2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Answerability Fields: Answerable Location Estimation via Diffusion Models

Daichi Azuma[1], Taiki Miyanishi[2,3,5], Shuhei Kurita[4,5], Koya Sakamoto[6,3] and Motoaki Kawanabe[3,5]

***Abstract*— We propose Answerability Fields (AnsFields), a novel approach for predicting the answerability of questions at different locations within indoor environments. AnsFields is represented as a map, where each grid's score reflects how well a question can be answered using the panoramic image at that location. Using a 3D question-answering dataset, we construct comprehensive AnsFields covering diverse scenes from ScanNet. Additionally, we employ a diffusion model to infer AnsFields from a scene's top-down view image and the question. We then conduct 3D question-answering using these predicted AnsFields and achieve a 24% improvement in accuracy over the standard 3D-QA method. Our results demonstrate the importance of object locations for answering questions in the environment, highlighting the potential of AnsFields for applications in robotics, augmented reality, and human-robot interaction.**

I. Introduction

The rapid advancements in applying deep neural networks to embodied agents have enabled capabilities such as navigating indoor environments following linguistic instructions [2], [4], [6], dexterous object manipulation [1], [18], [34], and answering questions within 3D environments [22], [33]. In particular, the capacity of embodied agents to interpret and respond to queries within 3D environments is essential for developing robots that can comprehend human language and execute tasks accordingly. In Embodied Question Answering (EQA) task [9], agents randomly placed in an unknown environment must explore it to find specific objects for providing accurate answers to posed questions. However, even though the layouts of indoor environments are often known in real-world scenarios and robots typically possess their own maps, most existing methods are unable to leverage this valuable map information. Thus, the question naturally arises, "Can we use an indoor 2D map to answer the question about 3D space?" There are several models to predict the location of objects in 3D space from 2D map which agent generated [11], [24]. However, in the QA task, the agent needs to find a location where agent can visually acquire not only the target object location in 3D space, but also the necessary information to answer the question, i.e., these are not only the target objects, but also other objects and object-position relationships included in the question. To this end, we have to focus on question answering using layouts and further investigate what locations the agent would need to spawn for QA.

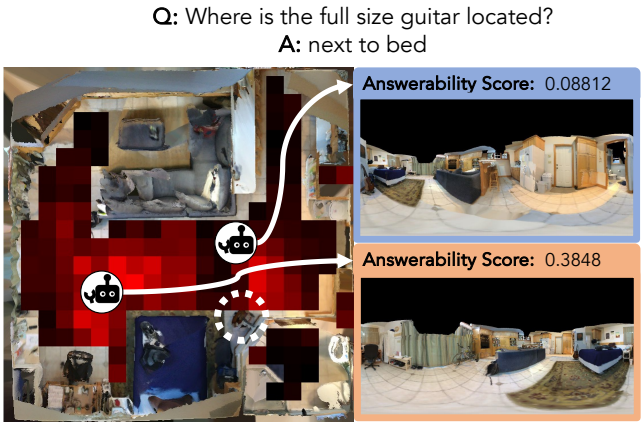

Fig. 1. We propose Answerability Fields to make agent efficiently understanding scenes. We compute the score of answerability to the questions in each location by using strong Visual Question Answering (VQA) model called OFA.

In this work, we propose Answerability Fields (AnsFields) that facilitates robots answering questions in 3D environments using indoor 2D maps. AnsFields is a grid representation where each cell indicates the possibility that the robot will be able to answer a given question at that specific location. We calculated the probability that the vision and language foundation model would return an appropriate answer based on the question and the agent-perspective image captured from each grid location when the agent was spawned there. We create AnsFields using the ScanQA [3] dataset, a 3D Question Answering dataset in indoor scenes from richly annotated 3D scans of ScanNet [8].

