Commonsense Storage Reasoning in Domestic Scenes: A Challenge for Vision-Language Models

Michaela Levi Richter [1] , **Reuth Mirsky** [1,2] , **Oren Glickman** [1]
[1]Bar-Ilan University, Israel
[2]Tufts University, USA
michaela.levirichter@live.biu.ac.il, oren.glickman@biu.ac.il, reuth.mirsky@tufts.edu

Abstract

To operate effectively in household environments, service robots must reason about object placement not just through direct visual perception, but also by drawing on context, prior knowledge, and commonsense expectations. For example, when asked to locate a spoon in an unfamiliar kitchen, a human might infer it is most likely stored in a drawer near the countertop. Enabling similar reasoning in multimodal models presents a significant challenge.

In this work, we introduce the Stored Household Item Challenge, a benchmark designed to evaluate commonsense spatial reasoning in domestic settings. The task requires models to predict the most likely storage location for a given item—such as a drawer or cabinet door—even when the item is not visible. We release two complementary datasets: a crowdsourced development set containing 6,500 annotated item-image pairs from kitchen scenes, and a real-world test set based on actual item storage in private homes.

We evaluate a range of state-of-the-art models, including vision-language models (Grounding-DINO) and multimodal large language models (Gemini, GPT-4o, Kosmos-2). Our results reveal that many models perform at or below random chance, and none come close to matching human-level performance. This highlights key limitations in current models' ability to integrate visual context with structured commonsense knowledge. By providing this task and dataset, we offer a novel testbed for advancing and benchmarking multimodal reasoning capabilities in real-world environments.

1 Introduction

Service robots and personal assistants are rapidly improving in their ability to perform domestic tasks, from navigating indoor spaces to manipulating household objects. Yet, they continue to struggle with deeper semantic understanding of their environments, particularly when reasoning about what cannot be directly seen. In this work, we address a specific challenge in this domain: detecting the likely storage locations of household items. Humans intuitively avoid searching for a fork in a high cabinet, instead drawing on prior experience and commonsense expectations about how kitchens are typically organized. Similarly, enabling robots or agents to reason about concealed storage spaces, such as drawers and cabinets, demands not just visual context but also the integration of structured knowledge and pragmatic inference. Existing object detection methods rely heavily on visible cues or explicit user input, which limits their applicability when target objects are occluded or hidden from view.

To begin addressing this gap, we introduce a new task: commonsense storage prediction, which involves inferring where an object is most likely stored within a household environment based solely on an image of the scene and the item name. We focus specifically on kitchens (Figure 1, right), where storage conventions are both common and diverse. Our core hypothesis is that item placements follow semantically meaningful patterns that can be learned from annotated data. For example, mugs are more often stored in upper cabinets, while cutlery is typically found in drawers. By leveraging these patterns, it becomes possible to make plausible, human-like predictions about where an object might be stored—even if it is not currently visible.

To support this task, we introduce the *Stored Household Item Challenge*, a new benchmark consisting of 6,500 item-image pairs, created by combining real-world kitchen scenes from a publicly available Scene Understanding dataset, object detections from a vision-language grounding model, and human annotations of likely storage locations. We evaluate several baseline models, including random selection, grounded vision-language models (e.g., Grounding-DINO), and multimodal large language models (e.g., GPT-4o), and compare them to human performance (Figure 1, top and middle left). Notably, many of these models—despite their strong performance on general vision-language tasks struggle to outperform a random baseline, highlighting a critical blind spot in their reasoning capabilities.

This work does not aim to show incremental accuracy gains, but rather to formalize and benchmark a form of spatial commonsense reasoning that is essential for embodied AI systems and is currently underrepresented in existing datasets. The Stored Household Item Challenge captures a class of inference that is underrepresented in existing benchmarks: reasoning about what is likely true but not directly observable. As large vision-language and multimodal models

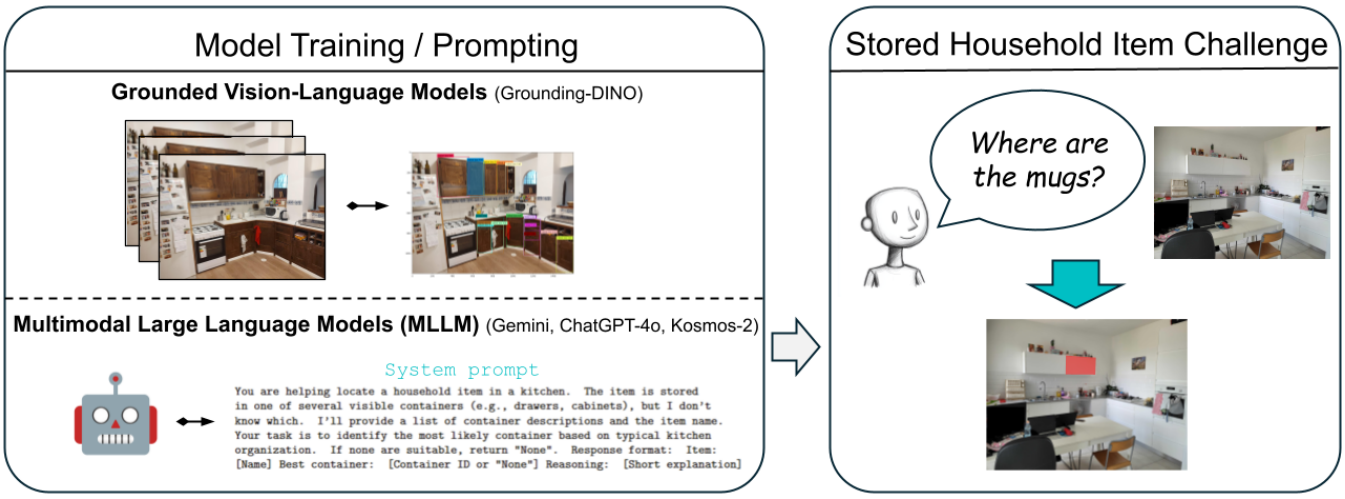

Figure 1: Depiction of the Stored Household Item Challenge (right). We compare the performance of various models on this task (left). This evaluation includes grounded vision-language models and multimodal large language models.

