content [28, 41, 58], impacting the performance of AIUTA. To address this, we propose a novel technique for estimating their perception uncertainty. Direct evaluation of this aspect is challenging and often requires architectural modifications. Instead, we employ a prompt-guided Shannon entropy-based method for effective assessment. Our goal is to measure the uncertainty $u \in [0, 1]$ of the VLM in perceiving specific aspects of a given image through visual question answering: the VLM answers to a specific question q with a response r and an associated uncertainty estimation u, *i.e.*, $(r, u) = \text{VLM}(O_t, q)$. Following the notation from [28], we consider an auto-regressive VLM, where $\mathbf{X}_I$ is the image representation (*i.e.*, image tokens), $\mathbf{X}_P$ is the prompt representation (*i.e.*, prompt text tokens), and $\mathbf{X}_H$ is the history representation (token generated at previous time-steps). During inference, the VLM generates a conditional probability distribution p over the vocabulary $\mathbf{y} \in \mathbb{R}^w$ at each time step, expressed as:

$$\begin{aligned} \mathbf{y} &\sim p_{\text{VLM}}(\mathbf{y} \mid \mathbf{X}_I, \mathbf{X}_V, \mathbf{X}_H), \\ &\propto \text{softmax}\left(\text{logit}_{\text{VLM}}(\mathbf{y} \mid \mathbf{X}_I, \mathbf{X}_V, \mathbf{X}_H)\right). \end{aligned} \tag{1}$$

Estimating the uncertainty of the VLM response is non-trivial as the VLM has an unbounded output space and its output probability distribution is over a (large) vocabulary of size w. To address this issue, we leverage the standard instruction-tuning [24] procedures for VLMs, utilizing a predefined set of templated answers to restrict the vocabulary size to a fixed, small w. In particular, during inference, we use the following prompt: *"<Question>? You must answer with* `Yes, No, or ?=I don't know`*."* In this way, we: *(i)* bound the auto-regressive nature to be essentially a one-step prediction, thus avoiding length-normalization; *(ii)* bound the vocabulary size, *i.e.*, $w = 3$. We then compute the Shannon entropy [49] H of a probability distribution p over vocabulary size w:

$$H(p_{\text{VLM}}) = -\sum_{i=1}^{w} p(y_i) \log p(y_i). \tag{2}$$

The VLM uncertainty u is then obtained by normalizing the entropy H within the range $[0, 1]$ as $u = \frac{H}{H_{\max}}$, where $H_{max} = \log(w)$ is the maximum entropy (*i.e.*, maximum uncertainty) over a vocabulary of size w.

Given a threshold τ, we can indicate if the answer is `Certain` or `Uncertain`, namely:

$$C(u, \tau) = \begin{cases} \texttt{Certain}, & u \leq \tau \\ \texttt{Uncertain}, & u > \tau \end{cases} \tag{3}$$

To reduce false positives, we use the prompt P_{check} = *"Does the image contain a <target object>? Answer with Yes, No or ?=I don't know."* (see *Supp. Mat.* Sec. I.3). This allows us to confirm the presence of the object, which we formally express as $(r_{check}, u_{check}) = \text{VLM}(O_t, P_{check})$. Following Eq. 3, we continue the AIUTA pipeline if response r_{check} = *"Yes"* and uncertainty $u_{check} = \texttt{Certain}$; otherwise, we continue exploring.

To remove uncertain attributes, we prompt the LLM to extract a set of attributes and values $K_t = \{(k_j, v_j)\}$ from the detection description $S_{enriched}$, where each attribute k_j is associated to a value v_j, *e.g.*, ("frame", "black"); ("content", "RGB image of a family"), etc. For each attribute k_j, we then prompt the LLM to generate a list of J questions, $Q^{attribute}_{a\to a} = \{q_j\}_{j=1}^{J}$ to be answered by the agent itself. Formally, we extract attributes list and self-questions in one prompt, $Q^{attribute}_{a\to a} = \text{LLM}(P_{selfquestions}, F, S_{enriched})$, where $P_{selfquestion}$ is the prompt for the LLM (*Supp. Mat.* Sec. I.4). For each question q_j, we access both the response r_j and the associated uncertainty u_j by evaluating $(r_j, u_j) = \text{VLM}(O_t, q_j)$. This process allows us to confirm or refine the attributes based on the VLM's responses, obtaining a final detailed description $S_{refined}$.