Figure 1 illustrates an example of AnsFields for the question "Where is the full size guitar located?", in which locations with the lighter red indicate higher answerability scores and the locations with darker red indicate lower answerability scores. In this example, in addition to "guitar" which is the subject of the answer, an image containing context that shows where the guitar is located is required. The answerability score is higher for locations that can be seen in the relationship between the "bed" and "guitar". We confirmed that the location that has the highest answerability score improves the QA performance compared to answering in front of the target object of the question, which is a primary objective of previous works [9], [33]. In contrast to previous EQA works [22], [33], once the location of the high answerability is known based on AnsFields, the question can be answered with high accuracy by moving to that location

[1] Sony Semiconductor Solutions, Tokyo, Japan
[2] The University of Tokyo, Japan
[3] Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan
[4] National Institue of Informatics, Tokyo, Japan
[5] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[6] Graduate School of Informatics, Kyoto University, Japan

and conducting visual question answering (VQA) based on the surrounding images there.

Therefore, we present a diffusion model tailored to predict AnsFields based on top-down view images of the environment and questions. It employs a technique known as InstructPix2Pix [5] designed for image-to-image translation tasks, guided by textual instructions. We train a diffusion model with the question and paired images that top-down view images and ground truth images of AnsFields. We also evaluate the VQA accuracy which is performed from the agent-perspective image taken on the highest-scoring location of the AnsFields predicted by the diffusion model. In the results, our method predicts the AnsFields that enhance QA performance, which outperforms the 3D QA method [3] and the 2D QA existing methods.

II. Related Work

Question Answering in 3D Space. EQA serves as the convergence point of visual navigation and visual question answering (VQA) within intricate 3D environments [9], [33]. In EQA, agents are tasked with navigating their surroundings to identify objects and provide accurate responses to posed questions. This task necessitates not only perceptual capabilities but also spatial reasoning and contextual understanding, underscoring the intricacy of embodied interactions in rich environments. In previous EQA methods [9], [33], the agent's behavior is trained by the trajectories to the vicinity of the target. Through the use of 3D point cloud data from ScanNet [8], a richly-annotated indoor 3D scans dataset, ScanQA [3] enhances the system's ability to comprehend spatial relationships, improving the identification and localization of objects within the scene. This capability enables more accurate question answering, especially for queries that involve spatial reasoning or require an understanding of the scene's layout.

Generative Models for Map Generation. In recent years, diffusion models have significantly improved in accuracy and can generate appropriate images under various conditions. Combined with advancements in Large Language Models (LLMs), it is now possible to include text, audio, and images as conditions and generate outputs. Noteworthy examples include InstructPix2Pix, Codi, and Controlnet [5], [30], [31], [35], [36]. These systems not only generate conditioned data but also comprehend the data and solve various tasks. For instance, DiffusionInst [12] and DiffusionDet [7] utilize diffusion models to recognize objects in images, predict segmentation, and bounding boxes, outperforming existing methods Diffusion models have also found application in scene understanding and robotics. MapPrior [37] employs a map image as input and utilizes a diffusion model to classify classes within the map. Models predicting the trajectory of an agent, such as PolyGRAD [26] and DiPPeR [17], leverage classifier guidance based on state and action. Look Outside the Room [25] generates an image of the agent's path from an image of an indoor environment. Additionally, there are models that generate house layouts using GANs, enabling scene understanding with diffusion models [23].

Semantic Fields. Semantic perception in the realm of embodied artificial intelligence involves understanding and interpreting the environment in terms of meaningful concepts and relations. Several recent advancements contribute to this area. CLIP-Fields [29] introduces an implicit scene model capable of tasks such as segmentation, instance identification, semantic search over space, and view localization. By learning a mapping from spatial locations to semantic embedding vectors, CLIP-Fields can perform these tasks without direct human supervision, outperforming traditional methods like Mask-RCNN [13] in tasks such as a few-shot instance identification or semantic segmentation with only a fraction of the examples. Additionally, when used as a scene memory, CLIP-Fields enables robots to navigate semantically in real-world environments, demonstrating practical utility. PONI [24] addresses the challenge of ObjectGoal navigation by separating the skills of 'where to look?' and 'how to navigate to (x, y)?'. By treating 'where to look?' as a perception problem, PONI learns to predict potential functions from semantic maps, facilitating efficient navigation to unseen objects. This modular approach achieves state-of-the-art performance on Object-Goal-Navigation tasks while significantly reducing computational costs for training, making it promising for real-world deployment. VLMAP [15] proposes a spatial map representation integrating pre-trained visual-language features with 3D reconstructions of physical environments. This approach allows natural language indexing of maps without additional labeled data, enabling more precise navigation according to complex language instructions. VLMAP facilitates autonomous map generation from video feeds, enhancing robot navigation capabilities in both simulated and real-world environments. ConceptFusion [16] introduces a scene representation that is open-set and multimodal, enabling reasoning about concepts across various modalities such as natural language, images, and audio. Leveraging foundation models pre-trained on internet-scale data. ConceptFusion achieves effective zero-shot spatial reasoning and retains long-tailed concepts better than supervised approaches. Extensive evaluations demonstrate its efficacy across different real-world applications, offering new possibilities for blending foundation models with 3D multimodal mapping. These semantic maps collectively contribute to enhancing semantic perception in embodied artificial intelligence systems, enabling more robust and contextually aware interactions with complex environments.