evolve, we expect their performance on this task to improve; however, the challenge itself—requiring the integration of spatial pragmatics, visual grounding, and structured world knowledge—will remain a useful diagnostic for evaluating deeper, human-like understanding in MLLMs and VLMs.

2 Related Work

Our research focuses on the intersection of three topics: Domestic Service Robots, Scene Understanding, and Object Detection. We utilize vision and language models for semantic understanding under object detection.

2.1 Domestic Service Robots

In recent years, the scope of research on cleaning robots has evolved beyond vacuum or floor cleaning robots to encompass service robots capable of performing diverse household tasks such as cleaning, tidying, organizing groceries, washing dishes, and setting tables.

In contrast to our work, existing research on this new generation of robots emphasizes aspects like learning processes tailored to specific houses [Hasegawa *et al.*, 2023; Kim *et al.*, 2019; Shah *et al.*, 2020] and user instructions posed before robot utilization [Kaneda *et al.*, 2024; Khan, 2022; Matsushima *et al.*, 2022; Ribeiro *et al.*, 2021; Wu *et al.*, 2023a; Wu *et al.*, 2023b; Yan *et al.*, 2021]. Notable approaches include using Graph Neural Networks (GNNs) to learn user preferences from observations [Kapelyukh and Johns, 2022] and learning user preferences from similarities between objects using object hierarchies and collaborative filtering [Abdo *et al.*, 2015]. While most research on pick-and-place tasks for service robots primarily addresses open shelves or tables [Hu *et al.*, 2023; Shiba *et al.*, 2023; Xu and Hsu, 2023], our work uniquely focuses on stored objects in enclosed containers. The work by Liu et al. [Liu *et al.*, 2022] represents a knowledge-based framework for object search in service planning, considering multi-domain knowledge and utilizing an ontology-based knowledge structure.

Ramrakhya et al. [Ramrakhya *et al.*, 2024] define a Semantic Placement (SP) task, predicting where an object could be placed in an image based on its name. While their methodology shares similarities with ours, we specifically concentrate on finding stored objects in containers rather than predicting visible placements. Furthermore, our approach is designed to apply to any common household object, while their work specifically targets 9 chosen items. Kant et al. [Kant *et al.*, 2022] focus on rearranging objects in a house based on human preferences without explicit goal specifications. Additionally, the researchers employed a simulator to validate the effectiveness of their model. While their approach also leverages commonsense reasoning in home environments using an LLM, it differs from ours by relying on partial observations and not addressing items that are already stored and thus not visible. Kurenkov's research [Kurenkov, 2023] discusses searching in containers, assuming the agent possesses the layout of the environment, storage locations, and observations of objects when exploring a container for the first time. Our work aims to contribute to both of these studies by reducing the deployment time of service robots in new houses.

Additional work explores contextual semantics for continual sweeping for a fleet of cleaning robots [Ahmadi and Stone, 2006], or for organizing household items based on user preferences [Akanimoh Oluwasanmi Adeleye, 2023; Kant *et al.*, 2022; Wu *et al.*, 2023a; Wu *et al.*, 2023b; Xu and Hsu, 2023]. Additionally, some studies adapt natural language processing (NLP) to the context of object arrangement [Hu *et al.*, 2023]. These works provide valuable insights for our task, facilitating the efficient integration of LLMs by generalizing user preferences without specifying particular ones.

2.2 Scene Understanding

Scene understanding, as articulated by Naseer et al. [Naseer *et al.*, 2019], involves the analytical process of evaluating a visual scene by considering both its geometric and seman-

tic context, along with the intrinsic relationships between its elements. This holistic comprehension is crucial for the seamless integration of robots into daily environments, promoting effective collaboration with humans [Aarthi and Chitrakala, 2017]. Recognizing the limitations of mere object recognition or classification, robots must possess a nuanced, functional understanding of the visual scene to enhance their overall performance [Ye *et al.*, 2017]. Within the broader landscape of scene understanding, subtasks like scene classification play a pivotal role in accurately labeling the elements within a given image [Patel *et al.*, 2020; Tong *et al.*, 2017]. Ongoing research in scene analysis for object understanding by assistive robots aims to simplify computational costs, ensuring a user-friendly interface for individuals to convey their needs [Bousquet-Jette *et al.*, 2017].

Furthermore, the intersections of semantics and Simultaneous Localization and Mapping (SLAM) are relevant to scene understanding. SLAM, a foundational technique in robotics, constructs a map of an unknown environment while concurrently determining the robot's location within that environment [Cadena *et al.*, 2016; Chen *et al.*, 2022; Garg *et al.*, 2020; Singandhupe and La, 2019]. Semantics in robotics enable robots to interpret and represent the meaning of objects and actions. The simultaneous use of semantics and SLAM equips robots with both the semantic context of targeted objects and the spatial mapping essential for comprehensive scene understanding. This approach provides a deep understanding of the environment, offering insights into the probable position of objects within specific frames and contributing crucial information for advancing robotic understanding and facilitating nuanced navigation.