Detection description refinement. To obtain the final detailed description $S_{refined}$, we let the LLM filter out uncertain attributes, given the enriched description $S_{enriched}$ and the set of questions, responses, and uncertainties $\{q_j, r_j, u_j\}$. More formally, $S_{refined} = \text{LLM}(P_{refined}, \{q_j, r_j, u_j\}, S_{enriched})$, where $P_{refined}$ is the prompt for the LLM (see *Supp. Mat.* Sec. I.5).

4.2. Interaction Trigger

Using the accurate and detailed description $S_{refined}$ of the detected object, the Interaction Trigger prompts the LLM to decide whether to pose a question to the human user or continue the navigation. Specifically, we prompt the LLM to estimate a similarity score s between scene description $S_{refined}$ and target object facts F_t. We instruct the LLM to

estimate the similarity score based on the alignment between the detection description and the known facts. Formally, $s = \text{LLM}(P_{score}, S_{refined}, F_t)$, where P_{score} is prompt instructing the LLM to produce the similarity score (*Supp. Mat.* Sec. I.6). Based on the LLM-estimated similarity score, the agent takes corresponding action based on the following intuition: *(i)* if $s \geq \tau_{stop}$, the navigation terminates as the agent deems the instance has been found; *(ii)* if $s < \tau_{skip}$, the agent deems the detected object is significantly different from the known target facts, thus skipping the agent-user interaction to reduce the user efforts in providing input. The agent will continue with the environment exploration; and *(iii)* if $\tau_{skip} \leq s < \tau_{stop}$, the description and facts are somewhat aligned, suggesting that posing a question to the user can effectively reduce uncertainty.

When taking the action `Ask`, we further leverage the capability of LLM to compose an effective question to the user, $q_{a \to u}$, aimed at maximizing information gain about the target instance, conditioned on the know target object facts F and the refined observation description $S_{refined}$. To minimize the number of LLM calls, we incorporate such question retrieval inside the P_{score} prompt. After receiving the corresponding response from the human, $r_{u \to a}$, we update the target object facts F_t with new information, maximizing the effectiveness of later agent-human interactions.

5. CoIN-Bench

To facilitate the evaluation of CoIN, we introduce *CoIN-Bench*, a curated dataset that features challenging multi-instance scenarios, supports both human evaluation and simulated agent-user interactions, and includes a new performance metric that accounts for agent-user interactions.

Dataset Construction. Our dataset is built upon the large-scale GOAT-Bench [19], which spans diverse scenarios from the HM3DSem [43] using the Habitat sim [48]. GOAT-Bench provides instance references in various formats, including category names and natural-language descriptions, making it a suitable source dataset. GOAT-Bench consist of a large `Train` split for policy training, and three eval splits: `Val Seen`, `Val Seen Synonyms` and `Val Unseen`. Specifically, `Val Seen` includes objects seen in `Train`, `Val Seen Synonyms` introduces synonymous object names, and `Val Unseen` contains only *novel* objects absent from `Train`. Since GOAT-Bench's `Train` split is dedicated to policy training, we design CoIN-Bench exclusively for evaluation.

We select episodes from the evaluation splits of GOAT-Bench, *i.e.*, `Val Seen`, `Val Seen Synonyms` and `Val Unseen` to ensure fair comparison with methods trained on GOAT-Bench. Since CoIN focuses on scenarios with multiple instances of the same target category (*i.e.*, *distractors*), we apply a filtering procedure to discard episodes with fewer than $d_{min} = 2$ distractors. After filtering the episodes, the simulator [48] sets random start positions to the agent, ensuring a geodesic distance of $[5m, 20m]$ between the start and target locations to vary navigation difficulty. Moreover, since visual observation are 3D renderings whose quality is dependent on the scene reconstruction, we manually filter out episodes to ensure high-quality visual observations, removing those where target instances have insufficient resolution, limited visual coverage, or are indistinguishable from distractors. Additionally, we ensure episodes are navigable without crossing floors, following [19, 61]. CoIN-Bench dataset includes 831 episodes in `Val Seen`, 459 in `Val Unseen` and 359 in `Val Seen Synonyms`, with a total of $1,649$ evaluation episodes, in line with the evaluation scale of well-known datasets [5, 19, 20]. As shown in Tab. 1, CoIN-Bench features an average of ~ 5 distractors per episode, and a mean path length > 7, forming a highly challenging multi-instance evaluation set. More details and statistics are provided in *Supp. Mat.* (Sec. A).

