Learning a Visually Grounded Memory Assistant

Meera Hahn[1,2] Kevin Carlberg[2] Ruta Desai[2] James Hillis[2]

[1]Georgia Institute of Technology [2]Facebook Reality Labs

meerahahn@gatech.edu, {carlberg, rutadesai, jmchillis}@fb.com

ABSTRACT

We introduce a novel interface for large scale collection of human memory and assistance. Using the 3D Matterport simulator we create a realistic indoor environments in which we have people perform specific embodied memory tasks that mimic household daily activities. This interface was then deployed on Amazon Mechanical Turk allowing us to test and record human memory, navigation and needs for assistance at a large scale that was previously impossible. Using the interface we collect the 'The Visually Grounded Memory Assistant Dataset' which is aimed at developing our understanding of (1) the information people encode during navigation of 3D environments and (2) conditions under which people ask for memory assistance. Additionally we experiment with with predicting when people will ask for assistance using models trained on hand-selected visual and semantic features. This provides an opportunity to build stronger ties between the machine-learning and cognitive-science communities through learned models of human perception, memory, and cognition.

KEYWORDS

Assistance, Navigation, Visual Memory, Visual Question Answering

1 INTRODUCTION

Automated interaction with humans in everyday activity remains a significant challenge for artificial intelligence (AI). Current interactive systems take many forms and operate on different time scales; examples include shopping recommendations, conversational AI, autonomous vehicles, and social robots. These models typically only work within a narrow range of conditions. For example, while autonomous vehicles can interact effectively with other drivers in highway conditions, they are not yet ready for city streets where environments are less predictable. An even more significant challenge is presented by the prospect of all-day wearable augmented reality (AR) glasses: the ideal is an automated system that that can offer assistance in any context. Unlike existing mobile devices, AR wearable glasses could have access to information from the internet *and* detailed information about local physical context, including user actions from a first person viewpoint. Simultaneous access to both sources of information opens new opportunities and challenges for the development of collaborative human–AI systems. The AI required for such a system would likely require "theory of mind" similar to that of humans, whereby people infer goals and cognitive states of others and take strategic, context-dependent actions that may be cooperative or adversarial in nature.

Correspondence: meerahahn@gatech.edu

Here we present a dataset aimed at providing insight into the conditions where people request assistance to recall facts about the local environment. We focus on memory because enhancing human memory is one of the primary uses of computing technology. Our data provides insight into (1) the kind of features people encode during navigation, (2) the difficulty of different types of questions, and (3) conditions under which people will ask for assistance.

To gain insight into the kind of local assistance people would like to receive, we gave participants exposure to a 3D environment (a "fly-through" of a Matterport3D environment [7]). They were then asked questions about the environment, as illustrated in Figure 1. For each question, participants could either (1) answer the question immediately, (2) navigate back to the location where the answer could be discerned, or (3) pay for an assistant to bring them back to that location.

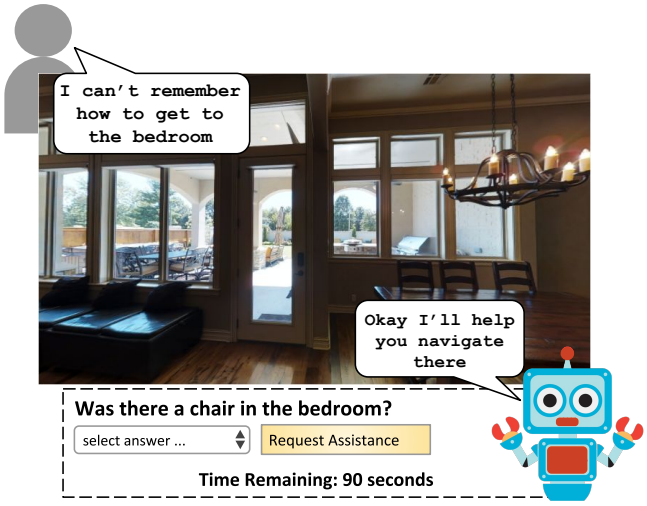


Figure 1: Conceptual depiction of the study with 3D environment, option for requesting assistance and human-assistant interaction.

In this paper, we present summary statistics and results of models (which employ hand-selected features) that predict whether participants will ask for assistance. The features used in these models were selected based primarily on intuition. Ultimately, we aim to formulate models of human perceptual and memory systems that are based on established computational models of human perception and cognition (e.g., [8, 20]). We hypothesize that this knowledge will help address a core challenge of developing accurate priors for when people will ask for assistance in memory and navigation tasks. Such priors provide a foundation for more generalizable models and inferring model parameters from less data (i.e., low-shot learning).

In summary, our main contributions are as follows:

(1) We introduce the Memory Question Answering (MemQA) task for humans, which tests human visual spatial memory. We created the Visually Grounded Memory Assistant Dataset, which contains over 6k instances of humans preforming the MemQA task. To the best of our knowledge, this is the largest dataset on visually-grounded memory assistance for humans.
(2) We perform in-depth analysis of conditions under which humans ask for assistance.
(3) We develop baseline models for the task of predicting whether participants will ask for assistance or navigate on their own as well as their accuracy on answering the MemQA questions.

