HOLM: Hallucinating Objects with Language Models for Referring Expression Recognition in Partially-Observed Scenes

Volkan Cirik[1] **Louis-Philippe Morency**[1] **Taylor Berg-Kirkpatrick**[2]
[1]Carnegie Mellon University [2]University of California San Diego
{vcirik,morency}@cs.cmu.edu {tberg}@eng.ucsd.edu

Abstract

AI systems embodied in the physical world face a fundamental challenge of partial observability; operating with only a limited view and knowledge of the environment. This creates challenges when AI systems try to reason about language and its relationship with the environment: objects referred to through language (e.g. giving many instructions) are *not immediately visible*. Actions by the AI system may be required to bring these objects in view. A good benchmark to study this challenge is Dynamic Referring Expression Recognition (dRER) task where the goal is to find a target location by dynamically adjusting the field of view (FoV) in a partially observed 360° scenes. In this paper, we introduce HOLM, **H**allucinating **O**bjects with **L**anguage **M**odels, to address the challenge of partial observability. HOLM uses large pre-trained language models (LMs) to infer object hallucinations for the unobserved part of the environment. Our core intuition is that if a pair of objects co-appear in an environment frequently, our usage of language should reflect this fact about the world. Based on this intuition, we prompt language models to extract knowledge about object affinities which gives us a proxy for spatial relationships of objects. Our experiments show that HOLM performs better than the state-of-the-art approaches on two datasets for dRER; allowing to study generalization for both indoor and outdoor settings.

1 Introduction

One of the fundamental challenges in building AI systems physically present in the world is addressing the issue of partial observability, the phenomenon where the entire state of the environment is not known or available to the system. People cope with partial observability by reasoning about what is not immediately visible (see example in Figure 1). People combine their general knowledge about the world and adapt their knowledge

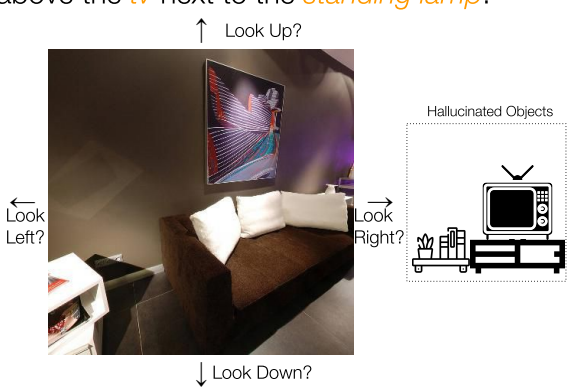

Figure 1: Illustration of our main contribution: **Hallucinating Objects**. Knowledge about object relationships is helpful when navigating in an unknown and partially observed environment. In the example above, the TV is not visible, but the couch hints that a TV might be in front of it because usually couches face TVs.

to specific contexts (Torralba et al., 2006). General knowledge about kitchens can help to know approximately where to look for pans or utensils in a kitchen that has never been seen before. How can an AI system build general knowledge about objects and their environment to help with a similar task? Even more interestingly, can we gather this information from language, using readily available resources such as language models trained on a large collection of unlabeled text?

In this paper, we introduce a method called HOLM, **H**allucinating **O**bjects with **L**anguage **M**odels, for reasoning about the unobserved parts of the environment. Inspired by the recent successes of large pre-trained language models (LM) extracting knowledge about the real world, we propose a methodology based on spatial prompts to extract knowledge from language models about object. HOLM extracts spatial knowledge about objects in the form of affinity scores, i.e., how often a pair of objects are observed together. This knowledge of objects are combined with observed spatial