Evaluation protocol. CoIN-Bench supports evaluation with both *real humans*, to assess the potential and limitation of genuine agent-human interactions, and simulated user-agent interactions, to enable extensive, reproducible and large-scale experiments. Simulating agent-human interactions is challenging due to: *(i)* the agent's open-ended, template-free questions about any target attribute, making it impractical to predefine a comprehensive question-answer dataset, and *(ii)* the huge question space in the simulated continuous environment [48]. To address this, we propose to simulate user responses via a VLM with access to a high-resolution image of the target object (1024×1024) at each episode. This setup is more effective than relying solely on instance descriptions [9], as the comprehensive visual coverage allows for diverse responses to the agent's questions.

Metrics. An episode is successful if the agent stops within 0.25m of the target goal viewpoints. If not located, the exploration ends after 500 actions. Following [2, 61], we use: Success Rate, `SR` (↑), our primary metric (in gray), and Success rate weighted by Path Length, `SPL` (↑). Additionally, we introduce the *average Number of Questions asked*, `NQ` (↓) in successful episodes to measure the amount of user input.

6. Experiments

We first benchmark AIUTA against state-of-the-art (SOTA) methods [19, 52, 63, 64] on CoIN-Bench, with simulated user-agent interactions, highlighting that CoIN-Bench present a challenging evaluation set for *training-free* and

Statistics	*Val Seen*	*Val Seen Synonyms*	Val Unseen
Avg. (std) number of distractors	4.58 (1.93)	6.01 (1.96)	5.15 (1.51)
Avg. (std) length (Geodesic)	9.32 (3.43)	9.13 (3.14)	9.86 (3.73)
Avg. (std) length (Euclidean)	7.48 (2.88)	7.50 (2.75)	7.78 (3.39)

Table 1. Avg. (std) number of distractors and distance to the goal.

Method	Model Condition		*Val Seen*			*Val Seen Synonyms*			Val Unseen		
	Input	Training-free	SR ↑	SPL ↑	NQ ↓	SR ↑	SPL ↑	NQ ↓	SR ↑	SPL ↑	NQ ↓
Monolithic† [19] (*CVPR-24*)	d	✗	6.62†	3.11	-	13.09†	6.45	-	0.22†	0.05	-
PSL [52] (*ECCV-24*)	d	✗	8.78	3.30	-	8.91	2.83	-	4.58	1.39	-
OVON† [64] (*IROS-24*)	c	✗	8.18†	5.24	-	15.88†	11.35	-	2.61†	1.29	-
VFLM [63] (*ICRA-24*)	c	✔	0.36	0.28	-	0.00	0.00	-	0.00	0.00	-
AIUTA (*ours*)	c	✔	**7.42**	2.92	1.67	**14.38**	7.99	1.36	**6.67**	2.30	1.13

Table 2. CoIN-Bench is challenging. AIUTA, while being *training-free*, achieves strong performance by outperforming trained policies (top rows) and significantly surpassing the zero-shot VLFM, across *all* splits, through effective user interaction. In contrast, policies trained on GOAT-Bench (denoted with †), the foundation of CoIN-Bench, fail to generalize to novel categories (Val Unseen). We report the SR (main metric, in **bold** *w.r.t* training free-methods), SPL, and the number of questions NQ. Input types: *c* for object category, *d* for its description.

training-based methods. Next, we conduct an evaluation on a small validation set using both real human and simulated user-agent interactions, demonstrating that the simulation setup serves as a viable alternative to real human evaluation, enabling scalable and reproducible experiments. Finally, ablation studies validate AIUTA design choices, and showcase the effectiveness of the Normalized-Entropy-based technique for estimating VLM uncertainty, outperforming recent baselines [29, 70] on the IDKVQA dataset.