III. Problem Fomulation

In this work, we address the task of map-based EQA in the 3D space, where a robot is placed within a previously seen 3D environment, moves to the best location for QA, and provides accurate responses to textual questions using the robot's egocentric RGB images. In contrast to the existing embodied QA tasks [9], [10], [33], the robot already has the 2D map (top-down image) of the placed environment that can

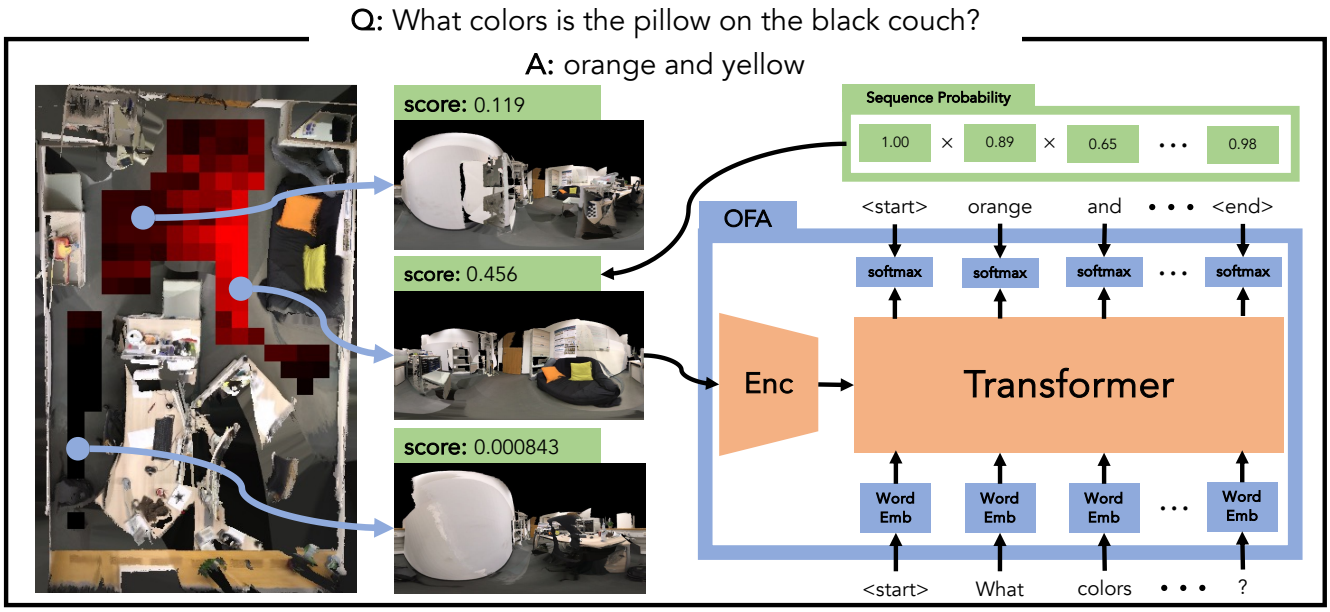

Fig. 2. Represents a method for generating answerability fields, which calculates the probability for each sequence to return an appropriate answer when a question and a image are entered into VQA model.

be used for this task. Thus, using a map to find a favorable location for answers is important for solving this task.