2 RELATED WORK

2.1 Research and models of human memory

Human memory is often classified based on the length of storage (sensory, short, and long term) and the ability to communicate the contents of memory. While contents of *declarative* memory can be stated explicitly, *procedural* or perceptual-motor skills cannot be so stated [3]. *Declarative* memory is often further classified into *episodic* (life events) and *semantic* (language and symbol-based knowledge). In addition to this classification scheme, computational theory has led to the development of process models for the encoding, storage, and retrieval of memories. These models, typically built on the basis of association networks, blur the lines in the typology and capture important patterns in data [16, 17]. In particular, the encoding and recall of memories is highly dependent on spatio-temporal context and the memory networks built with this structure seem to associate representations across the memory typology described above. For example, people can often report the context in which they learned to tie their shoe (procedural and episodic memory) and shoe brands are likely faster to recall when a person is tying their shoe than when they zipping up their jacket (indicating context specific effects of procedural and semantic memory).

The complexity of these association networks has, historically, been difficult to study in detail due to methodological limitations. In particular, studying memory at scale and in real-world, visually-rich contexts is a significant challenge. By collecting data on visual memory tasks at scale on Amazon Mechanical Turk (AMT) in complex 3D environments, our data set provide an important step toward overcoming these limitations. AMT studies have also been used in the past to study memorability of images [15, 19]. However, these studies focus on purely visual features underlying memorability and do not account for context-driven or task-driven visual memory encoding.

The most relevant cognitive-science research for our task focuses on how people learn to navigate environments. People build and store mental maps that allow for more efficient navigation on future visits to that location. These maps are built from mixtures of sensory cues that include landmarks, optical flow as well as non-visual cues [13, 14, 32]. What features are used and how they are encoded is not fully understood [6, 10, 18]. Our data provide a rich source of information to develop our understanding of what visual features are encoded and stored in human spatial maps. Understanding what features are used by humans may enable the development of better navigation systems in mechanical autonomous agents. Such agents are now being developed in a research program on embodied question answering (EQA) [11, 30, 31] which was the primary machine-learning inspiration for the present study.

2.2 Embodied Perception and Question Answering

Recently, the computer-perception community has opened a new field of embodied perception where agents learn to perform tasks in 3D simulated environments in an end-to-end manner from raw pixel data. These tasks include target-driven navigation [33], instruction-based visual navigation [2], and embodied and interactive question answering (EQA) [11, 30, 31]. In a typical EQA task set up, an agent is spawned in a random location of a novel building and asked a question about an object or room such as "What color is the car?". The agent has no prior knowledge or representation of the building or objects and it must navigate to find the object and then answer the question correctly. This task involves learning a robust navigational system and an accurate visual inference to answer the question. This task was designed as a good measure of an agent's ability to preform visually grounded navigation and semantic understanding of the environment. We hypothesize that observing humans in the EQA task will lend insight into human spatial memory and semantic understanding of the environment. To this end, we expanded the EQA dataset to include five types of questions: location, existence, color, count, and comparison. We then use the new EQA questions to create a new task for humans called Memory Question Answering (MemQA). In the MemQA task, participants are given a short video of a fly through of an environment and are then asked to solve multiple EQA questions about that environment within a time constraint.

Apart from enabling the study of how humans perform navigation tasks and encode spatio-temporal information, our task also allows humans to ask for assistance when needed. Recent research in embodied and autonomous agents has also explored the utility of seeking assistance [22, 23]. Nguyen and Daumé III developed a navigation task where agents could ask for natural language assistance [22]. Their goal was to develop mobile agents that can leverage help from humans potentially to accomplish more complex tasks than the agents could entirely on their own. They also explored using a cost to for each request for assistance. Their goal was to try and learn the optimal policy for requesting assistance with a limited budget of requests. They did not gather human assistance dialogue but instead supplemented the assistance with the instructions from the Room2Room task [2]. Unlike this work, our focus is on understanding when humans might seek assistance in tasks that require visual memory encoding. The ability to understand when a user has forgotten something about the local environment and thereby might need assistance would be crucial for the next generation of contextual, personal assistants.

2.3 Automated interaction and AI Assistance

Research in human–robot interaction and simulations of human–AI cooperative systems has helped identify and make progress toward some of the major challenges in automated interaction systems. As

noted above, our data provide a foundation for models that better predict what information people encode and remember during navigation. Under the assumption that humans use informative features to learn spatial maps, identifying and using these features to train autonomous AI agents is likely to lead to more efficient learning than learning from raw pixels. Some research toward this goal has used simulated human agents to answer questions from an autonomous agent to help it learn to navigate [22, 23]. A limitation of using simulated human agents is that AI–AI learning does not necessarily translate into improved human–AI performance [9], a fact that has been starkly revealed in the development of autonomous vehicles [27, 28]. This brings us to the second applications of research from our data set: The development of effective AI assistance using first-person camera data.

