# Look and Answer the Question: On the Role of Vision in Embodied Question Answering

**Nikolai Ilinykh** and **Yasmeen Emampoor** and **Simon Dobnik**
Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
University of Gothenburg, Sweden
{nikolai.ilinykh,simon.dobnik}@gu.se, gusemampya@student.gu.se

## Abstract

We focus on the Embodied Question Answering (EQA) task, the dataset and the models (Das et al., 2018). In particular, we examine the effects of vision perturbation at different levels by providing the model with either incongruent, black or random noise images. We observe that the model is still able to learn from general visual patterns, suggesting that they capture some common sense reasoning about the visual world. We argue that a better set of data and models are required to achieve better performance in predicting (generating) correct answers. The code is available here: https://github.com/GU-CLASP/embodied-qa.

## 1 Introduction

When language generation models are employed in real-world scenarios, they need to correctly perceive the environment, understand physics between objects and reason about the events in order to produce logical and correct descriptions (Lake et al., 2017). In order to study and ultimately construct such models, several language-and-vision tasks were developed including Visual Question Answering (VQA) (Antol et al., 2015; Gordon et al., 2018) and Visual Dialogue (Das et al., 2017). The advantage of such models is their ability to process visual information *jointly* with language. However, several papers following have found that **vision is often dismissed** by the model and language is much more attended to. Attempts were made to influence this bias on *the dataset side* and make the contributions of both modalities more equal. For example, Goyal et al. (2017) show that coupling questions in a VQA dataset with complementary images, which lead to different responses, makes the model learn more from vision and less from language biases. A different way of tackling the language bias in VQA datasets is to augment them with a larger variety of different question types, generated with either a template-based method or neural networks (Kafle et al., 2017). Caglayan et al. (2019) note that there exists a dataset structure bias realised through short and repetitive texts, which in principle could inhibit gains from vision. On the other hand, many papers have proposed *models* capable of better fusion between vision and language. Zheng et al. (2020) introduce a method to learn better alignment between language and vision spaces based on reasoning over entities in texts and objects in images for the VQA task. Work on multi-modal machine translation looked at the model performance when images are replaced with incongruent scenes (Elliott, 2018) or leveraging the importance of vision modality by testing different fusion techniques (Raunak et al., 2019).

VQA models cannot be directly applied in the real world scenario due to challenges that require direct interaction of the model with the environment. Therefore the task of Embodied Question Answering (EQA) has been proposed by Das et al. (2018) which is very much different from the standard VQA. It combines question answering with a preceding navigation task in the environment, first looking for a target object that the question is about. When the agent reaches the navigation endpoint, the system answers the question based on the view from its final position. Therefore, the success of the navigation directly affects the accuracy of question answering. EQA task is much harder than VQA, because (i) the robot does not contain a human model of attention (Dobnik and Kelleher, 2016), (ii) there is no guarantee that navigation will be successful, (iii) all questions relate to home environments, which are more similar to each other than unconstrained situations in the photographs used for VQA, and (iv) questions are limited in vocabulary, scope and complexity which restructs the language and makes it even a stronger predictor. To support the latter, Thomason et al. (2019) have shown that a language-only model outper-

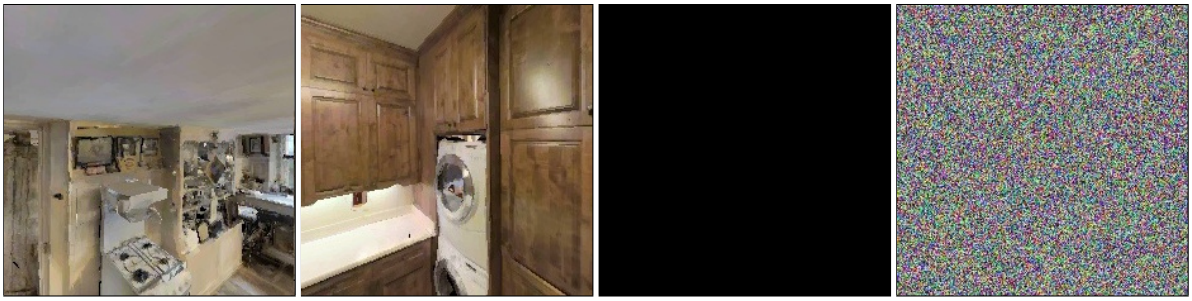

Figure 1: Example of successive removal of context, content and structure. For each removal type, we show the first frame from the set of frames that the model takes to answer the question "What color is the stove in the kitchen?". From left to right: **original** (nothing is removed), **shuffled** (structure and content are present, but context is incorrect), **blind** (no content and context, but structure), **random** (most disturbed representation).

forms multi-modal or vision-only system during QA in the EQA task. This demonstrates a stronger need for the deeper analysis of how and to what extent vision can be even utilised in the EQA model.