IV. Approach

In this section, we introduce the Answerability Fields (AnsFields) that represents the most efficient locations for answering questions in the 3D scenes. We hypothesized being too close or too far from the object related to the answer may not provide the best viewpoint; instead, it is crucial to view the object and its surrounding context to identify what the question is asking and determine where the relevant information is available to answer it. To explore the concept, we introduce the definition of AnsFields, then describe how to create AnsFields using vision-and-language foundation models, and then present the diffusion models for predicting AnsFields in unseen 3D environments.

A. Definition of Answerability Fields

Answerability Fields (AnsFields) is the grid representation of answerability for embodied agents. Answerability is calculated as the ability to provide appropriate answers to questions about panoramic images and scenes acquired at each location. As depicted in Figure 1, each cell within the grid is assigned answerability scores represented as the R value of RGB value, where a value closer to 255 indicates the higher probability of providing the correct answer to a question, while a value closer to 0 indicates a lower probability. We consider the probability of a visual question answering (VQA) at each cell as the answerability score.

B. Creation of Answerability Fields

To compute the AnsFields, we employed the powerful pre-trained transformer-based encoder-decoder framework, OFA [32], which was trained on a large corpus of image-text pair data, serving as the VQA model. It achieves remarkable performance in a series of cross-modal tasks such as visual grounding, VQA, and image captioning. Here, while any model can be used as the VQA model, in this work, we utilize OFA due to its high performance while being operable on a single GPU. Initially, we fine-tuned the OFA pre-trained model with the 3D-QA dataset, which takes a question and a panoramic image at each location on the grid of indoor scenes as input and predicts the token sequence of the answer. Then, OFA computes the answerability score, which represents how well a VQA model predicts answer tokens. This score is obtained as the probability of generating the correct answer tokens. We calculate scores for all grid cells within the range an agent can move (i.e., NavMesh area) to generate the AnsFields. Figure 2 depicts an example of how to compute the AnsFields within the range an agent can move. When addressing the question "What color is the pillow on the black couch?" with the answer "orange and yellow," the highest score is assigned to the location that captured a panoramic image showcasing both the orange and yellow pillows. Conversely, lower scores are given to locations that obtained images not clearly showing both pillows or not showing them at all. Using the VQA model, AnsFields can quantitatively express locations that are easier to answer.

C. Datasets

As a 3D-QA dataset, we utilized the indoor scenes from the ScanNet dataset [8] and obtained question-answer data related to these scenes from the ScanQA dataset [3]. The AnsFields' images are created according to the following steps. (i) We filled in missing parts of the floor in the point cloud of ScanNet data and then meshed the point cloud using Delaunay triangulation. (ii) We identified navigable areas corresponding to the floor and divided them into a grid based on the scene size. (iii) We then employed the VQA model to

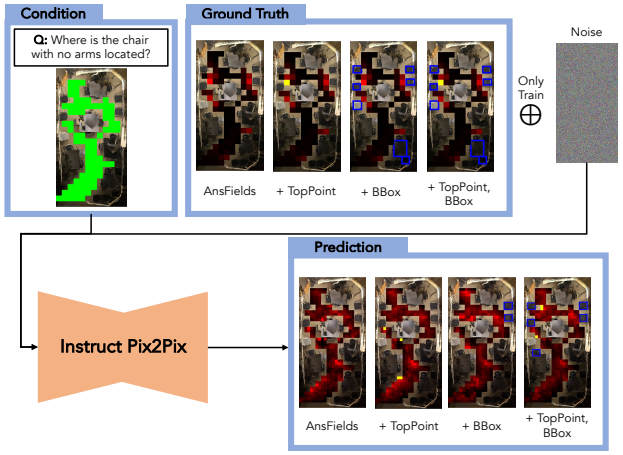

Fig. 3. This is the overview of generating Answerability Fields via Instruct Pix2Pix. On train, learning to predict noise by adding noise to the correct image. The correct images were trained and comparatively evaluated with only Answerability drawn and, TopPoints highlighted and BoundingBoxes on the top-down view images of the scene

compute answerability using panoramic images at the grid locations. (iv) We plotted the normalized answerability score on this top-down view image created from the scene's point cloud. We excluded scenes with navigable areas that were too small. Consequently, we created AnsFields images for 619 scenes and 31,741 question-answer pairs.