Implementation Details. We use [25] (LLaVA 1.6, Mistral 7B) as the VLM and GPT-4o [17] as the LLM. User interaction is limited to a maximum of 4 rounds. We empirically set $\tau = 0.75$ (Eq. 3), $\tau_{stop} = 7$ and $\tau_{skip} = 5$ as they yield the best result. In *Supp. Mat.*, see Sec. I for all prompts and Sec. D for AIUTA's computational analysis.

Baselines. We compare AIUTA against SOTA Instance Navigation and ObjectNav methods: the SenseAct-NN Monolithic Policy (Monolithic) [19], PSL [52], OVON [64] and the zero-shot, training-free VLFM [63].

To demonstrate the challenging nature of our dataset, we include two baselines, Monolithic [19] and OVON [64], which are trained on GOAT-Bench. Again, note that the "*Seen*" splits contains categories seen during training (Sec. 5). PSL is trained on the ImageNav task and transferred on the language-driven Instance navigation task. Notably, both Monolithic and PSL take a fully detailed description *d* of the target instance as input, while OVON [64] takes the target category *c*. Finally, VLFM operates in a zero-shot, training-free manner, while taking category *c* in input. All baselines are detailed in *Supp. Mat.* Sec. C. Tab. 2 summarizes the input types and training conditions on CoIN-Bench.

Results with simulated user-agent interaction. As shown in Tab. 2, training-based methods perform better on `Val Seen` and `Val Seen Synonyms` than on `Val Unseen`, highlighting their poor generalization to novel categories. This phenomenon is particularly pronounced on policies trained on GOAT-Bench (denoted with †), with performance dropping significantly—OVON's SR decreases from a maximum of 15.88 to 2.61, and Monolithic's SR drops from 13.09 to 0.22. In contrast, AIUTA, while being training-free, outperforms training-based methods on `Val Unseen`, with consistent and strong results in all the splits. Interestingly, on the `Val Seen Synonyms`, AIUTA is inferior to OVON, but outperforms PSL and Monolithich in SR and SPL. This is surprising, as PSL and Monolithic are training-based and operate with detailed instance descriptions. One possible explanation is that CLIP-based approaches is limited in encoding fine-grained instance description compared to category [11, 19]. Moreover, compared to the results reported on GOAT-Bench, the lower SR of the baselines, *e.g.* Monolithic [19] on CoIN-Bench, highlights the introduced challenge of multi-instance ambiguity.

In particular, our closest competitor VLFM [63], when using only the instance category as input, fails nearly all evaluation episodes, with almost 0% SR across all splits. This is expected, as the large amount of distractor objects (Tab. 1) poses significant challenges for ObjectNav methods, which lack instance-level discrimination capabilities. Further analysis of VLFM 0% SR can be found in *Supp. Mat.* Sec. F. In contrast, despite being built on top of VLFM and taking only the instance category, AIUTA effectively gather additional information from the user to identify the correct instance, requiring minimal agent-user interaction (NQ<2 for all splits). This results in a substantial improvement in SR, achieving an outstanding $\sim 14\times$ increase on `Val Seen Synonyms`, $\sim 7\times$ on `Val Seen`, and $\sim 7\times$ on `Val Unseen`. We show AIUTA's question diversity in *Supp. Mat.* Sec. G.

Validation with real human. To validate that simulated user-agent interactions yield credible results, we further conduct evaluation with real human on a small subset of CoIN-Bench. We randomly select 40 episodes with *detectable* target instances across all splits to minimize time and cognitive load. As a result, the SR for this set are higher compared to those reported in Tab. 2. We engage 20 participants of varying ages and backgrounds, each evaluating two episodes. Participants are provided an image depicting the target instance and interact with the agent via a chat-like interface (see **aiuta_demo.mp4** in the supplementary materials). They initiates the navigation via the fixed template "`Find the <category>`", and answer the agent's questions in natural

User type	CoIN-Bench subset		
	SR ↑	SPL ↑	NQ ↓
Simulated	42.50	15.48	1.10
Real Human	42.50	17.44	1.29

Table 3. Real human *vs* simulated user-agent interaction.