By understanding perceptual features that humans use to perform a task (i.e., a 'theory of mind' for human perception and memory in our case), an automated assistant will be able to better identify when and how to intervene with assistance. Recent studies where the AI agent has third-party observation have demonstrated that this 'theory of mind' approach with the assumption that humans will act rationally to achieve a goal, has advantages for the development of human–AI collaboration [4, 21]. When the agent has first-person video, this approach lays the foundation for understanding what features of the environment and what task a person is likely to attend to (see [5, 25] for recent reviews of computer-vision approaches to action understanding from first-person video). A machine-based representation space that is better aligned with human perceptual and memory representations allows for better grounded interaction and communication between the human and AI agents. We hypothesize that identifying this common representation space will lead to better generalization of models for visually grounded assistance. We further envision that our data would provide a foundation for testing current models of human perception and memory from cognitive science. If such models are predictive, they can and should be incorporated into a 'human-like' representation for visually-grounded AI assistants.

3 THE VISUAL ASSISTANCE DATASET

We now describe the Memory Question Answering task and the data set collection protocol for the Visually Grounded Assistant (VGA) Dataset. To re-iterate, the goals of our dataset collection were (1) to provide a basis for the development of an agent that can predict when a person is likely to request assistance and (2) to develop and test models of human visual and spatial memory.

3.1 Task Description

Task for Human We refer to the task for the human participants as Memory Question Answering with Navigation (MemQA). Please refer the video* to see a demonstration of the MemQA task, where the assistant is an oracle that can be used to help answer questions about the environment. The task was encapsulated in the 3D simulated indoor environments from the Matterport3D dataset [7] and was run on Amazon Mechanical Turk. The Matterport3D dataset has been used in many embodied perception tasks [2, 24, 30] and thus our data from humans complements existing data from AI agents. Additionally we used the Matterport3D dataset because it contains scans of real buildings with realistic settings which range from offices to houses. The setup of our data collection interface is a simulation which uses the actual RBG panoramic frames of the environment. This removes any negative impact that reconstruction errors or unrealistic looking scenes could have on a participants memory.

In each trial, the participant is exposed to a fly through of an indoor 3D simulated building as depicted in Figure 2. This fly through is on average a little longer than 70 seconds. After the fly through, the participant is teleported back to the starting point and is presented with four questions. The questions are always about objects and rooms which were passed during the fly through. The questions can concern the location, color, count, and existence of objects, as well as room comparisons and object comparisons (see Section 3.2 and Table 1 for more examples and details about the questions). Participants are allotted a total of 2.5 minutes to answer all four questions. They can take three distinct approaches to answer a question: (1) answer immediately without any form of navigation, (2) navigate through the environment themselves (presumably to return to the location where the answer could be discerned), or (3) request assistance, in which case an assistant would transport them to the location from which the answer could be discerned. Selecting the third option carried a time cost of 10(+) seconds, which was subtracted from the time limit. The exact cost of each request for assistance was a function of total distance the assistant would have to navigate them. When the participant requested assistance they would specify which question they wanted assistance with and if it was a multi location question such as object comparison they could select which object they wanted to be navigated to. For example for the question "were the stove and the bathtub the same color?" the participant would have to choose whether they wanted to be assisted with finding the bathtub or the stove. In order to incentivise the importance of answering correctly, the participant is given a monetary bonus for each correct answer.

Task for Assistant The main goal of the assistant is to work collaboratively with the human to answer the questions. The agent needs to be able to understand the humans behavior so that it can realize when the human needs and wants help while not being overly intrusive. To do this the agent needs to model an an accurate representation of what a human has encoded during the fly through and to predict the humans behavior during the question answering phase. The assistant agent is set to act as an oracle for the building and is allowed access to the annotated 3D mesh as well as the questions the human is answering. This dataset creates multiple interesting tasks for an assistant agent. The tasks we propose are most useful to realization of an AR assistant are:

(1) Take in the fly-through and a single question. Predict: correctness, navigation behavior, assistance request behavior.
(2) Take in the fly-through, all four questions and the sequence of frames during the answering phase. At each time step of the answering phase predict behavior: navigation, request for assistance, answer selection or nothing

*https://www.youtube.com/watch?v=T97r2leqFyQ

In Section 5 we define a method for the first task and give some insight of methodologies and metrics for solving and evaluating the second task.

3.2 Dataset Collection

Simulators

The Matterport3D dataset [7] includes over 10k panoramic RGB-D images over 90 real indoor buildings. These panoramic nodes are distributed on average 2.25 meters apart across the entire building. The Matterport simulator [2] creates an interactive and navigable environment for the Matterport3D dataset. We used the Habitat simulator [29] to extract additional information, such as semantic segmentation of the mesh.

Question and Fly through Creation

Following the methodology from [11], we generated the questions programmatically using the Matterport3D meshes and annotations. Each question was represented as a functional template as shown in Table 1. Each template defines the query-able rooms or objects. The original EQA dataset on Matterport3D contained the question types indicated by the $*$ superscript. To obtain a more diverse and representative question set, we added existence, count, and comparison questions. These questions provide additional data to determine how difficult different features of the environment are for people to encode and remember. In order to ensure consistency of answers, participants selected the answer from a drop down menu of all possible answers. The random chance accuracy for all questions in the dataset is 29.08%.