2.3 Object Detection

Object detection aims to identify instances of semantic objects in digital images and videos, addressing the fundamental question of "What objects are here?" [Zou *et al.*, 2023]. Traditional methods, relying on handcrafted features and shallow architectures, face performance limitations, leading to the emergence of powerful deep learning tools capable of learning semantic features with distinctions in architecture, training, and optimization [Zhao *et al.*, 2019]. This study explores image segmentation, encompassing semantic, instance, and panoptic segmentation. Semantic segmentation assigns pixel-level labels to object categories across the entire image, posing a more intricate challenge than whole-image classification [Minaee *et al.*, 2021]. Instance segmentation further enhances detailed scene interpretation by precisely detecting and delineating individual objects of interest [Hafiz and Bhat, 2020]. In our research, we leverage object detection and segmentation to identify *the containers themselves* such as drawers, closet doors, and cabinet doors in images. Specifically, we utilize Grounding-DINO [Liu *et al.*, 2023], an open-set object detector capable of localizing arbitrary objects based on textual prompts, including category names and referring expressions. For segmentation, we use the Segment Anything Model (SAM) [Kirillov *et al.*, 2023a], a prompt-driven segmentation model designed to generalize across diverse image distributions and tasks. We assess performance using standard evaluation metrics for detection and segmentation[Ayala *et al.*, 2024].

Concealed object detection (COD), a method that segments objects with similar patterns (e.g., texture, color, direction) to their natural or man-made environment [Fan *et al.*, 2022], is noteworthy. However, it cannot be leveraged in our task, as that problem assumes that the objects are visible in some location within the scene. Occlusion object detection, dealing with challenges of hidden or partially visible objects, presents solutions in controlled environments and robotics [Saleh *et al.*, 2021]. These are considered impractical for our scenario of stored items, as the item of interest is completely concealed. In specific environments like prisons and airports, monitoring completely occluded objects is essential [Kristoffersen *et al.*, 2016]. Yet, the proposed system's use of internet protocol and thermal cameras is irrelevant to our case, as thermal cameras cannot detect most locations of stored household items.

3 The Stored Household Item Challenge

We present the **Stored Household Item Challenge**, a challenge designed to probe the limitations of current AI systems when it comes to providing assistance in a household environment. This task was deliberately chosen because it encapsulates several core challenges that remain unresolved in today's models. It goes beyond standard object detection by requiring agents to reason about what is not directly visible, using indirect evidence, contextual cues, and prior semantic knowledge — all of which push the boundaries of current visual and vision-language systems.

First, the challenge directly targets the *gap between visual input and semantic understanding*. While many models can localize and label visible containers like drawers or cabinets, they rarely encode what those containers mean in context. The challenge forces systems to bridge this gap: they must map from an image of a household scene to an understanding of which storage locations are most appropriate for a given item, such that mugs are more likely found in upper kitchen cabinets than under the sink. This demands deeper reasoning than detection alone, encouraging models to internalize functional and cultural norms around household organization.

Second, the task is designed to require commonsense priors. It is not enough to extrapolate from local image features; a successful model must incorporate general knowledge about how items tend to be stored across many homes. Unlike traditional supervised learning tasks, where labels correspond directly to what is visible, the stored household item challenge rewards systems that can generalize across scenes and reason probabilistically about plausible storage locations, even with limited or ambiguous visual information. This requirement makes this challenge a natural testbed for combining visual grounding with large-scale pretraining and structured knowledge.

Finally, the challenge emphasizes interpretable and actionable predictions. It is not sufficient to name a likely container type; the model must select a specific instance of that container in the scene — one that is localized, segmentable, and ideally interactable by a robot. This requirement adds a layer of embodied relevance that connects abstract reasoning to ac-

tionable physical outcomes. By grounding each prediction to a concrete location within an image, the task supports real-world robotic use cases where storage inferences must translate into navigation and manipulation actions. Together, these aspects make this challenge a principled and practical test of the next generation of service robots.

4 Datasets and Data Collection

To develop a model capable of predicting where household items are typically stored, we constructed two complementary datasets: a **real-world test dataset** and a **crowdsourced development dataset**. Together, these datasets support both the training and rigorous evaluation of models targeting semantic understanding of storage locations in domestic environments.

The real-world test dataset consists of images collected from participants' actual home environments, where individuals manually annotated the precise storage locations (e.g., specific drawers, cabinets) of common household items. These annotations serve as ground truth labels for evaluating model performance in realistic, privacy-sensitive settings.

In contrast, the crowdsourced development dataset is designed to support model training and validation at scale. It is based on publicly available kitchen images from the SUN dataset [Xiao *et al.*, 2010; Xiao *et al.*, 2016], which avoids the privacy concerns of in-home data collection. We recruited annotators through a crowdsourcing platform to label the most likely storage container for a given item in each scene. To ensure data quality, we performed thorough preprocessing, including duplicate removal, inconsistency filtering, and analysis of inter-annotator agreement. The result is a high-quality dataset suitable for supervised learning and benchmarking, and our pipeline provides a reusable framework for future data collection efforts focused on storage reasoning.