Self-Questioner	Skip-Question	Ablation split		
		SR ↑	SPL ↑	NQ ↓
✗	✗	9.21	5.86	3.57
✗	✔	8.55	4.84	2.69
✔	✗	9.87	6.5	4.6
✔	✔	**14.47**	**7.22**	**1.68**

Table 4. Ablation of components in AIUTA on the `Train` split.

language. More details about human evaluation in *Supp. Mat.* Sec. E. The human results is compared against with simulated user-agent interactions in Tab. 3. We observe no significant differences in main metrics, confirming that the simulation setup is *reliable for reproducible evaluation*.

***Ablation I*: AIUTA components.** We introduce the Ablation split, derived from the largest GOAT-Bench `Train` split, following the procedure in Sec. 5. We select GOAT-Bench `Train` as it covers more semantic categories. Since AIUTA is *training-free*, validation remains fair. Tab. 4 highlights the importance of the *Self-Questioner* and *Skip-Question* (within the Interaction Trigger). Without both (row 1), `SR` drops to 9.21%, with a high number of questions `NQ`. Removing only the Self-Questioner (row 2) lowers the `SR`, reducing `NQ`, as expected. Enabling only the Self-Questioner (row 3) improves `SR` to 9.87%, but keeps `NQ` high. With both components active (row 4), `SR` peaks at 14.47%, and `NQ` drops to 1.68, proving both effectiveness and efficiency.

***Ablation II*: VLM uncertainty estimation on IDKVQA.** VLM uncertainty estimation is crucial for the Self-Questioner module, helping the agent to mitigate hallucinations and inaccuracies. For validating these techniques, we introduce IDKVQA, a VQA dataset with 502 questions and 102 images from GOAT-Bench [19]. Each question is answered by three annotators who choose from {`Yes, No, I Don't Know`}, allowing the agent to abstain when information is insufficient. We compare our *Normalized-Entropy*-based technique against three recent techniques: MaxProb (selects the answer with the highest predicted probability); an energy score-based framework for out-of-distribution detection [29]; and LP [70], a recent logistic regression model trained as a linear probe on the logits distribution of the first generated token. Tab. 5 reports the performance using the *Effective Reliability* metric Φ_c proposed in [60]. Our proposed technique achieves the best $\Phi_{c=1}$ score of 21.12, demonstrating its effectiveness.

VLM Model	Selection Function	$\Phi_{c=1}$
LLaVA llava-v1.6-mistral-7b-hf	MaxProb	15.94
	LP [70]	14.01
	Energy Score [29]	20.45
	Normalized Entropy (ours)	**21.12**

Table 5. Results of different selection functions and their corresponding *Effective Reliability* rate $\Phi_{c=1}$ on the IDKVQA dataset.

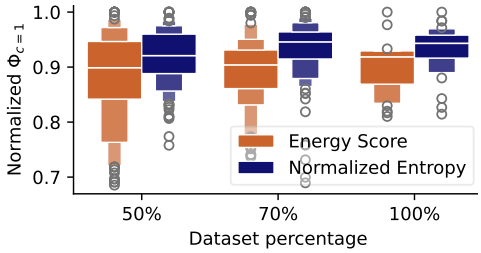

Figure 3. τ sensitivity results. For each method, 30 new τ values are sampled symmetrically around the optimal threshold τ^*. The x-axis shows the set size as a percentage of the original IDKVQA dataset size, while the y-axis displays the normalized ER $\Phi_{c=1}$.

Further details in the *Supp. Mat.* (Sec. H).

***Ablation III*: τ.** We analyze the sensitivity of the threshold (τ in Eq. 3) for our *Normalized-Entropy*-based technique and second-best performing Energy Score [29]. We subsample the datasets to 50%, 70%, and 100% of its original size. For each subsampled dataset, we find the optimal threshold τ^* and evaluate its sensitivity by testing $\Phi_{c=1}$ on 30 alternative thresholds around τ^*, normalizing it between 0 and 1. As shown in Fig. 3, our technique has a smaller interquartile range and a tighter distribution of $\Phi_{c=1}$, while [29] exhibits a greater degradation from τ^*, which worsens as the dataset size decreases. This proves that our technique is more robust in data-scarce situations, and is less sensitive to small variations in τ. Moreover, [29] depends on logits, thus being unbounded. On the contrary, our uncertainty is normalized, *i.e.* $u \in [0, 1]$, making optimal τ selection more efficient.