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5440 - 5453

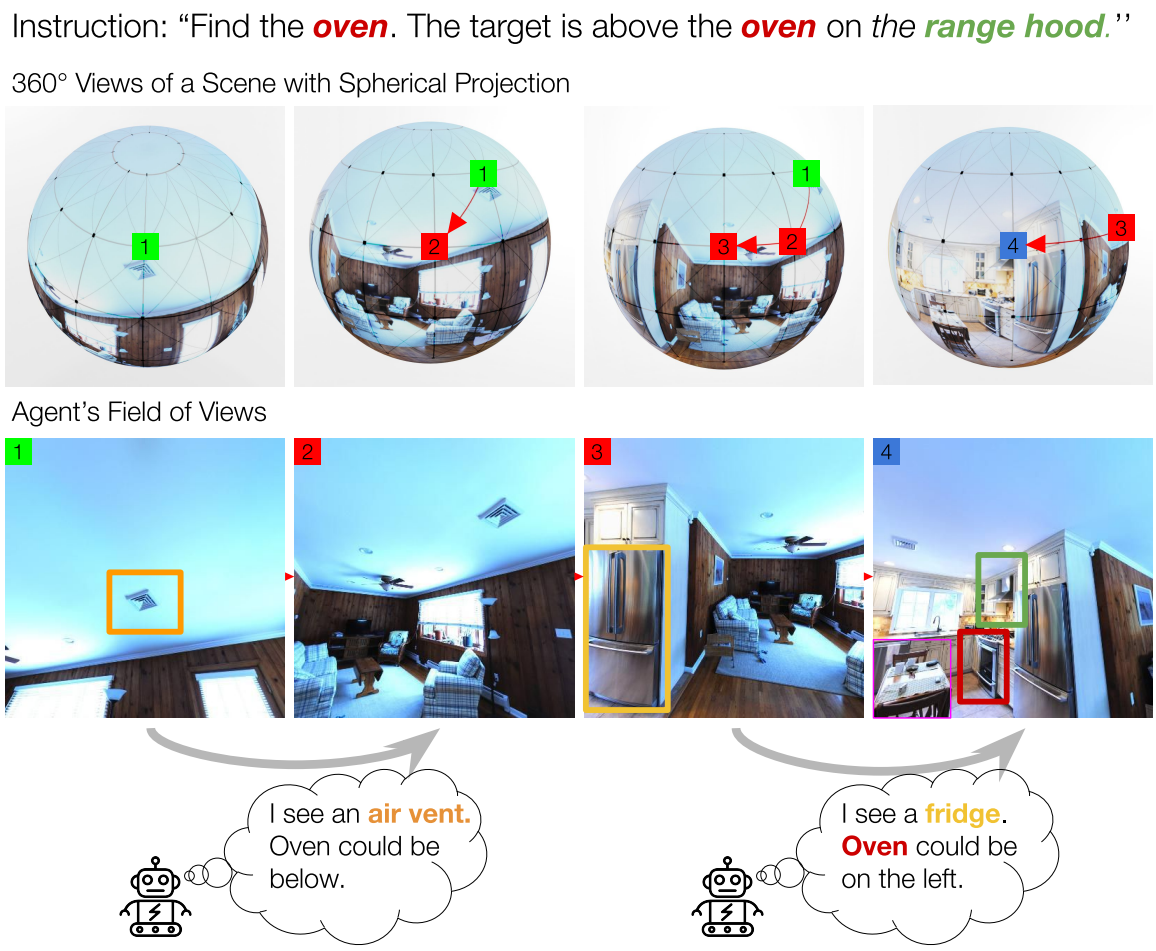

Figure 2: **Illustration of the dRER task with an example of language instruction and its recognition in four steps.**
The agent adjusts its FoV by looking at different directions and navigate on the graph in the spherical view. Note that objects mentioned in bold in the instruction are not visible at all until timestep 4. Thus, the agent needs to reason about possible locations of the mentioned object using its partial view of the scene.

layout to hallucinate what might appear in the unobserved part of the scene. We evaluate our HOLM approach on Dynamic Referring Expression Recognition (dRER) task where the goal is to find a target location by dynamically adjusting the field of view (FoV) in partially observed 360° scenes. We examine how HOLM compares with the state-of-the-art approaches on two publicly available datasets to study generalization for both indoor and outdoor settings.

2 Dynamic Referring Expression Recognition (dRER) Task

dRER task is designed to localize a target location in a dynamically observed 360° scene given natural language instruction. Unlike conventional referring expression recognition, which refers to an object in a static visual input, in dRER, only a small part of the scene is visible in a field of view. However, the system can adjust the field of view to find the described point in the scene. In Figure 2, we illustrate the dRER task and motivate our method. On top, natural language instruction is given. In the middle, the spherical view of the scene is illustrated – the agent explores only some portion of a 360° scene. FoVs on the sphere represented as square nodes form a graph. By *navigating to a neighboring node*, the agent adjusts its FoV and observes a different view of the scene. Note that objects mentioned in the instruction "oven" and "range hood" are not visible until the fourth timestep. However, we can reason about where to look using visible objects such as the air vent or the fridge. Thus, to perform well on this task, it is essential to reason about where objects might appear.

The dRER task can be formulated as a Markov Decision Process (MDP) (Howard, 1960) $\mathcal{M} =$