After the initial question generation, we discovered many errors in the Matterport3D annotations. To eliminate erroneous annotations, we ran a crowd-sourced study to verify annotation accuracy. Table 1, lists the number of original generated questions, the number that were filtered by the verification study and the remaining number of questions. This study resulted in the necessary filtering of 20% of the generated questions.

We generated the fly-through paths to ensure that they included the visual information needed to answer the questions with a minimum-distance criteria. The fly-through paths were generated after filtering and generating the questions. To generate each fly through, we first randomly sampled, without replacement, 4 questions about different objects from the same environment. We then computed the shortest path through the environment that visited each required location. A short random trajectory was added to the start and end of the path. For purposes of consistency, fly throughs under 45 seconds or over 75 seconds were discarded.

4 VISUAL ASSISTANCE DATASET ANALYSIS

General Statistics

We collected ≥ 5 annotations for 1250 unique MemQA tasks with a total of 6275 annotations. Note that each MemQA task consists of a single fly through and four questions as described in Section 3.1. Table 2 shows some basic statistics of the task and data collected.

While the average time taken on the task was 74.96 seconds, on 9.68% of annotations a participant reached the time limit. This almost always occurred on annotations where participants requested assistance. This shows that the time limit acts as an effective way of budgeting the total number of assistance requests thus disallowing participants from completely relying on the assistant.

The differences in question difficulty is coarsely exposed through Figure 4 that shows the proportion of requests for assistance and Figure 5 that shows accuracy by question type. The fewest number of requests came for room count and nonexistence questions. While this suggests that these questions were the easiest to answer, the latter fact likely reflects the fact that people recognized that they would not gain any information by being brought to the absence of an object. Comparison questions also had few requests. This could reflect either difficulty level *or* a lack of willingness to spend the time required to go to multiple locations to answer the question. This is where our dataset provides opportunity to examine trade-offs between the time/effort it would take to find out the right answer and risking a guess to save that time. The most assistance was requested for room in color and object count suggesting these were difficult questions. Similar trends are observed in the accuracy data, which we examine next.

The left panel of Figure 5 shows accuracy of all answers (with or without assistance) and the right panel shows accuracy only for questions when the participant neither asked for assistance nor navigated to the location to obtain the answer. The right panel reveals clearly that room count and object existence questions were the easiest for people to remember. Interestingly, there is a substantial difference between their ability to answer existence and non-existence questions. This may reflect participants' recognition that there is a good chance that they missed something that was present in the environment, making it more likely that they will guess it exists when, in fact, it doesn't exist.

Finally, we wanted to see if the distance to the target location effected choices to navigate or get assistance. Figure 6 shows the distribution of distances to all targets (1) from the starting point, (2) for questions that were answered without navigation or assistance, (3) for cases where participants navigated to the target and (4) for cases where assistance was requested. While there are long tails in each of these distributions, the cases where people navigated to the target is the only distribution with a clear mode at the shortest distance. This suggests that participants were reluctant to navigate to targets when travel distance was long. To further examine the effect of distance, we included 'time of first exposure', which is correlated with distance, as a feature in the models we present in the next section.

5 BASELINES AND METHODS

We explore modeling the first task described in Section 3.1: given the fly through and question, predict whether a participant will be able to answer the question correctly, and whether they will answer the question (1) without any form of navigation, (2) by navigating on their own, or (3) by requesting assistance. We adopt a modeling approach that consists of constructing four binary classification problems: one for answer correctness, and one corresponding to each of the three answer-strategy approaches mentioned above. While we could construct a single 3-class classifier to predict the participant's strategy, a collection of binary classifiers enables more nuanced study of which outcomes are easiest to predict; for example,

Figure 2: Frames from a fly through and visual question from the MemQA task. These panoramic RGB images of a Matterport3D building were rendered using the Matterport Simulator.

Table 1: MemQA question types and templates. The *superscript denotes question types that were included in the original Matterport EQA [30].

Question Type	# Generated	# After Filtering	Template
*location	203	116	*What room is the <OBJ> located in?*
*color	299	188	*What color is the <OBJ>?*
*color inroom	1432	943	*What color is the <OBJ> in the <ROOM>?*
existence	283	207	*Is there a <OBJ> in the <ROOM>?*
count object	2729	2463	*How many <OBJS> in the <ROOM>?*
count room	340	299	*How many <ROOMS> in the house?*
color compare inroom	320	288	*Does <OBJ1> share same color as <OBJ2> in <ROOM>?*
color compare xroom	86	70	*Does <OBJ1> in <ROOM1> share same color as <OBJ2> in <ROOM2>?*
object size inroom	300	272	*Is <OBJ1> bigger/smaller than <OBJ2> in <ROOM>?*
object size xroom	260	212	*Is <OBJ1> in <ROOM1> bigger/smaller than <OBJ2> in <ROOM2>?*
room size compare	1048	932	*Is <ROOM1> bigger/smaller than <ROOM2> in the house?*