7. Conclusion

We introduced the CoIN task, where the agent collaborates with the user during navigation to resolve uncertainties about the target instance. Trough extensive experiments, we show that existing trained method fails to generalize to unseen categories, while our training-free AIUTA, using a novel self-dialogue mechanism and uncertainty estimation, performs strongly across all validation splits. Moreover, our simulated user-agent interaction matches human eval, enabling scalable, reproducible experiments. Future works will investigate model optimization for embodied deployment, limiting inference cost, and extend interactions scope to action instructions.

8. Acknowledgment

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work made use of the OpenAI API (Researcher Access program). Moreover, this study was carried out within the PNRR research activities of the consortium iNEST (Interconnected North-East Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS_00000043). This manuscript reflects only the Authors' views and opinions. Neither the European Union nor the European Commission can be considered responsible for them.

References

[1] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On Evaluation of Embodied Navigation Agents. *arXiv preprint arXiv:1807.06757*, 2018. 2, 6, 4

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 2

[4] Luca Barsellotti, Roberto Bigazzi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Personalized Instance-based Navigation Toward User-Specific Objects in Realistic Environments. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 3

[5] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. *arXiv preprint arXiv:2006.13171*, 2020. 2, 6

[6] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *Advances in Neural Information Processing Systems*, pages 4247–4258. Curran Associates, Inc., 2020. 2

[7] Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-tur. Just Ask: An Interactive Learning Framework for Vision and Language Navigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03): 2459–2466, 2020. 3

[8] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, et al. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. *arXiv preprint arXiv:2407.07775*, 2024. 4, 5

[9] Yinpei Dai, Run Peng, Sikai Li, and Joyce Chai. Think, Act, and Ask: Open-World Interactive Personalized Robot Navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3296–3303, 2024. 3, 6, 1

[10] Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, Ranjay Krishna, Dustin Schwenk, Eli VanderBilt, and Aniruddha Kembhavi. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16238–16250, 2024. 3

[11] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 3, 7

[12] Qiaozi Gao, Govind Thattai, Suhaila Shakiah, Xiaofeng Gao, Shreyas Pansare, Vasu Sharma, Gaurav Sukhatme, Hangjie Shi, Bofei Yang, Desheng Zhang, Lucy Hu, Karthika Arumugam, Shui Hu, Matthew Wen, Dinakar Guthy, Shunan Chung, Rohan Khanna, Osman Ipek, Leslie Ball, Kate Bland, Heather Rocker, Michael Johnston, Reza Ghanadan, Dilek Hakkani-Tur, and Prem Natarajan. Alexa Arena: A User-Centric Interactive Platform for Embodied AI. In *Advances in Neural Information Processing Systems*, pages 19170–19194. Curran Associates, Inc., 2023. 2, 3

[13] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S. Sukhatme. DialFRED: Dialogue-Enabled Agents for Embodied Instruction Following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056, 2022. 3

[14] Groq. Groq - Accelerated AI Inference. https://groq.com/, 2024. Accessed: Mar. 7, 2025. 5

[15] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. 6

[16] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and brian ichter. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *Proceedings of The 6th Conference on Robot Learning*, pages 1769–1782. PMLR, 2023. 3

[17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[18] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12), 2023. 3

[19] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. GOAT-Bench: A Benchmark for Multi-Modal Lifelong Navigation. In *CVPR*, page 16373–16383. IEEE, 2024. 2, 3, 6, 7, 8, 4

[20] Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-Specific Image Goal Navigation: Training Embodied Agents to Find Object Instances. *arXiv preprint arXiv:2211.15876*, 2022. 3, 6

[21] Yuxuan Kuang, Hai Lin, and Meng Jiang. OpenFMNav: Towards Open-Set Zero-Shot Object Navigation via Vision-Language Foundation Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 338–351, Mexico City, Mexico, 2024. Association for Computational Linguistics. 3