4.1 Dataset Selection and Collection

To train a model to understand "commonsense" household storage norms, we required a dataset that captured where items are typically stored within home environments—specifically in containers such as drawers, cabinets, or closets. As no suitable dataset existed, we opted to build one. We initially considered sourcing data from volunteers who would submit labeled images of their homes. However, this approach proved challenging for two reasons: First, ethically, creating a dataset of images in homes, including kitchens or bedrooms, can potentially hold private information that is hard to obfuscate, such as personal documents or photos. It would require a significant amount of effort to make such a dataset public without risking privacy breaches. Second, from a logistical perspective, collecting a sufficiently large and diverse set of in-home images at the scale required for effective model training proved infeasible. To overcome these challenges, we eventually chose to maintain a smaller volunteer dataset as a ground truth benchmark and focused training on a semi-artificial dataset, by repurposing existing image datasets with additional human annotation.

To further maintain consistency and ensure a focused experimental scope, we restrict this research to kitchen environments, which present a diverse range of storage options compared to other rooms around the house. We selected 15 kitchen items, including both common and less common examples: *bottle opener, Tupperware containers, dish towels, cutting board, bowl, spices, spoon, mug, plate, pot, pan, cutting knife, cooking oil, screwdrivers* and *painkillers*. This selection was used to guide both image sampling and annotation efforts.

4.2 Dataset Preprocessing

For both of the datasets we collected, we applied a detection and segmentation pipeline using Grounding-DINO [Liu *et al.*, 2023] and SAM [Kirillov *et al.*, 2023b] to identify and segment all visible storage containers. There were ~ 16 containers on average in the train dataset and ~ 19 containers on average in the test dataset.

These automated segmentation serves two purposes: First, they provide structured visual cues to assist various algorithms in detecting the correct container. Second, by marking potential storage locations in each scene, we can simply automate the verification of the recognizer's answers (rather than evaluating the correctness of an open-ended answer like "the fork is likely to be in an upper drawer"). An institutional review board (IRB) was obtained for participant-based data collection, covering both datasets.

After surveying available resources, we selected the SUN dataset due to its diversity and relevance to indoor scene understanding [Xiao *et al.*, 2010; Xiao *et al.*, 2016]. We extracted images from the kitchen category. Storage containers within these images were automatically identified using Grounding-DINO and SAM. We then launched a crowdsourcing campaign on Upwork[1] in which human annotators selected the most plausible container for storing a specific household item in each image.

4.3 Development Dataset: Crowdsourced Annotations

Three annotators (one male, two female) were recruited, two from the U.S. and one from Ireland. A custom web application, hosted on GitHub pages, facilitated the annotation process. The interface presented the name of a household item alongside an image with marked storage containers. The interface can be viewed in Figure 2. Annotators either selected the most likely container or indicated that the item would not typically be stored in any of the shown containers. Total annotation times per participant were: 7.2, 9.6, and 10.2 hours per annotator.

From an initial pool of 1,746 kitchen images, we selected 1,656 that contained at least three detectable storage containers suitable for annotation. Anticipating the use of this dataset for training purposes, we withheld two items from our original 15-item list to enable future evaluation of model generalizability to previously unseen items. For each of the remaining 13 items, we sampled 500 images, yielding a total of 6,500 unique item-image pairs.

To assess annotation quality, 16% of the images for each item (80 out of 500) were labeled by all three annotators, while the remaining 420 images were evenly divided, with

[1] https://www.upwork.com/

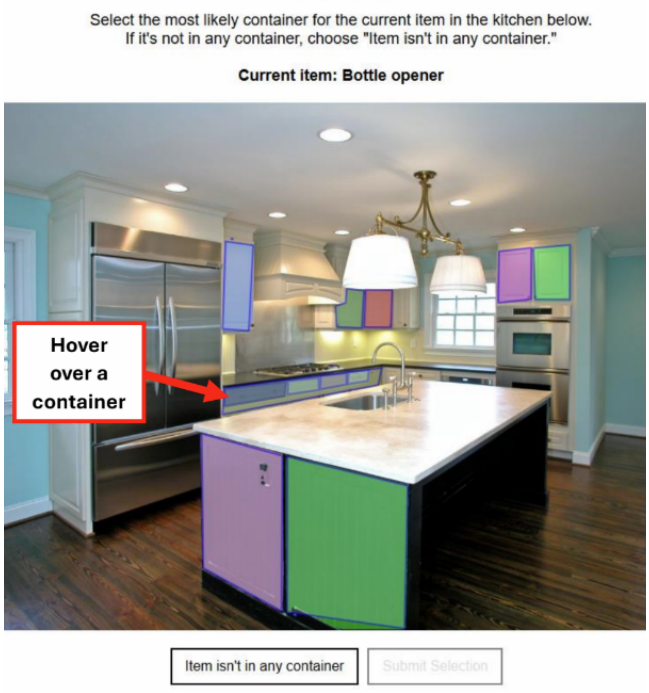

Figure 2: A screenshot from the annotation tool used to collect human-labeled data.

each annotator labeling 140. Annotation was conducted in four batches to enhance usability within the web application and maintain quality—each new batch was released only after satisfactory performance on the previous one. In total, each annotator completed 2,860 annotations, including 1,040 shared annotations. This process yielded 5,460 unique annotations and 3,120 overlapping ones. In total, we collected 8,580 item-image annotations, from which we constructed a training dataset of 6,500 pairs, as detailed in Section 4.5.