While most of the existing research on EQA has focused on the navigation subtask (Wijmans et al., 2019; Yu et al., 2019; Batra et al., 2020), in this work we examine **the general role of vision for the QA in the EQA task**. In particular, we investigate how EQA model is using visual information and whether it is sensitive to visual perturbations when answering the question. First, we confirm previous results, comparing models trained and tested in different uni-/multi-modal conditions showing that just as in the VQA task, the model in the EQA task tends to hallucinate and disregard vision. Second, we turn to the examination of *how different visual disturbances affect performance of the model*. We evaluate the model with images of different types exemplified in Figure 1. The effects of various disturbances reflected in the evaluation scores will tell us how much removing context, content and (or) structure from images impacts question answering.

Our study can be viewed as *a test bed* to understand how vision is used in the EQA task. Similar benchmarks were developed for VQA (Agrawal et al., 2018) and person-centric visual grounding (Luo et al., 2022). In terms of the EQA, most of the work examined what can be used *instead* of the visual features. For example, Hu et al. (2019) show that using route structures instead of visual representations is better for the task. Schumann and Riezler (2022) found out that the model relies on properties of the environment graph much more rather than on visual features in the EQA for outdoor scenes. Different from previous studies, here we do not completely remove visual modality or compare it against other modalities. Instead, we evaluate *the limits* of the existing EQA model when its vision is permuted. We also view the EQA task as a simple NLG task, e.g. the model is asked to map important parts in vision and language (content selection) followed by prediction of a *single* label (surface realisation). In general, the focus of this paper is to understand the interplay between different modalities used in this simple generation scenario which is also relevant for generation of longer sequences of descriptions.

## 2 Task Description

**Models** The EQA task is split into two subtasks: navigation and question answering. Below we briefly describe the models used for both subtasks, a more detailed scheme is provided in Appendix A. The navigation starts with an LSTM-based *planner* (Hochreiter and Schmidhuber, 1997) that selects an action from a pre-defined set (turn left, turn right, forward, stop) based on the question $\mathbf{Q}$, last action $\mathbf{a}_{t-1}$, last hidden state $\mathbf{h}_{t-1}$ and visual representation $\mathbf{V}_t = F(\mathbf{I}_t)$, where $F$ is a convolutional network (Cun et al., 1990) pre-trained on three tasks: RGB reconstruction, semantic segmentation, depth estimation. Next, the current hidden state of the planner $\mathbf{h}_t$, the predicted action $\mathbf{a}_t$ and the current visual input $\mathbf{V}_t$ are given to the *controller* that decides how many times the action has to be executed. The visual input $\mathbf{V}$ is updated for each reiteration of the action. The controller is a simple multi-layer perceptron that returns control to the planner once it concludes that it needs a new action. The question answering module is an information fusion network. The question $\mathbf{Q}$ is encoded by an LSTM network, while $F$ takes $N$ frames from the end of the navigation $I_{T-N}, \ldots, I_T$ once the agent has decided to stop (as predicted by the planner) or

the maximum number of actions $T = 100$ has been taken. Both representations are jointly attended and passed through a multi-layer classifier to predict a probability distribution across the answers.

**Dataset** The EQA dataset consists of automatically generated questions and answers from rules. The questions are made over visual scenes from the Matterport3D dataset (Chang et al., 2017) from which answers are generated. The authors use Habitat (Savva et al., 2019) to render the visual scenes. Each question in the dataset is replicated 15 times with different coordinates for the initial position of the agent as there is no single navigation path to the target object. There are three types of questions in the published dataset:

- colour: *What colour is the* `OBJ`*?*
- colour_room: *What colour is* `OBJ` *in the* `ROOM`*?*
- location: *What room is the* `OBJ` *located in?*

Nearly 70% of all questions are of colour_room type, ∼15% are of colour type and the rest (∼15%) are of location type. Placeholders $OBJ$ and $ROOM$ are filled with objects from dataset annotations (e.g., chair, plant) and room types (e.g., bathroom, kitchen) respectively.