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 4, 1

[23] Weijie Li, Xinhang Song, Yubing Bai, Sixian Zhang, and Shuqiang Jiang. ION: Instance-level Object Navigation. In *ACM MM*, pages 4343–4352. ACM, 2021. 2

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 5

[25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 7

[26] Qianyi Liu, Siqi Zhang, Yanyuan Qiao, Junyou Zhu, Xiang Li, Longteng Guo, Qunbo Wang, Xingjian He, Qi Wu, and Jing Liu. GroundingMate: Aiding Object Grounding for Goal-Oriented Vision-and-Language Navigation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1775–1784, 2025. 3

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 4

[28] Shi Liu, Kecheng Zheng, and Wei Chen. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. In *Computer Vision - ECCV 2024*. Springer Nature Switzerland, 2025. 2, 4, 5

[29] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, pages 21464–21475. Curran Associates, Inc., 2020. 2, 7, 8, 6

[30] Xiulong Liu, Sudipta Paul, Moitreya Chatterjee, and Anoop Cherian. CAVEN: An Embodied Conversational Agent for Efficient Audio-Visual Navigation in Noisy Environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3765–3773, 2024. 3

[31] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings. In *Advances in Neural Information Processing Systems*, pages 32340–32352. Curran Associates, Inc., 2022. 3, 4

[32] Arjun Majumdar, Fei Xia, Brian Ichter, Dhruv Batra, and Leonidas Guibas. FindThis: Language-Driven Object Disambiguation in Indoor Environments. In *Proceedings of The 7th Conference on Robot Learning*, pages 1335–1347. PMLR, 2023. 3

[33] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, 2018. 5

[34] Khanh Nguyen and Hal Daumé III. "Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning". In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 684–695, Hong Kong, China, 2019. Association for Computational Linguistics. 3

[35] Khanh Nguyen and Hal Daumé III. Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. 3

[36] Khanh Nguyen, Debadeepta Dey, and Bill Brockett, Chrnd Dolan. Vision-Based Navigation With Language-Based Assistance via Imitation Learning With Indirect Intervention. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. 3

[37] Khanh X Nguyen, Yonatan Bisk, and Hal Daumé Iii. A Framework for Learning to Request Rich and Contextually Useful Information from Humans. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16553–16568. PMLR, 2022. 3

[38] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. *arXiv preprint arXiv:2410.02707*, 2024. 3

[39] Jeongeun Park, Seungwon Lim, Joonhyung Lee, Sangbeom Park, Minsuk Chang, Youngjae Yu, and Sungjoon Choi. CLARA: Classifying and Disambiguating User Commands for Reliable Interactive Robotic Agents. *IEEE Robotics and Automation Letters*, 9(2):1059–1066, 2024. 4

[40] Sudipta Paul, Amit Roy-Chowdhury, and Anoop Cherian. AVLEN: Audio-Visual-Language Embodied Navigation in 3D Environments. In *Advances in Neural Information Processing Systems*, pages 6236–6249. Curran Associates, Inc., 2022. 3

[41] Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. How Easy is It to Fool Your Multimodal LLMs? An Empirical Analysis on Deceptive Prompts. In *Neurips Safe Generative AI Workshop 2024*, 2024. 2, 4, 5

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4

[43] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 4, 6, 1

[44] Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. PIRLNav: Pretraining with Imitation and RL Finetuning for OBJECTNAV. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 3

[45] Niyati Rawal, Roberto Bigazzi, Lorenzo Baraldi, and Rita Cucchiara. UNMuTe: Unifying Navigation and Multimodal Dialogue-like Text Generation. *arXiv preprint arXiv:2408.04423*, 2024. 3

[46] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. 5

[47] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. In *7th Annual Conference on Robot Learning*, 2023. 3

[48] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4, 6, 1

[49] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 5

[50] Ying Shen, Daniel Bis, Cynthia Lu, and Ismini Lourentzou. ELBA: Learning by Asking for Embodied Visual Navigation and Task Completion. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 5177–5186, 2025. 3