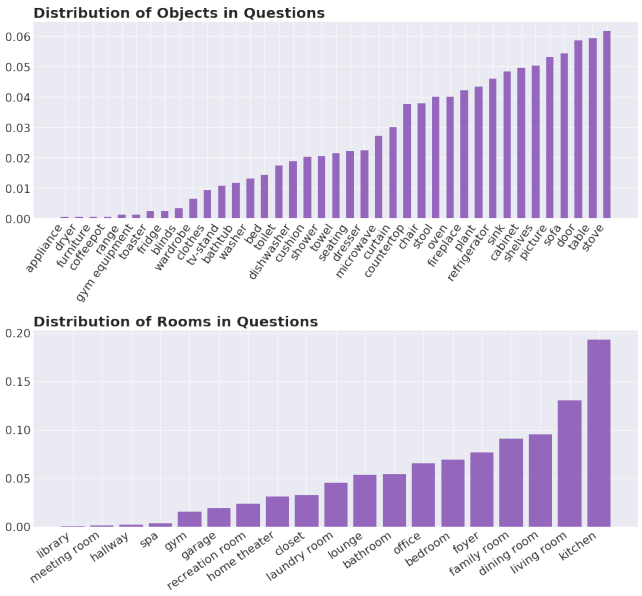


Figure 3: Distribution of the question topics: Each question type is generated from a template and query-able objects and rooms. The graphs show the distribution of query-able objects and rooms we used to generate questions. These distributions of questions were obtained after filtering out the questions that included erroneous annotations in the Matterport3D data set.

Table 2: VGA Dataset: General Statistics

Avg. length of fly through (seconds)	71.45s
Avg. time taken on task (seconds)	74.96s
Percent of annotations that reached the time limit	9.68%
Num. of unique workers	412
Avg. participant accuracy	70.34%
Random Chance accuracy	29.08%
Percent of annotations that used assistance	58.34%
Percent of questions with assistance requests	25.23%

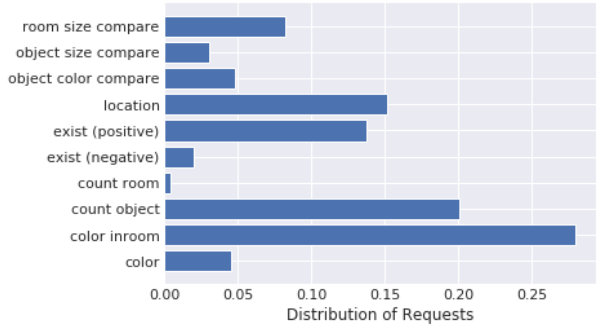


Figure 4: Distribution of Assistance Requests by Question Type: The frequency of requests for assistance is a good indicator of question difficulty.

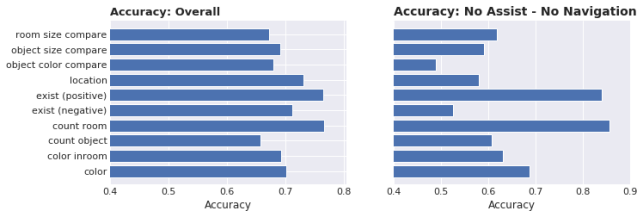


Figure 5: Accuracy by Question Type: The left panel shows the accuracy for each question type for the entire data set. The right panel shows accuracy for questions that were answered without assistance or navigation.

we can generate a receiver operating characteristic (ROC) curve for the classifier associated with the prediction of each outcome.

The goal of this study is to model human visual and spatial memory. The participant is always exposed to the correct answer of a question during the fly through. Whether the participant is able to answer the question after completing the fly through is a direct measure of their ability to encode and recall the relevant feature of the environment. Additionally, the participant's ability to navigate back to the location of the answer is informed by their spatial memory. Visually encoding all of the information into memory is outside the bounds of human memory ability. What features actually get encoded will depend on many factors such as object saliency and the duration of the fly through. By modeling the participants' performance based on the fly through alone, we seek to determine the most important factors in determining performance. For the initial model, we did not consider question types that dealt with multiple objects. This is, we omitted comparison and count questions, resulting in five remaining question types.

Let us call primary object of the question o_1. We started by characterizing each question with the following (hand-selected) features:

(1) Type of question (discrete).
(2) Length of fly through (continuous).
(3) Time of first exposure to o_1 (continuous).
(4) The temporal exposure of o_1.
(5) The spatial exposure of o_1 (continuous).
(6) Word embedding of object type of o_1 (continuous).

Features 3, 4, and 5 were drawn from the Habitat simulator over the frames of the fly. Using the instance and semantic segmentation over each frame, the object information from each view was extracted (Figure 7). To obtain a measure of exposure of objects, it was necessary to consider exposure both in temporal and spatial terms. Temporal exposure refers to the length of time the object was in view and spatial exposure refers to the amount of area the object occupied throughout the fly through. The spatial exposure is defined as the total of number of pixels of o_1 across all frames of the fly through.