**Dataset and model limitations** We describe several issues related to the EQA dataset. First, the quality of the rendered scenes is often poor, negatively affecting both navigation and question answering (Appendix B). Annotations of answers are sometimes questionable, including the ways the set of possible answers has been defined (e.g., limited set of possible colours in the scene) (Appendix C). A different concern is the "naturalness" of questions. Some questions are highly atypical of real interactions, e.g. why would one ask "What colour is the table in the living room?". Another problem is that house environments are visually similar, consisting of instances of the same object classes (e.g., sofas, plants) that often share the same attributes (e.g., sofas are brown, plants are green). This also leads to an unbalanced distribution of answers: some answers ("black" and "brown") are over-represented in the dataset, possibly allowing the model to exploit these priors, e.g. sofas are often brown. Although this dataset bias amplifies the model's ability to answer many questions about similar objects, artificially inflating accuracy on this dataset, the same biases prevent it from correctly answering questions about objects with specific properties, which require fine-grained visual understanding. Therefore in order to truly use vision to answer questions (e.g., when sofa is red, not brown), the model must have a *deeper* understanding of *fine-grained* visual representations, but as shown by Anand et al. (2018), the EQA models often struggle to utilise visual input. In the following sections, we will examine the level of visual understanding of the EQA model and overview problems on the dataset and modelling side that make it learn so little from vision.

## 3 Is language really stronger in EQA?

In the first set of experiments, we change the model's vision stream or visual input representations. **Vis-L** is the standard EQA model (Das et al., 2018) without any perturbations on the vision side. Given the question $\mathbf{Q}$ and $\mathbf{N}$ image frames, the model predicts the most probable answer $a^\star$:

$$a^\star = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathrm{P}(a|\mathbf{Q}, \mathbf{I}_{T-N}, \ldots, \mathbf{I}_T). \quad (1)$$

For the **Blind-L** model, we keep the vision stream in the model, but change visual representations. In particular, we replace them with arrays of zeros before they are passed to the CNN for pre-processing:

$$\mathbf{I}_t = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \quad \mathbf{I}_t \in \mathbb{R}^{3\times256\times256}. \quad (2)$$

Finally, in the **Ø-L** model, we completely remove the vision stream and train it on questions only:

$$a^\star = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathrm{P}(a|\mathbf{Q}). \quad (3)$$

We run all three models for 50 epochs using the official implementation[1] and choose the checkpoints with the lowest validation loss. For evaluation, we calculate accuracy (the top answer) and the mean rank (position of the correct answer in the ranked list of answers by the predicted probability distribution). We also compute Cohen's Kappa (Artstein and Poesio, 2005) which measures the agreement between the classifier and the ground truth dataset corrected by agreement by chance which is based on the distribution of labels. A kappa close to 0

[1] https://github.com/facebookresearch/habitat-lab/tree/main/habitat_baselines/il

(which ranges from 0 to 1 for agreement and 0 to −1 for disagreement) indicates that most agreement can be predicted only by knowing a distribution of labels. The higher the kappa the more the classifier is utilising additional knowledge that it has learned beyond a distribution of labels.

| Metric | Vis-L | Blind-L | Ø-L |
|---|---|---|---|
| ↓ Overall Mean Rank (MR) | 4.352 | 4.454 | **3.685** |
| MR, Color Room Questions | 3.611 | **3.157** | 3.247 |
| MR, Color Questions | 2.693 | **2.261** | 2.304 |
| MR, Location Questions | 10.137 | 13.667 | **7.611** |
| ↑ Overall Accuracy (A) | **0.38** | 0.323 | 0.362 |
| A, Color Room Questions | **0.374** | 0.348 | 0.337 |
| A, Color Questions | **0.528** | 0.478 | 0.522 |
| A, Location Questions | 0.222 | 0 | **0.278** |
| Kappa Score | -0.005 | 0.014 | **0.024** |

Table 1: Results for the models both *trained* and *evaluated* with the specified settings described in Section 3. We also report results per question type. The best scores are coloured in **blue**.