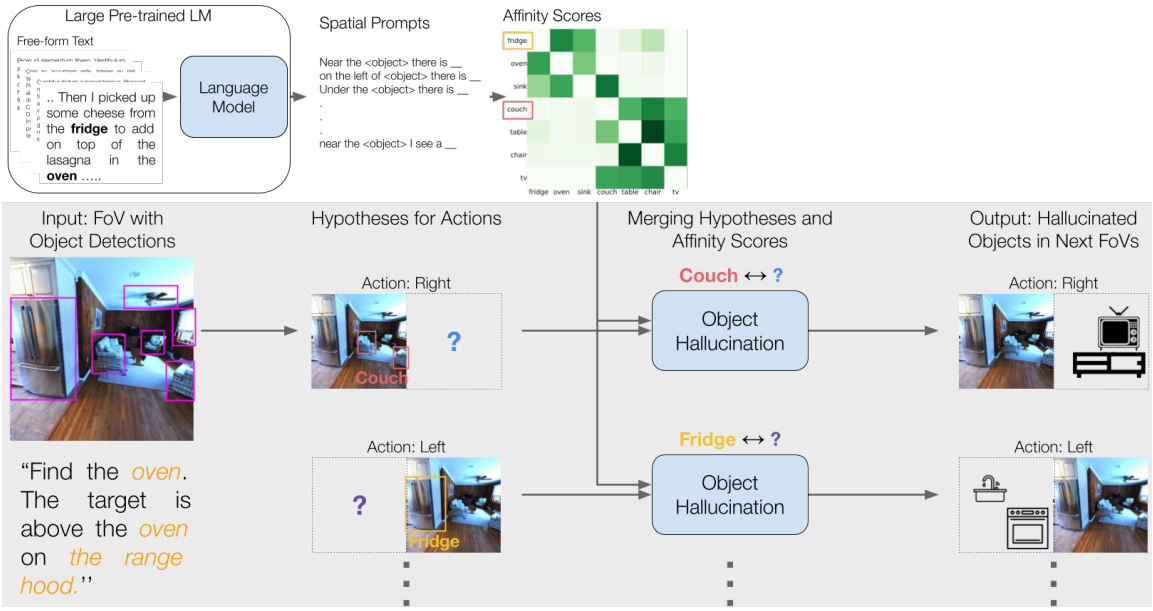

Figure 3: **HOLM for the dRER task**. **(Top)** We use language models trained on a large amount of text by prompting with the spatial relationship of objects to calculate co-occurrence statistics of objects. **(Bottom)** The flow of our hallucination method. We determine objects of interest for each action. Then, we combine objects of interest and co-occurrence table to hallucinate objects, i.e. what might appear after performing an action.

$\langle \mathcal{S}, \mathcal{A}, P_s, r \rangle$ where $\mathcal{S}$ is the visual state space, $\mathcal{A}$ is the discrete action space [1], P_s is the unknown environment probability distribution from which the next state is drawn, and $r \in \mathbb{R}$ is the reward function. For a time step t, the agent observes an image $s_t \in \mathcal{S}$, and performs and action $a_t \in \mathcal{A}$. As a result of this action, the environment generates a new observation $s_{t+1} \sim P_s(\cdot \mid s_t, a_t)$ as the next state. This interaction continues sequentially and ends when the agent performs a special `STOP` action or a pre-defined maximum episode length is reached. The resolution process is successful if the agent ends the episode at the target location.

In dRER, instructions are represented as N sequence of sentences represented as $x = \{x_i\}_{i=1}^{N}$. Each instruction sentence x_i consists of a sequence of L_i words, $x_i = [x_{i,1}, x_{i,2}, ..., x_{i,L_i},]$. The training dataset $\mathcal{D}_E = \{\mathcal{X}, \mathcal{T}\}$ consists of M pairs of the instruction sequence $x \in \mathcal{X}$ and its corresponding expert trajectory $\tau \in \mathcal{T}$. The agent learns to navigate by learning a policy π via maximum likelihood estimation (MLE):

$$\begin{aligned} \max_{\theta} \mathcal{L}_\theta(\mathcal{X}, \mathcal{T}) \text{ , where} \\ \mathcal{L}_\theta(\mathcal{X}, \mathcal{T}) = \log \pi_\theta(\mathcal{T}|\mathcal{X}) \\ \mathcal{L}_\theta(\mathcal{X}, \mathcal{T}) = \frac{1}{M} \sum_{k=1}^{M} \log \pi_\theta(\tau^k|x^k) \end{aligned} \tag{1}$$

3 HOLM

In dRER, the system observes the current FoV and does not see the resulting FoV before taking any actions. Thus, it is essential to reason what might appear in a future observation using what is currently visible to the system. Our core intuition is that objects visible in the current FoV and their locations in the FoV give us a clue about what might appear if a particular action is taken. Here, we propose an approach for reasoning about future observations using what is visible and some background knowledge of objects. Let us go through the illustration in Figure 3 to explain our HOLM method. In the top panel, we feed spatial prompts to pre-trained language models to extract knowledge about objects in the form of affinity scores. In the bottom panel, we see the input of the system where there are natural language instructions, an FoV of the scene, and detected objects. Next, we calculate which objects are relevant to each action. For instance, couch detections are on the right side;

[1] For computational efficiency, we picked discrete action space. It could be continuous as well.

thus, they are relevant to the right action. Similarly, the fridge is relevant for the left action because it is on the left side. Then on the third step, using the affinity score of a pair of objects, we predict what might appear after performing an action. For right action, our model hallucinates a tv and tv-stand might appear because the couch and tv have a high affinity score according to the LM.

3.1 Affinity Scores from Language Models

Language models process a large amount of text to learn regularities in natural language. They do so by predicting the next word or masked token given a sequence of words. Our intuition is that objects that frequently appear in an environment close to each other will have similar language usage. Thus, we hypothesize that language models' capability of learning affinity scores of words in language also reflects objects' spatial properties. In Figure 3's top panel, we illustrate how we extract this capability. We query language models trained on a large amount of free-form text with spatial relationship prompts. These spatial prompts aim to capture the usage of words when they appear together in the world. An example of these prompt templates is "Near the o_1, there is ___" where $o_1 \in O$ is an object label where O is a set of object labels. If object o_1 co-occurs with o_2 with high frequency, the language model would provide a high probability for the phrase "Near the o_1, there is o_2". Using all pairs in O and K[2] spatial templates, we generate queries q. We then calculate affinity scores C_{o_1,o_2}, i.e., observing o_2 when o_1 is present as follows:

$$C_{o_1,o_2} = \sum_{i=1}^{K} p_{\text{LM}}(o_2|q_i) \tag{2}$$

Where $p_{\text{LM}}(o_2|q)$ is a language model that calculates the probability of observing a token o_2 given a prefix sequence of tokens q.

3.2 Object Hallucination

Our main idea behind HOLM is to reason about what might be observed in a future observation by combining (1) which objects are visible in the current observation and (2) what we know about the spatial properties of those objects. We explain the details of our approach in this section.