4.4 Real-World Evaluation Data

In addition to the crowdsourced annotations, we collected a small set of real-world examples submitted by volunteers. These included images of home interiors, each paired with a labeled storage container for a specific item. Participants were informed that their photos would be made public, and they were instructed to hide any personal items or information they did not want to be published. The images were further manually checked for any personal identifiers, such as documents or personal photos.

From the list of 15 items mentioned earlier, participants provided storage locations for four items: *bottle openers, Tupperware containers, painkillers, and screwdrivers*. The latter two are the ones that were intentionally excluded from the development dataset to be able to test the models' generalizability. In total, the evaluation dataset comprises 100 item-image pairs collected from 74 distinct kitchens, with some participants contributing locations for multiple items.

This dataset served dual purposes: it was used to validate annotator reliability and will also be used to compare against future model predictions.

4.5 Data Cleaning and Processing

To ensure dataset quality, we first removed duplicate entries and conflicting responses due to server lag in recording users' responses. We then consolidated annotations from multiple users: if two annotators agreed, we retained the majority vote. If no agreement was reached, one label was selected at random.

Preliminary analysis revealed variation in annotator agreement across different object categories. Overall, the Fleiss' Kappa inter-annotator agreement was $\kappa = 0.372$ for all items in the real-world evaluation set, and $\kappa = 0.354$ (fair agreement) for the development dataset. As only 16% of each item's item-image pairs in the development set were annotated by all annotators, the presented agreement only covers that segment of the data. Items with more standardized storage locations, such as bottle openers and plates, exhibited higher consistency in responses with moderate agreement ($\kappa = 0.494$ and $\kappa = 0.478$, respectively). In contrast, more subjective categories like Tupperware containers showed slight agreement ($\kappa = 0.155$). These results highlight both the inherent difficulty and variability of the task, while also suggesting that people rely on more than random reasoning when searching for hidden household items.

5 Experimental Setup

We provide a preliminary evaluation of current state-of-the-art models on this task, using the real-world evaluation data described in Section 4.4. We tested two categories of models: grounded vision-language models (Grounding-DINO) and multimodal large language models (MLLMs) (Gemini, ChatGPT-4o, and Kosmos-2). To train all models, we used the development dataset presented in Section 4.3. Approximately 5% of the dataset (370 out of the 6,500 item-image pairs) was used to train the vision-language models, and the remainder was used for evaluation and parameter tuning of all models. All baseline models receive an image and a prompt as input and return a bounding box. Models were prompted with a query about a specified household item, and were evaluated on their ability to identify the correct storage location among drawers and cabinet doors. Predicted bounding boxes were compared to ground truth annotations using Intersection over Union (IoU) and binary accuracy thresholds.

All experiments were conducted on a standard Intel i7 machine, while model training requiring a GPU (e.g., Kosmos-2) was performed on a university computing cluster.

5.1 Evaluation Metrics

Each model's performance was evaluated separately on the development and evaluation datasets. Model performance was assessed using the following metrics:

- **Accuracy:** The percentage of predictions with IoU ≥ 0.2 compared to ground truth.
- **IoU (Intersection over Union):** The overlap between the predicted and true bounding boxes, computed as:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

5.2 Baseline Comparisons

Random For each item-image pair, we took the list of containers provided by Grounding-DINO during the annotation phase and picked one at random.

Kosmos-2 We used the Kosmos-2 MLLM with an image and the following prompt format:

```
<grounding> In which<phrase>
drawer</phrase> or<phrase>
cabinet door</phrase> is<phrase>
a {item}</phrase> stored?
```

In the prompt above, `{item}` is a placeholder that is dynamically replaced with the name of the queried object (e.g., `spoon`, `plate`, `bottle opener`) for each example.

The model outputs captions and entity bounding boxes relevant to the query. Since Kosmos-2 does not use a pre-defined list of container locations (unlike human annotators), we considered a prediction correct if the IoU between the model's output and the ground truth was ≥ 0.2 (meaning that the bounding box overlapped with the true container in at least 20% of its area).

Grounding-DINO and SAM We used Grounding-DINO in combination with SAM for detection and segmentation. The prompt was:

```
drawer for {item}.  cabinet door
for {item}.
```

Notice that this baseline is a second, separate use of Grounding-DINO from the first use for data annotation: In the annotation phase, we used Grounding-DINO to output all polygons representing cabinets, closets, and drawers. Here, we ask the model to output a bounding box for a stored item. From the resulting set of polygons, we selected the one with the highest confidence score and compared it to the ground truth. Because these models were also used during dataset construction, we had to ensure that we did not provide these models with an unfair bias during evaluation. Thus, we only considered a prediction correct for these models if IoU was exactly 1.0.

Gemini (Google MLLM) We used the Gemini API to evaluate vision-language performance. As with Kosmos-2, predictions with IoU ≥ 0.2 were considered correct. For each item-image pair, we submitted the image with the following prompt:

```
Analyze the provided image of
a kitchen.  Identify the item:
'{item}'.  Determine the most
likely storage location for this
item, considering only drawers
or cabinet doors visible in the
image.  Provide the bounding
box coordinates as a Python
list of four integers:  [x_min,
y_min, x_max, y_max].  If you
cannot determine a likely storage
location, output:  [0, 0, 0, 0]
```

ChatGPT-4o We used the GPT-4o model with vision capabilities through the OpenAI API. The prompt format was:

```
You are analyzing a kitchen
image.
The visible containers are
drawers and cabinet doors.
The item to store is:  {item}
Determine the most likely storage
location among visible drawers
and cabinet doors only.
Return a list of 4-point bounding
box coordinates.
If no suitable location is found,
return an empty list [].
```

As in the other models, predictions were counted as correct if the IoU with the ground truth polygon was ≥ 0.2.