D. Generation of Answerability Fields by Diffusion Models

We generate the AnsFields for unknown questions using a diffusion model called InstructPix2Pix [5]. It is an extension of the latent diffusion model (LDM) [27], which samples data in latent space via a variational autoencoder [19] with UNet [28] architecture. Formally, given an RGB image C_i and text prompt C_p, the InstructPix2Pix model processes a noisy image Z_t as its input and produces a new image Z_0 using the Classifier-Free Guidance [14]. In the Classifier-Free Guidance, the diffusion model predicts the amount of noise present in the input image Z_t, using the denoising U-Net $\tilde{e}\theta$ which is divided into two networks, conditional noise predictions $e_\theta(Z_t, c)$ and unconditional noise predictions $e_\theta(Z_t, \phi, \phi)$ as:

$$\begin{aligned}\tilde{e}\theta(Zt, C_i, C_p) = e_\theta(Z_t, \phi, \phi) \\ + S_i \cdot (e_\theta(Z_t, C_i, \phi) - e_\theta(Z_t, \phi, \phi)) \\ + S_t \cdot (e_\theta(Z_t, C_i, C_p) - e_\theta(Z_t, C_i, \phi))\end{aligned}$$

where S_i and S_t are the scale factors in adding up the conditional scores of the image and text, respectively. A higher scale factors indicate a stronger influence of the condition on the generated image. In our experiments, the image condition scale factor S_i is set to 1.5 and the image and text condition scale factor is set to 7. Using InstructPix2Pix, our method learns to predict the AnsFields conditioned on textual questions and top-down view images of indoor scenes.

Highlight Top Answerability Score Point. In the AnsFields, answerability is normalized and plotted based on the size of the R value of the RGB value (0 to 255). Therefore, the RGB value of the point with the highest answerability is represented as $[255, 0, 0]$. To improve prediction, we change the color of this point to $[255, 255, 0]$. By explicitly learning the points where the most visual information to answer the question is gathered, we anticipated an increase in inference accuracy.

Visualize Object BBox. To infer answerability, understanding the location of the target objects mentioned in the question and the relationship between their locations is considered necessary. Hence, bounding boxes were applied with RGB values $[0, 0, 255]$ to the objects related to the question in the training image. By allowing the model to infer the location of the target object along with the answer possibilities, we aimed to enhance the understanding of the environment and measure the improvement in accuracy.

The overview of AnsFields predictions in InstructPixPix is shown in Fig 3. Random noise is added to the correct AnsFields image and, as a further condition, the question and scene top-down view image are added to predict the noise with UNet [28], and compute the Mean Square Error (MSE) with the prediction results and correct image to train the model. The correct AnsFields image is trained with each of the above conditions: only AnsFields are added to the top-down view of the scene, TopPoints are illustrated and the BoundingBox is annotated. During sampling, noise images, questions, and scene images are used to generate AnsFields images.

V. Experiments

A. Estimation of the Best Viewpoint

From the image inferred by the model, we calculated the point with the highest score of answerability to the question and obtained a panoramic image of the environment taken at that location. Specifically, we estimated the location of navigable areas in the generated images using a transformation matrix between the pre-prepared scene and the top-down view images. The point with the highest RGB value of R in that area was then identified. Furthermore, we calculated the position of that point in the scene using the transformation matrix, placed the agent, and acquired the panoramic image. In essence, this point represents the most probable position to answer the question inferred by the model, allowing us to validate whether the correct inference was made by evaluating the image of that point.

B. Implementation Details

We utilized a newly created collection of AnsFields images and question and answer pairs on the ScanQA [3] dataset to train the InstructPix2Pix model [5]. The model was trained using AdamW [21] with an initial learning rate of 1e-5 and a batch size of 16 for 150 epochs with train data. Subsequently, AnsFields were generated on the test data.

C. Evaluation

To verify the accuracy of the locations predicted by AnsField as the most likely to contain the answer to a question, a VQA test was conducted using a panoramic image taken at

TABLE I

VQA PERFORMANCE COMPARISON TO THE EXISTING METHOD.