The results are shown in Table 1. The **Vis-L** model has the highest overall accuracy. However, the kappa score close to 0 shows that the model has a similar performance to a model that has memorised the distribution of labels. The lower mean ranks for **Blind-L** and **Ø-L** show that they are better at approximating the correct answer than the **Vis-L** model. These models strongly learn from language since the lack of vision does not prevent them from learning from biases in the dataset, leading to higher ranks. The **Vis-L** model however needs to process vision, but it is not capable of doing that (Thomason et al., 2019). Thus vision interferes and obstructs it from learning from language biases, confusing the problematic model and leading to lower ranks of the correct answers. When breaking the results based on question types, colour question are generally the easiest to answer, followed by the colour_room and location questions. The location questions are the hardest to predict in terms of accuracy and ranking overall. Furthermore, they are also most affected by different model configurations. In particular, the results suggest that the location questions are better predicted from language alone (**Ø-L**). The **Blind-L** model has the worst ranks and the worst accuracy overall. Its inconsistent performance across question types is hard to explain. Possibly, irrelevant visual information (black images) makes it more unpredictable than no vision at all or complete vision. Although the **Blind-L** is not the optimal model, it is still not far off from the other two models due to the second source of information - language.

Overall, we partially replicate the results of Thomason et al. (2019) and observe that vision is not that crucial. The role of language is much stronger than the role of vision, as demonstrated by the performance of the **Ø-L** model that predicts answers from questions alone. However, Frank et al. (2021) show that diminishing the importance of vision is detrimental for language tasks. Therefore in the second experiment we investigate *how different visual perturbations are utilised by the model* and *what are the model's limits in learning from vision*. We are particularly interested in examining if the model is able to understand complex high-level patterns from images or does it only learn lower-level information, which is present in some form in different visual permutations.

## 4 "How much" vision is required?

To understand the limits of the model when utilising vision, we ask the following question: how much information can the model extract from different visual representations? We train the model according to Eqn. 1, but *evaluate* it on the vision with various levels of perturbations. In the **Eval-Shuffled** set-up, the model is provided with incorrect images for a specific question. In this case, the model gets structurally plausible representations which do not contain object(s) that the question asks about since the images depict a different house or room. We give more details about shuffling in the Appendix D. The **Eval-Blind** model has been evaluated on images which were transformed into arrays of zeros, following Eqn. 2. In **Eval-Random**, the model has been given arrays of random noise as its visual input. The image vectors were replaced by an array of the specified shape $(3 \times 256 \times 256)$ that was populated with random samples from a uniform distribution:

$$\mathbf{I}_t = \begin{bmatrix} \mathbf{v} & \cdots & \mathbf{v} \\ \vdots & \ddots & \vdots \\ \mathbf{v} & \cdots & \mathbf{v} \end{bmatrix}, \quad \mathbf{v} \in [0, \ldots, 1]. \quad (4)$$

Results in Table 2 demonstrate that each of the **Eval-** configurations results in lower performance compared to the baseline (**Vis-L**). However, the model performs better on both incongruent (**Eval-Shuffled**) and black (**Eval-Blind**) images rather than random noise (**Eval-Random**). This suggests that the model is using *visual patterns* to support its

| Metric | Vis-L | Eval-Shuffled | Eval-Blind | Eval-Random |
|---|---|---|---|---|
| ↓ Overall Mean Rank (MR) | 4.352 | 5.145 | 5.508 | 6.899 |
| MR, Color Room Questions | 3.611 | 4.157 | 4.562 | 5.512 |
| MR, Color Questions | 2.693 | 3.035 | 3.087 | 3.319 |
| MR, Location Questions | 10.137 | 12.722 | 13.278 | 18.33 |
| ↑ Overall Accuracy (A) | 0.38 | 0.266 | 0.246 | 0.211 |
| A, Color Room Questions | 0.374 | 0.264 | 0.258 | 0.258 |
| A, Color Questions | 0.528 | 0.307 | 0.217 | 0.194 |
| A, Location Questions | 0.222 | 0.222 | 0.222 | 0 |
| Kappa Score | -0.005 | 0.013 | 0.004 | -0.005 |

Table 2: Results for the models trained with original data (as **Vis-L**), but *evaluated* with specified conditions, described in Sec. 4). We also report results per question type. Intensity of the blue colour indicates performance of the model for the specific metric (more intensity means better performance).

prediction in some way. The performance across question types is similar to the results for models from the first set of experiments in Table 1: location questions are the hardest, colour questions are the easiest. Both experiments suggest that the visual information is not used as much as one would hope - disturbing vision or completely removing it has little effect on the overall performance, suggesting that the model exploits language more. In terms of accuracy, location questions (which have the lowest accuracy on the baseline) are affected the least by different visual input. One reason could be that the baseline is bad so there is not much room for decrease in performance. Another reason could be that there are only 15 *distinct* location question-answer pairs in the evaluation set, seven of which are also found in the training. This may be the reason for a more exploitable language bias for location questions compared to other types.