6 Results and Analysis

We begin by reporting the overall performance of the various algorithms on the real-world evaluation dataset. As shown in Table 1, Grounding-DINO substantially outperforms all other approaches. To further assess whether the observed performance differences between models were statistically significant, we conducted a one-way ANOVA on accuracy scores, followed by post-hoc pairwise comparisons using the Bonferroni correction. The results confirmed a significant effect of model type on performance ($p < 0.05$). Moreover, no model was able to reach results similar to those of the human annotators ($p < 0.05$).

We continue by reporting the accuracy and IoU results on the annotated development dataset, as shown in Table 2. While our primary focus is on evaluation set results, the development set performance helps reveal whether a model has learned meaningful patterns or is simply underfitting. For example, models like Kosmos-2 and Gemini achieve relatively low accuracy even on the development set (lower than random from a given set of containers), suggesting limited capacity or misalignment with the task rather than overfitting. Including development results thus helps distinguish between models that fail to generalize and those that fail to learn effectively in the first place. Human performance on this dataset is estimated based on the subset of overlapping annotations used to assess inter-annotator agreement, and may slightly differ from actual human performance on the full development set.

Table 1: Evaluation Set Accuracy and Average IoU for Various Models.

Model	**Accuracy (%)**	**Average IoU**
Human Annotator 1 (IoU = 1.0)	38.00	0.380
Human Annotator 2 (IoU = 1.0)	27.00	0.271
Human Annotator 3 (IoU = 1.0)	36.00	0.361
Grounding-DINO (IoU = 1.0)	**13.00**	**0.188**
Random (IoU = 1.0)	6.00	0.062
Kosmos-2 (IoU $\geq$ 0.2)	4.00	0.042
Gemini-1.5-flash (IoU $\geq$ 0.2)	3.00	0.034
GPT-4o API (IoU $\geq$ 0.2)	8.00	0.082

Table 2: Development Set Accuracy and Average IoU for Various Models.

Model	**Accuracy (%)**	**Average IoU**
Human Annotator (Estimated)	35.83	0.361
Grounding-DINO (IoU = 1.0)	5.89	0.112
Random (IoU = 1.0)	7.55	0.083
Kosmos-2 (IoU $\geq$ 0.2)	2.25	0.013
Gemini-1.5-flash (IoU $\geq$ 0.2)	8.65	0.083
GPT-4o API (IoU $\geq$ 0.2)	13.78	0.138

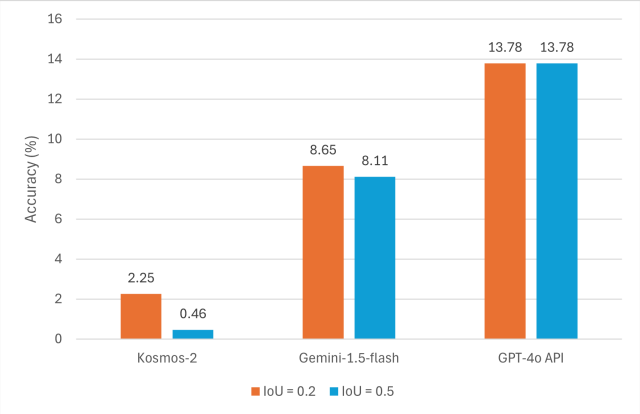

Figure 3: Accuracy of the different models using different intersection over union (IoU) values.

Lastly, to validate our evaluation framework further, we analyze the effect of varying the Intersection over Union (IoU) threshold on model accuracy. Recall that all of the evaluated models output a bounding box that may or may not overlap with the container bounding box provided by Grounding-DINO for annotation. As expected, stricter IoU thresholds impose more demanding localization criteria, making it harder for models to achieve high accuracy. Conversely, lower thresholds (e.g., IoU $\geq$ 0.2) allow for more lenient matches. Figure 3 illustrates this trade-off for the models evaluated above. Notably, the results remain relatively consistent across thresholds, suggesting that our findings are robust to the exact choice of IoU cutoff and reinforcing the reliability of our evaluation methodology.

7 Conclusion and Future Work

This paper introduces the Stored Household Item Challenge, a novel problem designed to evaluate semantic spatial reasoning about the likely locations of non-visible objects in household environments. Unlike conventional object detection tasks, this challenge targets a critical but underexplored capability: inferring hidden item locations based on context and commonsense knowledge. We present a new dataset of labeled item-image pairs featuring concealed storage scenarios, along with a suite of baseline evaluations. Preliminary results reveal a significant gap between current state-of-the-art methods and human-level reasoning.

Our work highlights several key directions for future research. First, integrating this task into physical robots would bring the challenge closer to real-world deployment, introducing new considerations such as actuation constraints and real-time perception. Richer scene understanding may also emerge through interactive or multimodal inputs and adaptation to user preferences. Second, future models may benefit from learning hierarchical spatial priors, such as typical room layouts or common sub-region uses (e.g., “mugs are often stored above counters”). Another promising direction is embedding this task in embodied systems, allowing robots to refine their beliefs about storage through feedback and exploration.