[51] Kunal Pratap Singh, Luca Weihs, Alvaro Herrasti, Jonghyun Choi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Ask4Help: Learning to Leverage an Expert for Embodied Tasks. In *Advances in Neural Information Processing Systems*, pages 16221–16232. Curran Associates, Inc., 2022. 3

[52] Xander Sun, Louis Lau, Hoyard Zhi, Ronghe Qiu, and Junwei Liang. Prioritized Semantic Learning for Zero-shot Instance Navigation. In *Computer Vision - ECCV 2024*. Springer Nature Switzerland, 2025. 3, 6, 7, 4

[53] Francesco Taioli, Federico Cunico, Federico Girella, Riccardo Bologna, Alessandro Farinelli, and Marco Cristani. Language-enhanced RNR-Map: Querying Renderable Neural Radiance Field maps with natural language. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, page 4671–4676. IEEE, 2023. 3

[54] Francesco Taioli, Francesco Giuliari, Yiming Wang, Riccardo Berra, Alberto Castellini, Alessio Del Bue, Alessandro Farinelli, Marco Cristani, and Francesco Setti. Unsupervised Active Visual Search With Monte Carlo Planning Under Uncertain Detections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11047–11058, 2024. 3

[55] Francesco Taioli, Stefano Rosa, Alberto Castellini, Lorenzo Natale, Alessio Del Bue, Alessandro Farinelli, Marco Cristani, and Yiming Wang. I2EDL: Interactive Instruction Error Detection and Localization. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 1872–1877, 2024. 3

[56] Francesco Taioli, Stefano Rosa, Alberto Castellini, Lorenzo Natale, Alessio Del Bue, Alessandro Farinelli, Marco Cristani, and Yiming Wang. Mind the Error! Detection and Localization of Instruction Errors in Vision-and-Language Navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12993–13000, 2024. 2

[57] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-Dialog Navigation. In *Proceedings of the Conference on Robot Learning*, pages 394–406. PMLR, 2020. 3

[58] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9568–9578. IEEE, 2024. 2, 4, 5

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4

[60] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly. In *Computer Vision – ECCV 2022*, page 148–166. Springer Nature Switzerland, 2022. 8, 5

[61] Karmesh Yadav, Jacob Krantz, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Jimmy Yang, Austin Wang, John Turner, Aaron Gokaslan, Vincent-Pierre Berges, Roozbeh Mootaghi, Oleksandr Maksymets, Angel X Chang, Manolis Savva, Alexander Clegg, Devendra Singh Chaplot, and Dhruv Batra. Habitat Challenge 2023, 2023. 6

[62] B. Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. 'Towards New Computational Principles for Robotics and Automation'*, pages 146–151, 1997. 3

[63] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-Language Frontier

Maps for Zero-Shot Semantic Navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, page 42–48. IEEE, 2024. 3, 4, 6, 7, 8

[64] Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. HM3D-OVON: A Dataset and Benchmark for Open-Vocabulary Object Goal Navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5543–5550, 2024. 3, 6, 7, 1, 4

[65] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3MVN: Leveraging Large Language Models for Visual Target Navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023. 3

[66] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE, 2023. 4

[67] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv preprint arXiv:2306.14289*, 2023. 4

[68] Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. TriHelper: Zero-Shot Object Navigation with Dynamic Assistance. *arXiv preprint arXiv:2403.15223*, 2024. 3

[69] Michael JQ Zhang and Eunsol Choi. "Clarify When Necessary: Resolving Ambiguity Through Interaction with LMs". In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5526–5543, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 3

[70] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The First to Know: How Token Distributions Reveal Hidden Knowledge in Large Vision-Language Models? In *Computer Vision - ECCV 2024*. Springer Nature Switzerland, 2025. 4, 7, 8, 6

[71] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. ESC: Exploration with Soft Commonsense Constraints for Zero-shot Object Navigation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 42829–42842. PMLR, 2023. 3

[72] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions. *Transactions on Machine Learning Research*, 2024. 4

[73] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, page 3357–3364. IEEE, 2017. 2, 3

[74] Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. Self-Motivated Communication Agent for Real-World Vision-Dialog Navigation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021. 3