MODEL	EM@1	EM@10
2D-QA and 3D-QA		
TopDownQA	35.00	71.32
ScanQA	31.18	68.35
Agent-perspective QA		
Random Spawn	37.75	73.17
Top-down View Attention	38.38	74.36
AnsFields (Ours)	**38.77**	**74.41**

the predicted location (referred to as Agent-perspective QA). We used EM@1 and EM@10 for the VQA metric, where EM@K represents the percentage of predictions where the top K predicted answers exactly match one of the ground-truth answers.

D. Baselines

We compared the AnsFields inferred by our diffusion model against various other methods. Note that no existing models are specifically tailored for the inference of Ans-Fields. Therefore, we selected a range of closely related methods for this comparison.

Top-Down View VQA. We utilized mesh images captured from a top-down view (referred to as TopDownQA) to capture the entire room with a single image for performing 2D-QA. Given a question and top-down view image, the method directly produces the answer. We use the same VQA model, OFA, as the proposed method to see whether an agent-perspective panoramic or top-down view image is more effective.

3D-QA. For comparison, we introduce a 3D-QA method that employs a point cloud of the entire scene to answer questions. Unlike 2D-QA which uses top-down images, 3D-QA can consider 3D spatial relationships of objects in a room. We use ScanQA [3], the most commonly used method for 3D-QA on indoor scenes.

VQA at Random Location. To investigate whether it is effective to VQA at appropriate positions, we will prepare a method to QA at random positions (referred to as Random Spawn.)

VQA on Top-down View Attention. This method uses the text-image matching model BLIP [20] to determine the most salient location within the Navmesh area in a top-down view image based on a given question using the attention mechanism. Subsequently, an agent is positioned at this identified location to capture a panoramic image. The image is used for conducting VQA using the OFA model [32].

VI. RESULTS

A. Comparison to Baselins

Table I shows the performance of the baselines and our method on our QA dataset. Our method significantly outperformed both the 2D-QA method using top-down view images and the 3D-QA method. The results suggest that panoramic images from the agent's viewpoint are more effective for QA of indoor spatial reasoning than point clouds or top-down view images of the entire scene. Additionally, our method achieving EM@1 and EM@10 38.77 % and 74.41 %, respectively outperformied "Random Spawn" and "Top-down View Attention" with 38.38 % for EM@1 and 74.36 % for EM@10. This indicates that in agent-perspective QA, conducting VQA at locations relevant to the question is more effective than performing VQA at random locations. In addition, the proposed method for generating AnsFields can predict locations more useful for QA than those predicted by image-text matching of a question and a top-down view image.

TABLE II

ABLATION STUDY OF THE IMAGES WHICH INPUT TO INSTRUNC-PIX2PIX AS CONDITIONS.

TopPoint	BBox	EM@1	EM@10
✓		37.85	73.78
	✓	36.73	73.51
✓	✓	**38.77**	**74.41**

B. Ablation Study

We investigated the extent to which predicting object bounding boxes (refer to BBox) and the locations with the highest answerability scores (refer to TopPoint) affect the performance of the final agent perspective QA. Table II shows the ablation results. The results show the VQA performance with both TopPoint and BBox had the highest accuracy, 38.77 % for EM@1 and 74.41 % for EM@10. This result can be attributed to the diffusion model's ability to learn and predict the locations of objects and areas with high answerability scores, enabling more accurate predictions of Answerability Fields.

C. Qualitative Analysis

Finally, we show the results of the predicted answerability for various scenes and questions in Figure 4. In the first example, the question "What is next to the coffee table?" is given and the AnsFields show the score of answerability to predict "shelf" as the answer. This question requires the identification of a location where the "coffee table" and surrounding objects are well visible. The correct panoramic image shows that an image has been obtained that clearly shows the positional relationship between "coffee table" and "shelf". However, the prediction shows that object recognition is working well, with higher scores for locations near the "coffee table" where the positional relationship of the surrounding objects is easy to identify. However, the panoramic image in the predicted location is difficult to recognize the "shelf." In the second example, the model is required to predict the location of "kitchen cabinets". In this example, TopPoint is highlighted in AnsFields. As with the correct answer image, the prediction results show higher scores for locations where "kitchen cabinets" and