Let $p_a \in \mathbb{R}^{|O|}$ be the vector of probabilities of observing an object among a set of all objects O after performing an action a. We calculate p_a as follows:

$$p_a = (p_{\text{FoV}} \odot \mathbb{1}_a)C \tag{3}$$

Where $p_{\text{FoV}} \in \mathbb{R}^{|O|}$ is a vector of confidence values for objects detected in the current FoV. We use an off-the-shelf object detection system (Anderson et al., 2018a) to calculate p_{FoV}. C is the affinity scores of size $|O| \times |O|$. C represents how often a pair of object appear in a spatial relationship and represents the background knowledge of objects. $\mathbb{1}_a \in \{0,1\}^{|O|}$ is a binary vector representing spatially related objects for a direction a. This vector is calculated with an indicator function to determine whether an object is spatially related to action a.

We calculate the indicator function as follows. First, we separate the FoV into 4 imaginary regions called quadrants where each quadrant determines how a region in observed FoV is spatially relevant for canonical directions (i.e., up, down, left, right). In other words, quadrants are "hot-spots" for each direction i.e., the left side of the image is more relevant to the right side of the image if we are interested in what might appear on the left. For 8 directions (left, right, down, up, down-left, down-right, up-left, up-right), we calculate how much each objects' bounding box overlaps with these quadrants. If intersection-over-union is above a fixed threshold we keep this object for the hallucination process.

4 Experiments

We designed our experiments to study and evaluate our proposed HOLM approach under five different research questions. **RQ1:** What is the performance of HOLM when compared to other state-of-the-art approaches? **RQ2:** what is the impact of LM as a source of knowledge for HOLM when compared to other more conventional sources (e.g., images)? **RQ3:** How essential are external sources of data for learning knowledge about objects compared to in domain data? **RQ4:** How accurate is HOLM for predicting objects in future observations? **RQ5:** How do annotation-free language-based knowledge sources i.e., LMs and word embeddings compare for HOLM?

The following section explains the details of experimental setup. Our results are presented and discussed in Section 4.2.

[2]Please see Appendix A.1 for the full list of spatial prompt templates.

4.1 Experimental Setup

To study the research questions previously mentioned, we used two publicly available datasets and state-of-the-art methods as baselines to compare with.

Datasets. We selected the following two datasets to see if our method generalizes to both indoor and outdoor settings. The Refer360° dataset (Cirik et al., 2020) consists of 17K natural language instructions and ground-truth trajectory pairs for localizing a target point in 360° scenes. The ground-truth trajectories are annotated by human annotators in the form of successive FoVs in partially observed 360° scenes. The dataset uses a subset of the SUN360 dataset (Xiao et al., 2012) as the source of scenes and these scenes are from both indoor and two outdoor locations.

Touchdown (Chen et al., 2018) consists of 9K natural language instruction and ground-truth location pairs for 360° scenes on Google Streetview. Unlike the Refer360° dataset, Touchdown does not have expert trajectories – only expert predictions for the target location are provided. Thus, we generated ground-truth trajectories by calculating shortest path trajectories between a randomly selected starting point [3] and the target location.

Baselines Models. We compare our method with the state-of-the-art models and also few simple baselines (i.e., no parameter learning).

- The Self Monitoring Navigation Agent (SMNA) (Ma et al., 2019) model is trained with a co-grounding module where both visual and textual input is attended at the same time. The agent also measures its progress with a progress monitor module.
- FAST (Ke et al., 2019) stands for Frontier Aware Search with backTracking. The FAST model learns to score partial trajectories of an agent for efficiently backtracking to a previous location after a mistake.
- Speaker-Follower (Fried et al., 2018) uses a sequence-to-sequence speaker model to rerank a follower model's candidate trajectories. This pragmatic reasoning model has been shown to improve navigation agents' performance significantly.
- LingUNet (Misra et al., 2018) is an image-to-image encoder-decoder model for learning image-to-image mappings conditioned on language. We should emphasize that, unlike the previous methods, LingUNet is not a navigation model; instead, it predicts regions over an image.
- RANDOM agent randomly picks an action.
- STOP agent predicts the starting FoV as the target FoV.

For a fair comparison, the same model was used as the basis for all the compared models. For our proposed approach HOLM is used to enhance the SMNA baseline by hallucinating objects for unseen regions. After getting object hallucinations for each neighboring FoVs, we use the sum of word embeddings for object labels as the input representation for the neighboring FoV. In the oracle "Next FoV" scenario, we use ground-truth FoVs to do the same process. For a fair comparison, we use SMNA as the base agent for learning to recover from a mistake during navigation process with FAST and as the follower model for pragmatic reasoning with Speaker-Follower.

Evaluation Metrics. Our main evaluation metric for methods is FoV accuracy: the percentage of the time the target location is visible in the final FoV. The FoV accuracy sets an upper bound on the localization accuracy for predicting the pixel location of the target point, i.e., if the target is not visible, it is impossible to predict the exact location. Thus, we focus on this metric to compare systems.

Implementation. All models are trained for 100K iterations. We use Adam (Kingma and Ba, 2015) for optimization with a learning rate 0.0001 and weight decay parameter 0.0005 (Krogh and Hertz, 1992). For each model, we perform a grid-search over their hyperparameters (e.g., number of hidden units, number of layers, dropout rate) and pick the best performing model based on validation score [4]. All models are implemented using PyTorch (Paszke et al., 2019) and publicly available[5].