5.1 Model Description

As mentioned above, we used these features to train four binary classifiers. We now describe the features and classification models we consider:

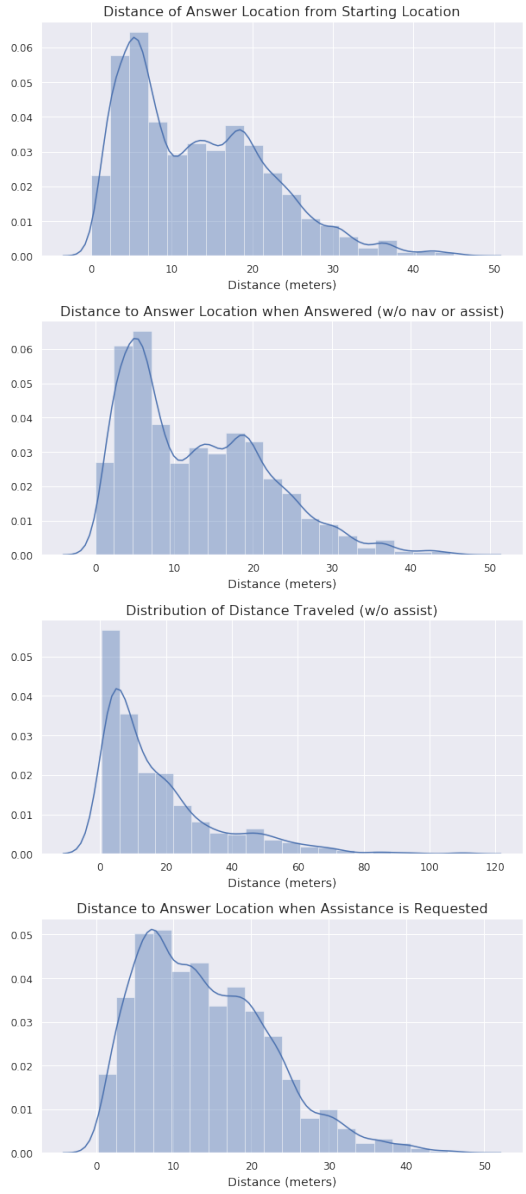
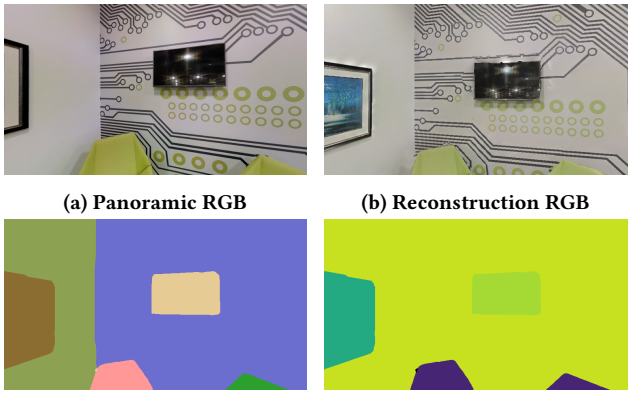


Figure 6: Distribution of Distance to Target under different condition: The distance in meters from the participant to the location answer. Illustrates the location conditions under which participants ask for assistance.

Features. We employ 10 features in the models: the features 1–5 above, as well as a 5-dimensional embedding of the object type derived from feature 6 above; note that the dimension of the embedding space is a hyperparameter, which we set to five to keep the number of features relatively small. To compute this 5-dimensional embedding, we (1) compute the GloVe embedding [26] of all 35 unique object types into a 50-dimensional latent space, (2) apply principal component analysis (PCA) [1] and project the embeddings onto the 5-dimensional linear subspace of this 50-dimensional latent



(a) Panoramic RGB (b) Reconstruction RGB

(c) Instance Segmentation (d) Semantic Segmentation

Figure 7: Example of four frames types for the same camera view point. (a) is an example of the view that participants had in the MemQA task. (b) shows the Habitat Simulator reconstruction from the environment mesh. The mesh and it's annotations were used to construct the MemQA questions as well as to create the features for modeling accuracy and assistance requests. The (c) and (d) show segmentation as annotated by the environment mesh.

space spanned by the first five principal components, and (3) treat the resulting five PCA coordinates of the object type as features.

Classification models. We consider three different classification models as implemented in Scikit-learn:

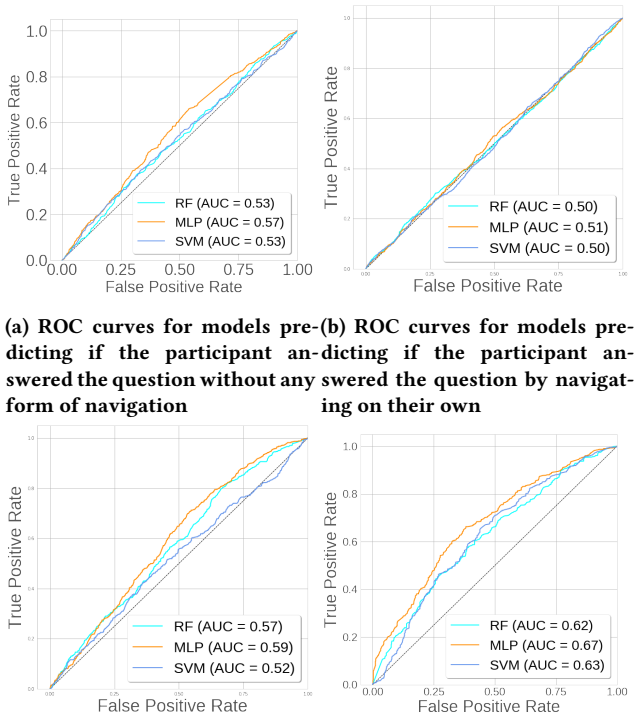
(1) **Random-forest (RF) classifier**. We employ an ensemble of 10 trees and we use the Gini impurity measure in tree construction; all remaining parameters correspond to the default values in Scikit-learn.
(2) **Multilayer-perceptron (MLP) classifier** (i.e., feedforward, fully connected neural network). We employ one hidden layer with 100 neurons, ReLU activations, and the Adam optimizer; all remaining parameters correspond to the default values in Scikit-learn.
(3) **Support-vector-machine (SVM) classifier**. We employ a radial-basis-function kernel with coefficient $\gamma = 2$ and penalty parameter of $C = 1$; all remaining parameters correspond to the default values in Scikit-learn.

Models are trained using 80% of the data, and are tested on the remaining 20%. We ensure that the training and testing sets contain different questions, such that results on the test set assess generalization of the models across both participants and questions.

5.2 Results

Figure 8 reports the receiver operating characteristic (ROC) curves for each of the candidate classification models on each of the four binary prediction tasks computed on the test set, with Table 3 reporting the associated area under the curve (AUC).

These results demonstrate that—using the ten features specified above—the models are best able to discriminate whether the participant correctly answered the question; the MLP classifier achieved an AUC–ROC value of 0.67 in this case. The models' next best



(a) ROC curves for models predicting if the participant answered the question without any form of navigation

(b) ROC curves for models predicting if the participant answered the question by navigating on their own

(c) ROC curves for models predicting if the participant answered the question by requesting assistance

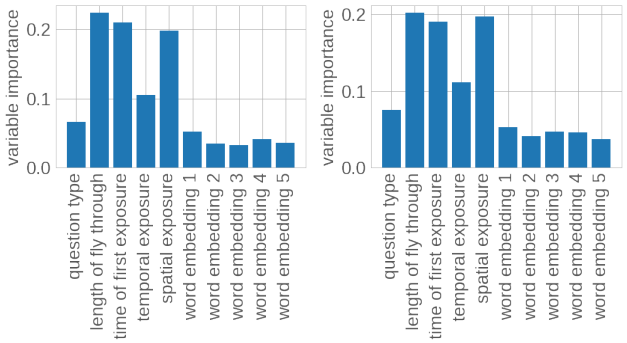
(d) ROC curves for models predicting if the participant answered the question correctly

Figure 8: ROC curves for the four types of predictions and three considered classification models

Response	RF	MLP	SVM
no navigation	0.53	0.57	0.53
self navigation	0.50	0.51	0.50
assistance	0.57	0.59	0.52
answered correctly	0.62	0.67	0.63

Table 3: AUC–ROC values for the candidate classifiers and prediction tasks (RF: random forest; MLP: multilayer perceptron; SVM: support vector machine)

performance occurs for predicting either the participants' ability to answer with no navigation or with assistance; the MLP classifier achieved AUC–ROC values of 0.57 and 0.59, respectively, in these cases. None of the models performed better than chance in predicting whether the participant employed self-guided navigation to answer the question. Note that currently, none of our features capture the spatial mapping of the environment or the complexity of the fly through path, both of which might be important for predicting whether the users would employ self-navigation. Apart from developing more sophisticated models, we also plan to explore human perception and cognition inspired features in the future. For instance, object saliency has been found to be important for recalling the object [12]. Likewise, landmarks, optical flow, and



(a) Variable importance for "assistance"

(b) Variable importance for "answered correctly"

Figure 9: Variable importance generated by the random-forest classifier for two responses.

other non-visual cues have been found to be important for navigation [13, 14, 32].

We now turn to the question of which features are most important for determining the above outcomes. For this purpose, Figure 9 reports the feature importances arising from the random-forest classifier for responses corresponding to "assistance" and "answered correctly"; we note that the importances for "no navigation" and "self navigation" are nearly identical to the former. These figures demonstrate that the most important features driving the construction of the decision trees correspond to (1) the total number of time steps in the fly through, (2) the time step at which the object is exposed, and (3) the spatial exposure of the object. Interestingly, the semantic meaning of the objects—as represented by the word embeddings—are characterized by low variable importance; the question type also is relatively unimportant.

6 DISCUSSION

The primary purpose of this paper was to introduce a new data set. The data provide an opportunity to develop models of human visually-grounded memory that can serve as a basis for an automated memory assistant. The assistant in our task was an oracle that only responds when called for. However, the simple model we presented above demonstrates promise of developing a model from a richer feature set that can predict when assistance is needed without an explicit request. In addition to the opportunity to develop machine representations that align with human perceptual and memory systems, our data offer an opportunity to examine how people trade off the cost of time to obtain the right answer with the risk of getting the wrong answer (and, hence, not receiving the monetary reward). This could provide a rich avenue for gaining insight into how people value the travel time associated with obtaining the right answer, money, and risk (i.e. the probability of being wrong when guessing an answer without confirmation). Similar situations arise, for example, when people travel to different shops for items at lower prices than the store they are in.