## 5 EQA: biases and limitations

Recently, Hirota et al. (2022) have discovered social and gender biases in the VQA dataset. In the EQA, on the other hand, the model acts in the house environments with household objects without any humans, meaning that there are no biases towards any social group. The nature of dataset problems in the EQA task is different from VQA. One of the primary problems of the EQA is the lack of the perfect navigation module that would select correct images as input to the QA module. In addition, even if navigation is perfect, there is a chance for an image to be badly rendered (Appendix B). These problems combined make the task harder and bridge it with the likes of captioning of images taken by visually impaired people (Gurari et al., 2018) instead of VQA where images are fixed and taken in perfect conditions to answer the question. Another problem is of the limited scope of automatically generated questions and distribution of answers. In our view this directly forces the model to rely on language (which is limited and predictable) and to consider only basic visual patterns.

## 6 Conclusion

We looked at the Embodied Question Answering task and the corresponding dataset, focusing on how much vision is exploited by the QA module. The novelty of our study is the examination of *how* and *what* does the model learn from different types of images. Our results suggest that even if vision is not properly used, the model can extract general patterns from different visual permutations that are helpful to some degree. This means that the model could be looking at incongruent images or images with homogeneous structure (black) and answer questions correctly. Overall, we show that the model captures low-level knowledge of vision but is not capable of identifying and reasoning about specific high-level visual contexts that require understanding of scenes at a fine-grained level. Future work can improve model's vision by implementing cognitive attention (Dobnik and Kelleher, 2016; Kruijff-Korbayová et al., 2015) or splitting the QA task into more subtasks because QA involves several inference steps and is not a simple pattern matching procedure. Using pre-trained multi-modal transformers such as LXMERT (Tan and Bansal, 2019) could also tell us whether these models are able to overcome problems related to dataset construction and image selection for the QA task in the EQA. If a performance of such a model improves then it must be the case that transformers capture some common sense knowledge through pre-training, but this could also be a hallucination of a different kind: it is hallucination because it is general V&L knowledge not the specific one arising from a particular image and text.

## Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

## References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980.

Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron C. Courville. 2018. Blindfold baselines for embodied QA. *CoRR*, abs/1811.05013.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Ron Artstein and Massimo Poesio. 2005. Kappa$^3$ = alpha (or beta). Technical report, University of Essex Department of Computer Science. Available at: `http://ron.artstein.org/publications/kappa3.pdf`.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. 2020. Objectnav revisited: On evaluation of embodied agents navigating to objects. *CoRR*, abs/2006.13171.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. 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 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

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. Cite arxiv:1709.06158.

Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. 1990. *Handwritten Digit Recognition with a Back-Propagation Network*, page 396–404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, New Brunswick, NJ. SEMDIAL.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4089–4098.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Gender and racial bias in visual question answering datasets. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1280–1292. ACM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. 2019. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6551–6557, Florence, Italy. Association for Computational Linguistics.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202, Santiago de Compostela, Spain. Association for Computational Linguistics.

Kenneth L. Kelly. 1965. Twenty-two colors of maximum contrast. *Color Engineering*, 3:26–27.

Ivana Kruijff-Korbayová, Francis Colas, Koen Hindriks, Mark Neerincx, Petter Ögren, Mario Gianni, Tomáš Svoboda, and Rainer Worst. 2015. TRADR Project: Long-Term Human-Robot Teaming for Robot Assisted Disaster Response. *KI - Künstliche Intelligenz*, 29(2):193–201.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Yiran Luo, Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2022. To find waldo you need contextual cues: Debiasing who's waldo. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 355–361, Dublin, Ireland. Association for Computational Linguistics.

Vikas Raunak, Sang Keun Choe, Quanyang Lu, Yi Xu, and Florian Metze. 2019. On leveraging the visual modality for neural machine translation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 147–151, Tokyo, Japan. Association for Computational Linguistics.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9339–9347.