Although this paper focuses on kitchens, the underlying reasoning principles extend to a broad range of domestic and industrial environments. We plan to expand the dataset and task to include additional spaces and item categories. By advancing hidden object reasoning, this work supports the development of service robots that operate effectively in unfamiliar, dynamic human environments.

Finally, deploying these models on physical robots introduces practical challenges, including perception noise, varied cabinet designs, shifting viewpoints, and the physical cost of interaction. For example, a robot may need to weigh the likelihood of an item being in a distant, hard-to-reach cabinet against a closer drawer that is slightly less probable. Balancing prediction confidence with physical effort will be essential for efficient, real-world deployment.

Ethical Statement

This work involves no experiments with human subjects, personal data, or animals. All visual data used for training and evaluation was sourced from the publicly available SUN dataset, which consists of images collected from the internet for scene understanding research. To avoid potential privacy concerns associated with collecting in-home photos, we did not solicit or share any private images of participant homes beyond a small, non-public evaluation set. Annotations for the larger dataset were obtained via crowdsourcing using a standardized task description and anonymized images, and responses were screened for quality and consistency. The study design ensures that all human annotation work was voluntary, compensated fairly, and involved no sensitive or identifying information.

References

[Aarthi and Chitrakala, 2017] S. Aarthi and S. Chitrakala. Scene understanding — a survey. In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 1–4, 2017.

[Abdo *et al.*, 2015] Nichola Abdo, Cyrill Stachniss, Luciano Spinello, and Wolfram Burgard. Robot, organize my shelves! Tidying up objects by predicting user preferences. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1557–1564, Seattle, WA, USA, May 2015. IEEE.

[Ahmadi and Stone, 2006] Mazda Ahmadi and Peter Stone. A multi-robot system for continuous area sweeping tasks. In *International Conference on Robotics and Automation, ICRA.*, pages 1724–1729. IEEE, 2006.

[Akanimoh Oluwasanmi Adeleye, 2023] Henrik Christensen Akanimoh Oluwasanmi Adeleye. *Enabling Assistive Service Robots to Contextually Organize Household Objects*. PhD thesis, UC San Diego, 2023.

[Ayala *et al.*, 2024] Christian Ayala, Carlos Aranda, and Mikel Galar. Guidelines to Compare Semantic Segmentation Maps at Different Resolutions. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

[Bousquet-Jette *et al.*, 2017] C. Bousquet-Jette, Sofiane Achiche, Dominique Beaini, Yann-Seing Law-Kam Cio, Cédric Leblond Ménard, and Maxime Raison. Fast scene analysis using vision and artificial intelligence for object prehension by an assistive robot. *Engineering Applications of Artificial Intelligence*, 63:33–44, August 2017.

[Cadena *et al.*, 2016] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Transactions on Robotics*, 32(6):1309–1332, December 2016.

[Chen *et al.*, 2022] Weifeng Chen, Guangtao Shang, Aihong Ji, Chengjun Zhou, Xiyang Wang, Chonghui Xu, Zhenxiong Li, and Kai Hu. An Overview on Visual SLAM: From Tradition to Semantic. *Remote Sensing*, 14(13):3010, June 2022.

[Fan *et al.*, 2022] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, October 2022.

[Garg *et al.*, 2020] Sourav Garg, Niko Sünderhauf, Feras Dayoub, Douglas Morrison, Akansel Cosgun, Gustavo Carneiro, Qi Wu, Tat-Jun Chin, Ian Reid, Stephen Gould, Peter Corke, and Michael Milford. Semantics for robotic mapping, perception and interaction: A survey. *Foundations and Trends® in Robotics*, 8(1–2):1–224, 2020.

[Hafiz and Bhat, 2020] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International Journal of Multimedia Information Retrieval*, 9(3):171–189, September 2020.

[Hasegawa *et al.*, 2023] Shoichi Hasegawa, Akira Taniguchi, Yoshinobu Hagiwara, Lotfi El Hafi, and Tadahiro Taniguchi. Integrating probabilistic logic and multimodal spatial concepts for efficient robotic object search in home environments. *SICE Journal of Control, Measurement, and System Integration*, 16(1):400–422, 2023.

[Hu *et al.*, 2023] Yuhang Hu, Zhizhuo Zhang, Ruibo Liu, Philippe Wyder, and Hod Lipson. Knolling bot: A Transformer-based Approach to Organizing a Messy Table, October 2023.

[Kaneda *et al.*, 2024] Kanta Kaneda, Shunya Nagashima, Ryosuke Korekata, Motonari Kambara, and Komei Sugiura. Learning-to-rank approach for identifying everyday objects using a physical-world search engine, 2024.

[Kant *et al.*, 2022] Yash Kant, Arun Ramachandran, Sriram Yenamandra, Igor Gilitschenski, Dhruv Batra, Andrew Szot, and Harsh Agrawal. Housekeep: Tidying Virtual Households Using Commonsense Reasoning. In *Computer Vision – ECCV 2022*, pages 355–373. Springer Nature Switzerland, Cham, 2022.

[Kapelyukh and Johns, 2022] Ivan Kapelyukh and Edward Johns. My house, my rules: Learning tidying preferences with graph neural networks. In *5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 740–749. PMLR, 08–11 Nov 2022.

[Khan, 2022] Tareq Khan. Towards a Low-Cost Object Collecting and Organizing Household Robot using Deep Learning. *European Journal of Electrical Engineering and Computer Science*, 6(6):16–25, November 2022.