To speed up the training procedure, we used fixed a grid of FoVs for all 360° images where each FoV is connected to its neighboring FoVs. This grid forms the navigation graph depicted in

[3] Following (Cirik et al., 2020), we set the initial random point to be a fix heading and random yaw.

[4] For Refer360° we use validation unseen split. Touchdown does not have seen-unseen distinction.

[5] https://github.com/volkancirik/HOLM

Method	Oracle	Refer360°	Touchdown
Stop Agent		14.1	0.0
Random Agent		12.1	6.8
SMNA (Ma et al., 2019)		27.1	45.9
+ HOLM (this work)		**32.2**	**49.8**
SMNA (Ma et al., 2019)	Next FoV	33.5	50.2
LingUNet* (Chen et al., 2018)	Full Panorama	21.4	47.2

Table 1: FoV accuracy results for Refer360° and Touchdown with no hallucination baseline, best performing models, and Next FoV oracle model, i.e. the ability to look ahead for neighbor FoVs, and observing full 360° scenes. Our method outperforms the baseline models from the literature.

the Figure 2. We use 30° of separation between successive FoVs which provides enough overlap to reveal relevant information about successive FoVs yet distant enough so that the model needs to reason about future steps. We then pre-calculated the rectilinear projection of each of the FoVs on the grid for all scenes.

4.2 Results and Discussion

In this section we present and discuss experimental results and analyses.

(RQ1) HOLM Improves performance. Our main results are presented in Table 1. In the first row block, we see that simple non-learning baselines fail to perform on the dRER. In the second row block, we compare our method with the baseline where the agent does not have any visual input from the next FoVs. HOLM improves the baseline by hallucinating objects for the next FoVs. In the third row block, we provide results for oracle scenarios. For SMNA, we feed ground-truth FoV as the input of the system. This result sets the upper bound on HOLM, because it cannot achieve better hallucination than the ground-truth FoVs. However, HOLM achieves pretty close to this upper bound and show that it can provide useful predictions for this task. For LingUNet, we feed the full 360° scenes as the visual input. Since LingUNet is not a navigation agent i.e. predicts the target location using full 360° scenes, we calculate FoV accuracy by drawing an FoV around the prediction, which explains '*'.

In Table 2, we compare HOLM with FAST and Speaker-Follower methods, both of which use beam search. During the beam search, these methods use multiple trajectories while deciding on a trajectory. However, this is not plausible in a real-world scenario, i.e. a robot would not generate many trajectories before performing action. HOLM, on the other hand completes the task on a single trajectory while predicting possible future states. FAST improves SMNA for Touchdown but not for Refer360° , which might be due to the richness of scenes in Refer360° whereas in Touchdown , the scenes are always in the same domain. Speaker-Model's decreases the score for SMNA possibly due to the Speaker models' poor performance where the BLEU score is around 6. HOLM consistently improves for both datasets and does not perform any expensive look-ahead operations such as beam search.

Method	Beam Search	Refer360°	Touchdown
Baseline SMNA (Ma et al., 2019)		27.1	45.9
+ HOLM (this work)		**+5.1**	+3.9
+ FAST (Ke et al., 2019)	✓	-6.4	**+4.7**
+ Speaker-Follower (Fried et al., 2018)	✓	-4.6	-11.1

Table 2: FoV accuracy results for Refer360° and Touchdown for methods using beam search or single candidate trajectory. HOLM consistently improves the baseline and does not use multiple trajectories.

Knowledge Type	Human Annotation	Affinity Scores	Refer360°	Touchdown
Baseline	✓	Uniform	27.8	45.2
Baseline	✓	Identity	29.3	45.9
Visual	✓	VisualGenome	30.8	48.4
Knowledge Base	✓	WordNet	29.5	48.4
Pre-trained LM		XLM	**32.2**	**49.8**

Table 3: FoV accuracy results for Refer360° and Touchdown for different methods for calculating affinity scores for HOLM. XLM-based affinity scores achieve the best performance.

(RQ2) Pre-trained LM produces better affinity scores compared to other sources. In Table 3, we compare several baseline methods for calculating the affinity scores. First, we use uniform (i.e., each object pair has the same affinity score) and identity (i.e., object x can only have affinity score with itself) baselines. We also study calculating affinity scores using data annotated by humans. First, we use object annotations in VisualGenome (Krishna et al., 2017). VisualGenome provides a large collection of fine-grained annotations for objects and their spatial relationships. Second, ideally we would like to use human annotations for calculating the affinity score. However, this requires annotation of $|O|^2$ annotations. Instead, as a proxy, we use WordNet (Miller, 1995), a knowledge-base hierarchy annotated by experts. We use NLTK (Bird et al., 2009) to calculate the WordNet similarity to extract the affinity scores between ob-

jects. XLM-based HOLM achieves the best results among these baselines. This result shows that without using human annotations, we can extract useful knowledge about objects using pre-trained LMs.

Method	Data Source	Refer360°	Touchdown
HOLM with XLM	External	**32.2**	**49.8**
HOLM with Objects Counts	Internal	30.3	48.7
Hallucinating with 3-Layer MLP	Internal	27.5	46.3

Table 4: FoV accuracy results for Refer360° and Touchdown when task data is used for object hallucination. The limitation of the domain data can be addressed using external resources such as pre-trained LMs.