Raphael Schumann and Stefan Riezler. 2022. Analyzing generalization of vision and language navigation to unseen outdoor areas. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7519–7532, Dublin, Ireland. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the baseline: Single modality performance on visual navigation & QA. 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 1977–1983, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6659–6668.

Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, and Dhruv Batra. 2019. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6309–6318.

Chen Zheng, Quan Guo, and Parisa Kordjamshidi. 2020. Cross-modality relevance for reasoning on language and vision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7642–7651, Online. Association for Computational Linguistics.

## A Baseline QA Model

Fig. 2 shows the architecture of the baseline model for question answering in the EQA task. The model consists of three parts: language encoder, vision encoder and attention across both modalities. Questions are processed by a standard LSTM network (Hochreiter and Schmidhuber, 1997) that also learns word embeddings from scratch. $B = 20$ stands for the batch size, $N = 5$ is the number of used image frames taken from the last steps of navigation, $L = 11$ is the question maximum length, and $M = 64$ is the dimension size. Note that each question representation is repeated $N$ times. Images are represented as three-channel (RGB) $256 \times 256$ egocentric scenes from the Habitat's image renderer. A CNN network that has been pre-trained for RGB reconstruction, semantic segmentation and depth estimation is used to process images. The fully connected layer refers to a sequence of a linear layer, a ReLU layer, and a dropout layer with $p = 0.5$. $D = 4608$ is the dimension size of the visual processing network. The output representations from the language and

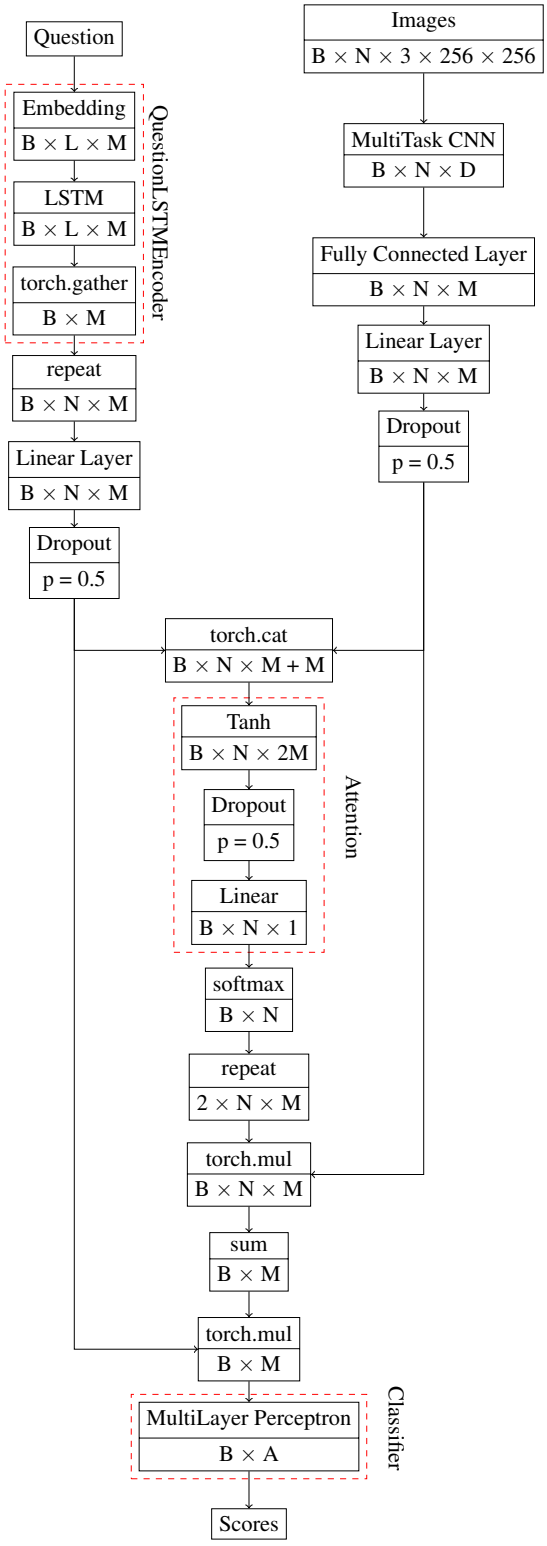


Figure 2: The baseline question answering model described in Das et al. (2018) with available implementation in Habitat-Lab, link: https://github.com/facebookresearch/habitat-lab/tree/main/habitat_baselines/il. We schematically show the key components of the model: QuestionLSTMEncoder, Attention, and Answer Classifier. The stream in the top right side corresponds to the processing of visual information.