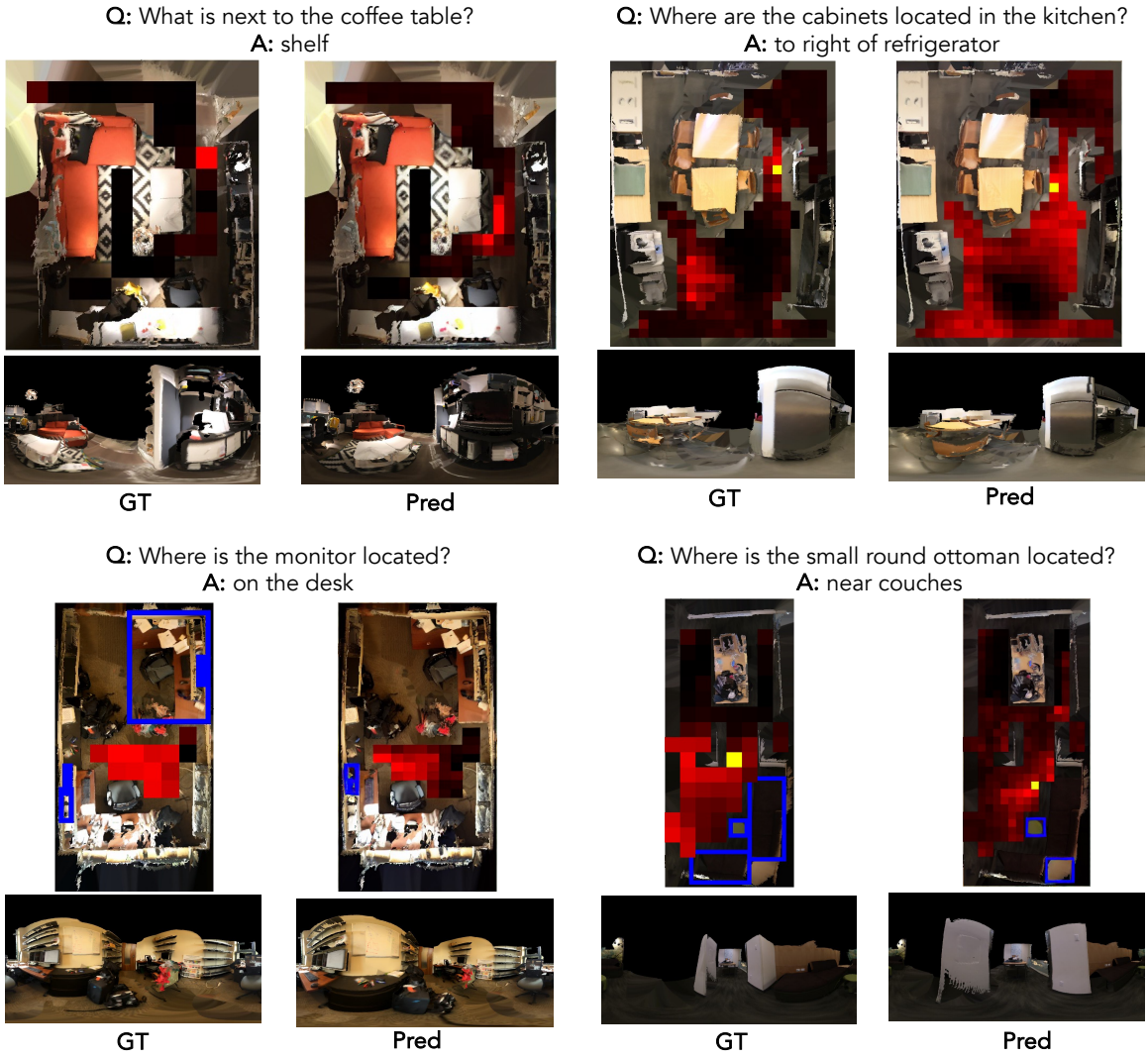

Fig. 4. We used InstructPix2Pix to predict answerability fields for a variety of unknown scenes. The figure shows AnsFields to the question and the panoramic image taken in the location of the highest score of answerability.