(RQ3) External sources may provide better information compared to task data. In Table 4, we compare methods that only use task data for object hallucination and HOLM with external sources such as pre-trained LM. For the second row in the table), we use the BUTD model (Anderson et al., 2018a) to annotate training images with object bounding boxes. Using bounding boxes of objects, we calculate affinity scores. For the third row in the table, we design a model that takes FoV and an object type as an input and predicts a direction (i.e., hallucinate where it might appear) as output. We pass the final feature map layer of 152-layer ResNet (He et al., 2016) as input to a 3-layer feed-forward neural network to predict objects that might appear in neighboring FoVs. This model achieves an F1 score of 40.3 for direction prediction. Both of these methods improve over the SMNA baseline but are worse than the pre-trained LM. This result indicates that task data may have limitations, and external sources such as a pre-trained LM may provide a signal for knowledge about objects.

Knowledge Type	Affinity Scores	Refer360°	Touchdown
Visual	VisualGenome	P 1.4 R 55.3 F1 2.7	P 1.5 R 55.2 F1 2.9
Knowledge Base	WordNet	P 1.3 R 55.4 F1 2.6	P 1.4 R 55.3 F1 2.8
Pre-trained LM	XLM	P 2.0 R 49.5 **F1 3.9**	P 2.2 R 63.2 **F1 4.3**

Table 5: Precision (P), Recall (R), and F1 scores for Refer360° and Touchdown for hallucinating objects in neighboring FoVs. Similar to the downstream task results, pre-trained LM performs the best.

(RQ4) Accuracy of HOLM translates to dRER So far, we measure the performance of HOLM for the downstream dRER task. We can also measure how accurate HOLM is at predicting the presence of an object in neighboring FoVs. We annotate each neighboring ground-truth FoVs with detections from BUTD. If the p_a^i for object $o_i \in O$ is above $\frac{1}{|O|}$, we count that as a prediction of an object in the neighboring FoV after performing action a. In Table 5, we provide precision, recall, and F1 score for the performance of different methods for calculating affinity scores for HOLM. XLM achieves the best performance among the methods we compare. We conclude that the performance for the intrinsic task (i.e., predicting the presence of objects) translates to dRER performance.

Method	Model	Refer360°	Touchdown
Baseline	SMNA	27.1	45.9
WE	+ HOLM with FastText (Mikolov et al., 2018)	31.6	46.8
WE	+ HOLM with GloVe (Pennington et al., 2014)	31.0	49.2
WE	+ HOLM with word2vec (Mikolov et al., 2013)	29.3	46.2
LM	+ HOLM with GPT3 (Brown et al., 2020)	31.1	46.3
LM	+ HOLM with Roberta (Liu et al., 2019c)	30.3	46.0
LM	+ HOLM with XLM (Conneau and Lample, 2019)	**32.2**	**49.8**

Table 6: FoV accuracy results for Refer360° and Touchdown for models processing unlabeled text. **WE** and **LM** are abbreviations for word embeddings and language models. All hallucination-based methods perform better than the baseline. XLM achieves the best performance in both datasets.

(RQ5) Both word embeddings and LMs are good sources of general knowledge of objects In Table 6, we compare word embedding methods and different language models. We use cosine similarities between pairs of objects to calculate the affinity scores. For language models, we compare Open AI's GPT3 (Brown et al., 2020) using their online API[6]. We use Transformers Library (Wolf et al., 2020) for RoBERTa (Liu et al., 2019c) and XLM (Conneau and Lample, 2019). All methods consistently improve over the baseline SMNA model, however, we achieve the best performance using XLM. This result indicates that we can extract useful knowledge about objects with methods relying on large amount of unlabeled text.

5 Related Work

Our work on dRER is closely related to previous studies focusing on Referring Expression Recognition (RER), Vision-and-Language Navigation (VLN), and methods we propose are related to pre-training language models for vision-and-language tasks, model-based reinforcement learning, and co-occcurrence modeling for computer vision. We review these studies in this section.

RER is the task of localizing a target object or a point in an image described by a natural language expression. The most of existing datasets

[6]https://beta.openai.com/

poses the task in 2D images with objects as being the target (Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016; Strub et al., 2017; Liu et al., 2019a; Akula et al., 2020; Chen et al., 2020). Several lines of work are proposed to address RER (Mao et al., 2016; Nagaraja et al., 2016; Yu et al., 2016; Hu et al., 2016; Fukui et al., 2016; Luo and Shakhnarovich, 2017; Liu et al., 2017; Yu et al., 2017; Zhang et al., 2018; Zhuang et al., 2018; Deng et al., 2018; Yu et al., 2018; Cirik et al., 2018; Liu et al., 2019b).

In Touchdown (Chen et al., 2018) and Refer360° (Cirik et al., 2020) the target is a point not an object in a 360° image. In the dRER setup, we also use 360° images of Touchdown and Refer360° , but we do not provide the full panoramic view of the scene. Instead, in a more realistic scenario, the agent observes a partial and dynamic view of the scene, i.e. the agent needs to adjust its FoV to find the target location. Closer to our work, in REVERIE (Qi et al., 2020b) an embodied setup is proposed where the agent needs to first navigate to a location where the target object is visible. Similar to Touchdown and Refer360° , at the final position, the full 360° view is visible to the agent. Unlike ours and similar to 2D image-based RER, the target is an object rather than a point in the scene.