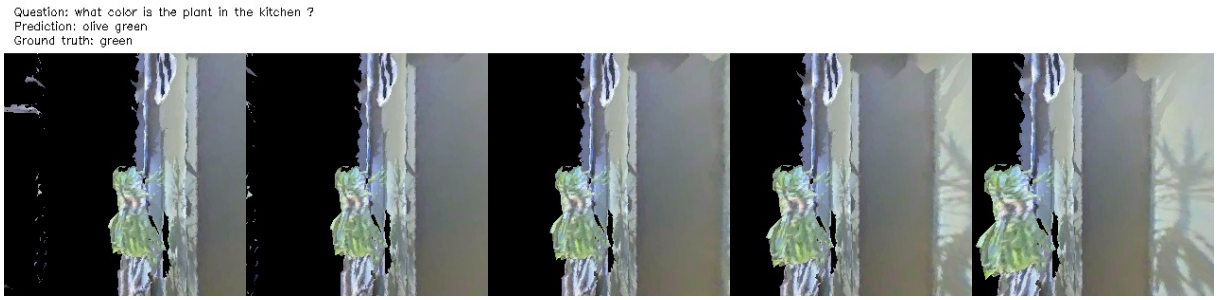


Figure 3: Example of a badly rendered scene from the EQA dataset.

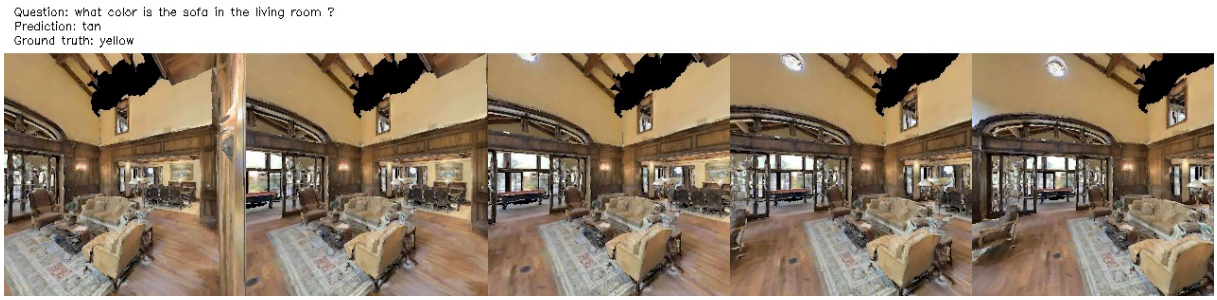


Figure 4: Example of a sequence of images, the question, the predicted answer and the ground-truth answer.

vision encoder are jointly attended and summed across $N$ frames. The resulting representation is passed to a multi-layer perceptron to predict the scores across $A = 35$ possible answers. We ran all models on 4 NVIDIA GeForce GTX 1080 Ti GPUs, running time was approximately 4 hours per model. In all experiments we report results for the models with the minimal loss across 50 epochs. In our experiments we did not use any explicit tools except Habitat-Lab[1], release version 0.1.7, MIT license.

## B Image Rendering Problem

While the majority of scenes are rendered properly, some of the scenes could be of poor quality. An example is shown in Fig. 3, demonstrating that the last five image frames used to answer the question include a lot of visual noise which makes the scene very confusing for a human eye. One wonders how an agent processes such poorly rendered scenes: does it rely on language information to answer the question? Note that scene annotations often include an object named "void" which is simply a black space. It is possible that the agent will encounter such confusing and uninformative space at the end of its navigation path. This could either confuse the agent or enforce better learning from the language stream. Or can the agent infer the answer from the general colours in the scene, given that the navigation often finishes at a close proximity to the target object? The example that we show is intended to demonstrate that due to the quality of the visual input, the agent might be biased to strongly learn from language and dataset biases.