the appropriate answer object, "refrigerator", are properly shown, and lower scores for locations where either is not shown. The third example is the result when train images are annotated with bounding boxes. The question asks for the location of the "monitor". There are several monitors in this scene and all of them are on the desk. The correct image annotates all of the monitors and the desk where they are located. In the prediction results, although there is only one, the monitor is surrounded by a blue frame and the AnsFields prediction is close to the correct image. Finally, we show an example when images used for model training are annotated with bounding boxes and drawn with the top point highlighting. The locations of "ottomans" and "coaches" need to be properly predicted from within the scene to identify where they can be seen well. Some parts of bounding box prediction were successful and the result of AnsFields are predicted "ottomans" and "coaches" were well visible. Thus, we have successfully used the diffusion model to understand the character of the objects and their location in the scene needed to answer a question and to predict the appropriate location to answer the question.

VII. Conclusion

This paper introduced Answerability Fields as a novel approach to predicting answerability within complex indoor environments. Leveraging data from the ScanQA dataset, we constructed a comprehensive AnsFields data encompassing diverse scenes and questions from ScanNet. Utilizing a diffusion model, we successfully inferred and evaluated the AnsFields, shedding light on objects' importance and locations in answering questions within a scene. Our results highlight the efficacy of our approach in guiding scene navigation tasks by utilizing the gradient of AnsFields. By demonstrating the potential application of this method, we pave the way for enhanced interactions between intelligent agents and their environments. Moving forward, we envision further advancements and applications of AnsFields in various domains, ultimately contributing to the advancement of embodied artificial intelligence.

VIII. Acknowledgements

This work was supported by JST PRESTO Grant Number JPMJPR22P8, and JSPS KAKENHI Grant Numbers JP21K12055, JP22K12159, JP22K17983, JP22KK0184, Japan. This research project has benefitted from the Microsoft Accelerate Foundation Models Research (AFMR) grant program through which leading foundation models hosted by Microsoft Azure along with access to Azure credits were provided to conduct the research.

References

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

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

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, June 2022.

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

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[6] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *In Neural Information Processing Systems*, 2020.

[7] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022.

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural Modular Control for Embodied Question Answering. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2018.

[11] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Clip on wheels: Zero-shot object navigation as object localization and exploration. *ArXiv*, abs/2203.10421, 2022.

[12] Zhangxuan Gu, Haoxing Chen, Zhuoer Xu, Jun Lan, Changhua Meng, and Weiqiang Wang. Diffusioninst: Diffusion model for instance segmentation. *arXiv preprint arXiv:2212.02773*, 2022.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[15] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[16] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023.

[17] Dimitrios Kanoulas Jianwei Liu, Maria Stamatopoulou. Dipper: Diffusion-based 2d path planner applied on legged robots. *arXiv preprint arXiv:2309.14341*, 2024.

[18] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *CoRL*, pages 651–673, 2018.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[20] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada, July 2023. Association for Computational Linguistics.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[22] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[23] Nelson Nauata, Sepidehsadat Hosseini, Kai-Hung Chang, Hang Chu, Chin-Yi Cheng, and Yasutaka Furukawa. House-gan++: Generative adversarial layout refinement network towards intelligent computational agent for professional architects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13632–13641, 2021.

[24] Santhosh K. Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Computer Vision and Pattern Recognition (CVPR), 2022 IEEE Conference on*. IEEE, 2022.

[25] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *CVPR*, 2022.

[26] Marc Rigter, Jun Yamada, and Ingmar Posner. World models via policy-guided trajectory diffusion. *arXiv preprint arXiv:2312.08533*, 2023.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[29] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv: Arxiv-2210.05663*, 2022.

[30] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023.

[31] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[32] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022.

[33] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[34] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. Transporter networks: Rearranging the visual world for robotic manipulation. *Conference on Robot Learning (CoRL)*, 2020.

[35] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[36] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens, 2023.

[37] Xiyue Zhu, Vlas Zyrianov, Zhijian Liu, and Shenlong Wang. Mapprior: Bird's-eye view perception with generative models, 2023.