VLN is a vision-and-language task where an agent in a simulated environment observes a visual input and is given a natural language instruction to navigate to a target location. The earlier work (MacMahon et al., 2006; Shimizu and Haas, 2009; Chen and Mooney, 2011) studies the task with synthetic images or in a very small scale (Vogel and Jurafsky, 2010). Anderson et al. (2018b) proposes Room-to-room (R2R) benchmark and revisit VLN task with a modern look. In R2R, the agent observes panoramic scans of a house (Chang et al., 2017) and needs to carry out the natural language instruction. EnvDrop (Tan et al., 2019) model shows generalization to unseen environments by dropping visual features. PREVALENT (Hao et al., 2020) tackles the data sparsity problem with a pre-training scheme. Hong et al. (2021) show that a pre-trained multi-modal can be enhanced with a memory state for the VLN task by recurrently feeding a contextualized state feature after each time step. dRER also poses a navigation task where locations in physical space in VLN correspond to FoVs in a fixed location. In dRER, a trajectory of the agent corresponds to its resolution process for finding the goal location.

Pre-trained models for Vision-and-Language has been recently studied after the huge success of transformer-based models (Vaswani et al., 2017) in NLP (Devlin et al., 2018; Liu et al., 2019c; Conneau and Lample, 2019; Sun et al., 2019b; Poerner et al., 2020; Raffel et al., 2020; Brown et al., 2020). Numerous studies extend these approaches to the multimodal domain (Tan and Bansal, 2019; Lu et al., 2019; Sun et al., 2019a; Su et al., 2020; Li et al., 2020; Qi et al., 2020a; Hu and Singh, 2021). They achieve the-state-of-the-art results in several tasks such as image captioning, text-to-image retrieval, or referring expression recognition. Our work differs from these studies in the sense that the previous approaches use large scaled paired image-text data (Chen et al., 2013; Divvala et al., 2014; Sadeghi et al., 2015; Radford et al., 2021; Jia et al., 2021) to learn efficient representations (Frome et al., 2013; Kottur et al., 2016) for visual and textual modalities whereas we are interested in spatial information learned in unimodal text representations.

Language priors for vision were explored in recent studies. Lu et al. (2016) use word embeddings in a language module to learn a representation for a object-predicate-object triplet for visual relationship detection task. Kiela et al. (2019) propose an approach to extend pre-trained transformer-based LMs for multimodal tasks. Similarly, Lu et al. (2021); Tsimpoukelli et al. (2021) show that pre-trained LMs can be finetuned to perform well in few-shot settings for image classification and open-domain Visual Question Answering (Marino et al., 2019). Marino et al. (2021) also show that multimodal transformer architectures capture implicit knowledge for a pair of objects. Our work differs from these studies (1) we use only unimodal models, (2) we do not finetune models – we do not update models during training. The most similar work to ours, Scialom et al. (2020) show that pre-trained LMs can perform reasonably well on Visual Question Generating (Yang et al., 2015; Mostafazadeh et al., 2016) out of the box. One difference is that we use object labels rather than object features or the appearance of objects to query the language model; however, they use object features as a visual token to the language model. Prompts we use in our work shares similarities with prompts designed in PIQA (Paranjape et al., 2021), but our work is evaluated in a multimodal setup. In con-

trast, PIQA is evaluated for textual commonsense reasoning tasks.

Hallucination idea is also related the work on predicting future observations in long horizons (Villegas et al., 2019) which has been studied in the context of learning planning (Hafner et al., 2019) and acquiring skills for control problems (Hafner et al., 2020), and efficient policy learning (Ha and Schmidhuber, 2018), and vision-and-language navigation (Koh et al., 2021). All these approaches are interested in longer horizons; however, in our work, we study predicting single-step future observation. More recent work (Hu et al., 2021; Rombach et al., 2021; Rockwell et al., 2021) study view synthesis from a single visual observation. Unlike these approaches, HOLM does not generate pixel-level views rather abstractions of views with object labels.

Affinity scores are mainly studied in computer vision tasks in the form of object co-occurrences. Previous studies have shown that object co-occurrences are efficient representations of visual prior for object categorization for object segmentation (Rabinovich et al., 2007; Galleguillos et al., 2008; Ladicky et al., 2010) and zero shot object-recognition (Mensink et al., 2014), and scene understanding (Wu et al., 2014). Our work differs from these studies: we do not calculate co-occurrence statistics, i.e. we do not count the frequency of times they appear together; instead, we calculate a probability measure using language models.

6 Conclusion

In this paper, we introduced HOLM – a model that can extract prior knowledge about objects from LMs and hallucinate objects in future observations. Our experiments showed that HOLM approach improves over various baselines from the literature. Surprisingly, our model which used background knowledge from LMs outperformed models with knowledge from human-annotated data showing that LMs learn useful knowledge about the world without requiring any visual observations. We also showed that out approach generalizes to both indoor and outdoor scenarios.

Our work has limitations in the following ways. First, the hallucination process solely conditions on the current field of view. However, the instruction and the previous observations are available to the system. Conditioning on these sources of information could improve the hallucination accuracy by getting more targeted information from the language model. Second, we assume a fixed lexicon of object labels for hallucination. For both the visual side i.e., the object detector, and the language side i.e., the language model, when an unknown object appears the system cannot use this object for hallucination. Another issue is the scalability, i.e, the affinity scores scale with $O(N^2)$ where N is the number of objects, which might be challenging when N is large. We hope the follow-up work could address these limitations.