[Kim *et al.*, 2019] Jaeseok Kim, Anand Kumar Mishra, Raffaele Limosani, Marco Scafuro, Nino Cauli, Jose Santos-Victor, Barbara Mazzolai, and Filippo Cavallo. Control strategies for cleaning robots in domestic applications: A comprehensive review. *International Journal of Advanced Robotic Systems*, 16(4), July 2019.

[Kirillov *et al.*, 2023a] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[Kirillov *et al.*, 2023b] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything, 2023.

[Kristoffersen *et al.*, 2016] Miklas Kristoffersen, Jacob Dueholm, Rikke Gade, and Thomas Moeslund. Pedestrian Counting with Occlusion Handling Using Stereo Thermal Cameras. *Sensors*, 16(1):62, January 2016.

[Kurenkov, 2023] Andrey Kurenkov. *Manipulation and Reasoning Methods for Embodied Object Search*. PhD thesis, Stanford University, 2023.

[Liu *et al.*, 2022] Shaopeng Liu, Guohui Tian, Ying Zhang, Mengyang Zhang, and Shuo Liu. Service planning oriented efficient object search: A knowledge-based frame-

work for home service robot. *Expert Systems with Applications*, 187, January 2022.

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

[Matsushima *et al.*, 2022] Tatsuya Matsushima, Yuki Noguchi, Jumpei Arima, Toshiki Aoki, Yuki Okita, Yuya Ikeda, Koki Ishimoto, Shohei Taniguchi, Yuki Yamashita, Shoichi Seto, Shixiang Shane Gu, Yusuke Iwasawa, and Yutaka Matsuo. World robot challenge 2020 – partner robot: a data-driven approach for room tidying with mobile manipulator. *Advanced Robotics*, 36(17-18):850–869, September 2022.

[Minaee *et al.*, 2021] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image Segmentation Using Deep Learning: A Survey. *Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[Naseer *et al.*, 2019] Muzammal Naseer, Salman Khan, and Fatih Porikli. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE Access*, 7, 2019.

[Patel *et al.*, 2020] Tanvi A. Patel, Vipul K. Dabhi, and Harshadkumar B. Prajapati. Survey on Scene Classification techniques. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 452–458, Coimbatore, India, March 2020. IEEE.

[Ramrakhya *et al.*, 2024] Ram Ramrakhya, Aniruddha Kembhavi, Dhruv Batra, Zsolt Kira, Kuo-Hao Zeng, and Luca Weihs. Seeing the unseen: Visual common sense for semantic placement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16273–16283, 2024.

[Ribeiro *et al.*, 2021] Tiago Ribeiro, Fernando Gonçalves, Inês S. Garcia, Gil Lopes, and António F. Ribeiro. CHARMIE: A Collaborative Healthcare and Home Service and Assistant Robot for Elderly Care. *Applied Sciences*, 11(16):7248, August 2021.

[Saleh *et al.*, 2021] Kaziwa Saleh, Sandor Szenasi, and Zoltan Vamossy. Occlusion Handling in Generic Object Detection: A Review. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000477–000484, Herl'any, Slovakia, January 2021. IEEE.

[Shah *et al.*, 2020] Rishi Shah, Yuqian Jiang, Justin Hart, and Peter Stone. Deep R-Learning for Continual Area Sweeping. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5542–5547. IEEE, October 2020.

[Shiba *et al.*, 2023] Tomoya Shiba, Tomohiro Ono, and Hakaru Tamukoh. Object Search and Empty Space Detection System for Home Service Robot. *Proceedings of International Conference on Artificial Life and Robotics*, 28:400–403, February 2023.

[Singandhupe and La, 2019] Ashutosh Singandhupe and Hung Manh La. A Review of SLAM Techniques and Security in Autonomous Driving. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, Naples, Italy, February 2019. IEEE.

[Tong *et al.*, 2017] Zhehang Tong, Dianxi Shi, Bingzheng Yan, and Jing Wei. A Review of Indoor-Outdoor Scene Classification. In *2nd International Conference on Control, Automation and Artificial Intelligence (CAAI 2017)*, Sanya, China, 2017.

[Wu *et al.*, 2023a] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, November 2023.

[Wu *et al.*, 2023b] Zhanxin Wu, Bo Ai, and David Hsu. Integrating common sense and planning with large language models for room tidying. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[Xiao *et al.*, 2010] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[Xiao *et al.*, 2016] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *Int. J. Comput. Vision*, 119(1):3–22, August 2016.

[Xu and Hsu, 2023] Yiqing Xu and David Hsu. " tidy up the table": Grounding common-sense objective for tabletop object rearrangement. *arXiv preprint arXiv:2307.11319*, 2023.

[Yan *et al.*, 2021] Zhi Yan, Nathan Crombez, Jocelyn Buisson, Yassine Ruichck, Tomas Krajnik, and Li Sun. A Quantifiable Stratification Strategy for Tidy-up in Service Robotics. In *Advanced Robotics and Its Social Impacts (ARSO)*, pages 182–187, Tokoname, Japan, July 2021. IEEE.

[Ye *et al.*, 2017] Chengxi Ye, Yezhou Yang, Ren Mao, Cornelia Fermüller, and Yiannis Aloimonos. What can i do around here? deep functional scene understanding for cognitive robots. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4604–4611. IEEE, 2017.

[Zhao *et al.*, 2019] Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu. Object Detection With Deep Learning: A Review. *Neural Networks and Learning Systems*, 30(11):3212–3232, November 2019.

[Zou *et al.*, 2023] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, 111(3):257–276, March 2023.