7 CONCLUSION

The primary aim of this paper was to introduce the 'The Visually Grounded Memory Assistant Dataset'. The summary statistics and baseline model demonstrate the potential of using these data to develop models that can predict when people will ask for visual and spatial memory assistance.

Our dataset creates a rich set of tasks to explore including predicting human performance, behavior and memory. We explored modeling predicting human performance on the MemQA task from what participants were visually exposed to in the fly through as well as the context of the task they were solving. Future work can explore looking at their behavior during the question answering phase and model predictions from raw pixels.

REFERENCES

[1] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

[3] Alan David Baddeley. 1999. *Essentials of human memory*.

[4] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Publishing Group* 1, 4 (March 2017), 1–10.

[5] Alejandro Betancourt, Pietro Morerio, Carlo S Regazzoni, and Matthias Rauterberg. 2015. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 5 (May 2015), 744–760.

[6] E Chan, O Baumann, Mark Bellgrove, and Jason Mattingley. 2012. From objects to landmarks: the function of visual location information in spatial navigation. *Frontiers in Psychology* (Aug. 2012).

[7] 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. In *arXiv preprint arXiv:1709.06158*.

[8] Nick Chater, Joshua B Tenenbaum, and Alan Yuille. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences* 10, 7 (July 2006), 287–291.

[9] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating Visual Conversational Agents via Cooperative Human-AI Games. In *Fifth Conference on Human Computation and Crowdsourcing*.

[10] Matthew Collett, Lars Chittka, and Thomas S Collett. 2013. Spatial Memory in Insect Navigation Review. *Current Biology* 23, 17 (Sept. 2013), R789–R800.

[11] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *CVPR*.

[12] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. 2008. Objects predict fixations better than early saliency. *Journal of Vision* 8, 14 (2008), 18–18.

[13] Ariane S Etienne and Kathryn J Jeffery. 2004. Path integration in mammals. *Hippocampus* 14, 2 (2004), 180–192.

[14] Sabine Gillner and Hanspeter Mallot. 1998. Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of Cognitive Neuroscience* 10, 4 (July 1998), 445–463.

[15] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. 2019. GANalyze: Toward Visual Definitions of Cognitive Image Properties. *arXiv preprint arXiv:1906.10112* (2019).

[16] Marc W Howard. 2017. Temporal and spatial context in the mind and brain. *Current Opinion in Behavioral Sciences* 17 (Oct. 2017), 14–19.

[17] M W Howard, C J MacDonald, Z Tiganj, K H Shankar, Q Du, M E Hasselmo, and H Eichenbaum. 2014. A Unified Mathematical Framework for Coding Time, Space, and Sequences in the Hippocampal Region. *Journal of Neuroscience* 34, 13 (March 2014), 4692–4707.

[18] Simon Jetzschke, Marc O Ernst, Julia Froehlich, and Norbert Boeddeker. 2017. Finding Home: Landmark Ambiguity in Human Navigation. *Frontiers in Behavioral Neuroscience* 11 (July 2017), 12301–15.

[19] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*. 2390–2398.

[20] David C. Knill and Whitman Richards (Eds.). 1996. *Perception As Bayesian Inference*. Cambridge University Press, New York, NY, USA.

[21] Chang Liu, J Karl Hedrick, S Shankar Sastry, Thomas L Griffiths, Jessica B Hamrick, Jaime F fisac, Anca D Dragan, and J Karl Hedrick. 2016. Goal Inference

Improves Objective and Perceived Performance in Human-Robot Collaboration. In *Proceedings of the th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 940–948.

[22] Khanh Nguyen and Hal Daumé III. 2019. Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning. In *International Joint Conference on Natural Language Processing*. 684–695.

[23] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. 2019. Vision-Based Navigation With Language-Based Assistance via Imitation Learning With Indirect Intervention. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 12527–12537.

[24] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. 2019. Vision-Based Navigation With Language-Based Assistance via Imitation Learning With Indirect Intervention. In *CVPR*.

[25] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, and Francisco Florez-Revuelta. 2016. Recognition of Activities of Daily Living with Egocentric Vision: A Review. *Sensors* 16, 1 (Jan. 2016), 72–24.

[26] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[27] Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. 2017. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems XIII*. Robotics: Science and Systems Foundation.

[28] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. 2016. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*. University of California, Berkeley, Berkeley, United States.

[29] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. *arXiv preprint arXiv:1904.01201* (2019).

[30] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. 2019. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *CVPR*.

[31] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, and Dhruv Batra. 2019. Multi-Target Embodied Question Answering. In *CVPR*.

[32] M Zhao and William H Warren. [n. d.]. How you get there from here: Interaction of visual landmarks and path integration in human navigation. *Psychological Science* 26, 6 ([n. d.]), 915–924.

[33] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*.