Future work will explore the use of background knowledge in other domains such as vision-and-language navigation (Anderson et al., 2018c) and dialog (Thomason et al., 2020). We also believe background knowledge of objects would be handy in complex scenarios such as manipulating objects in a simulated environment (Shridhar et al., 2020). Our method examines extracting background knowledge in a zero-shot manner. However, the literature shows that learning how to prompt could be helpful in finding better (Liu et al., 2021). We strictly compared unimodal approaches for hallucination. Future work extend our work by comparing multimodal models (Tan and Bansal, 2019; Lu et al., 2019; Sun et al., 2019a; Su et al., 2020; Li et al., 2020; Qi et al., 2020a; Hu and Singh, 2021).

Another interesting direction would be to study the capability of transferring knowledge from indoor to outdoor settings and vise versa. Finally, the success of PREVALENT (Hao et al., 2020) and other pre-training approaches for VLN could stem from their ability to *implicitly* encode prior knowledge about objects. Hopefully, future studies examines this phenomenon.

Acknowledgements

This material is based upon work partially supported by National Science Foundation awards 1722822 and 1750439, and National Institutes of Health awards R01MH125740, R01MH096951 and U01MH116925. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred. We also thank anonymous reviewers of ACL Rolling Review for their valuable feedback.

References

Arjun R Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. *arXiv preprint arXiv:2005.01655*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018c. 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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*.

David Chen and Raymond Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25.

Howard Chen, Alane Shur, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2018. Touchdown: Natural language navigation and spatial reasoning in visual street environments. *arXiv preprint arXiv:1811.12354*.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE international conference on computer vision*, pages 1409–1416.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. 2020. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095.

Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2020. Refer360: A referring expression recognition dataset in 360: A referring expression recognition dataset in 360 images images. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7189–7202.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2018. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. 2014. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3314–3325. Curran Associates, Inc.

Andrea Frome, Greg S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: a deep visual-semantic embedding model. In *Proceedings of*

the 26th International Conference on Neural Information Processing Systems-Volume 2, pages 2121–2129.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas. Association for Computational Linguistics.

Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. 2008. Object categorization using co-occurrence, location and appearance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2455–2467.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2020. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 2019. v. In *International Conference on Machine Learning*, pages 2555–2565. PMLR.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. A recurrent vision-and-language bert for navigation. In *CVPR*.

Ronald A Howard. 1960. Dynamic programming and markov processes.

Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. 2021. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12528–12537.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*.

Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha Srinivasa. 2019. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6741–6749.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. 2021. Pathdreamer: A world model for indoor navigation.

Satwik Kottur, Ramakrishna Vedantam, José MF Moura, and Devi Parikh. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4985–4994.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Anders Krogh and John A Hertz. 1992. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957.

Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip HS Torr. 2010. Graph cut based inference with co-occurrence statistics. In *European conference on computer vision*, pages 239–253. Springer.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv e-prints*, pages arXiv–2107.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. 2019a. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4185–4194.

Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019b. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1950–1959.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*.

Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019. Self-monitoring navigation agent via auxiliary progress estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. *Def*, 2(6):4.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. 2014. Cosrarta: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2441–2448.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3D environments with visual goal prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2667–2678, Brussels, Belgium. Association for Computational Linguistics.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany. Association for Computational Linguistics.

Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. 2016. Modeling context between objects for referring expression understanding. In *ECCV*.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-bert: Efficient-yet-effective entity embeddings for bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 803–818.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020a. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

Yuanka Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020b. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. 2007. Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Chris Rockwell, David F. Fouhey, and Justin Johnson. 2021. Pixelsynth: Generating a 3d-consistent experience from a single image. In *ICCV*.

Robin Rombach, Patrick Esser, and Björn Ommer. 2021. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

Thomas Scialom, Patrick Bordes, Paul-Alexis Dray, Jacopo Staiano, and Patrick Gallinari. 2020. What bert sees: Cross-modal transfer for visual question generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 327–337.

Nobuyuki Shimizu and Andrew Haas. 2009. Learning to follow navigational route instructions. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron C. Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019a. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. 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 5100–5111.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.

Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *arXiv preprint arXiv:2106.13884*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. 2019. High fidelity video prediction with large stochastic recurrent neural networks. *Advances in Neural Information Processing Systems*, 32:81–91.

Adam Vogel and Dan Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 806–814.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chia-Chien Wu, Farahnaz Ahmed Wick, and Marc Pomplun. 2014. Guidance of visual attention by semantic information in real-world scenes. *Frontiers in psychology*, 5:54.

Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2012. Recognizing scene viewpoint using panoramic place representation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2695–2702. IEEE.

Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. 2015. Neural self talk: Image understanding via continuous questioning and answering. *arXiv preprint arXiv:1512.03460*.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166.

Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. 2018. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261.

A Appendix

This section presents details omitted in the main document.

A.1 Spatial Prompts

We use a fixed set of spatial prompts to query pretrained language models. The list is in Table 7

near the `object` there is
near the `object` I see a
near the `object` there should be a
the object near the `object` is
on the left of `object` there is
on the right of `object` there is
on top of `object` there is
under the `object` there is
across the `object` there is
close the `object` there is

Table 7: Spatial Prompt Templates