## C Colour Problem

The EQA dataset has been generated automatically which means that it might contain errors. An example is shown in Fig. 4, where the question answering model has answered "tan" when asked about the colour of the sofa in the living room. One could say that, when looking at the image, the sofa is indeed tan, while there is an yellow armchair next to it. It could be that the model is actually correct in its prediction for a good reason and annotations are incorrect. The problem with colour annotations is also related to the set of colours used by annotators, that is coming from Kenneth L. Kelly's "Twenty-two colours of maximum contrast" (Kelly, 1965) with two additional colours: "off-white" and "slate grey". This set of colours has been designed to describe situations when contrast is needed (e.g., colour coding of graphs), not necessarily to depict colours in real world with natural descriptions. For example, the set introduces "buff" and "yellowish pink", the former one is replaced with "tan" in the EQA dataset and the latter one is simply replaced with "yellow pink", which makes the dataset even

less natural. In addition, many colours in this set are easily confused under different lighting conditions ("white" and "off-white", "grey" and "slate grey"), complicating the task for the question answering model.

## D Example Episode

An example episode structure from the EQA dataset. We display only a part of the shortest path coordinates and viewpoint lists. In **Eval-Shuffle**, shuffling is performed by modifying the original set of image frames and creating a new one. We show an example of one navigation episode from the EQA dataset below. A single episode includes a `question` field, which includes the question, answer, question type, and answer token IDs. We shuffle these `question` fields (line 68 in the example structure) across different episodes. Note that the authors of the dataset duplicated questions across multiple episodes, which, however, have different navigation paths to the target. This has been implemented in order to ease the navigation task since there is no single correct navigation to the target object. We acknowledge that it could be possible that an episode with a shuffled question still has a valid set of last $N$ image frames, but this possibility is low – for a single question, this probability is less than one percent.

```
1  {'episode_id': '640',
2   'scene_id': 'mp3d/5LpN3gDmAk7/5LpN3gDmAk7.glb',
3   'start_position': [15.50573335967819,
        ↪ -0.7660300302505512, 8.392731789742543],
4   'start_rotation': [-5.312086480921031e-17,
5    -0.8526401643962381,
6    -0.0,
7    0.522498564647173],
8   'info': {'bboxes': [{'type': 'object',
9      'box': {'centroid': [13.2358, -14.5238, 0.497693],
10      'a0': [1.0, 0.0, 0.0],
11      'a1': [0.0, 1.0, 0.0],
12      'a2': [0.0, 0.0, 1.0],
13      'radii': [0.593273, 0.243441, 1.68627],
14      'obj_id': 305,
15      'level': 0,
16      'room_id': 18},
17     'name': 'door',
18     'target': True},
19    {'type': 'room',
20     'box': {'centroid': [10.874245, -11.97072, 0
            ↪ .5380600000000001],
21      'a0': [1.0, 0.0, 0.0],
22      'a1': [0.0, 1.0, 0.0],
23      'a2': [0.0, 0.0, 1.0],
24      'radii': [3.1686549999999998, 3.26178, 1.95437],
25      'room_id': 18,
26      'level': 0},
27     'name': ['kitchen'],
28     'target': False}],
29    'question_meta': [{'name': 'colour', 'diffuse':
        ↪ 'grey'}],
30    'question_answers_entropy': 0.8303560860446519,
31    'level': 0},
32   'goals': [{'position': [13.2358, 0.4976929999999973, 14
        ↪ .5238],
33     'radius': 0.6412771421234348,
34     'object_id': 305,
35     'object_name': 'door',
36     'object_category': 'object',
37     'room_id': 18,
38     'room_name': 'kitchen',
39     'view_points': [{'position': [12.985883260576134,
40        -1.246680130110505,
41        14.494095338174798],
42       'rotation': [-2.855981544936522e-28,
43        -0.7071067811874078,
44        -0.0,
45        0.7071067811856873]},
46          ...
47      {'position': [13.089462756345679,
            ↪ -1.246680130110505, 13.9761978593327065],
48       'rotation': [-1.2227381688226952e-16,
49        -0.8910065241891411,
50        -0.0,
51        0.453990499738029 35]}]}],
52   'start_room': 'R22',
53   'shortest_paths': [[{'position': [15.50573335967819,
54       -0.7660300302505512,
55       8.392731789742543],
56      'rotation': [-5.312086480921031e-17,
57       -0.8526401643962381,
58       -0.0,
59       0.522498564647173],
60      'action': 2},
61          ...
62     {'position': [13.042462387438766,
            ↪ -0.7660300302505512, 13.951177365325918],
63      'rotation': [-1.2227381690007914e-16,
64       -0.8910065242228339,
65       -0.0,
66       0.45399049967190386],
67      'action': 3}]],
68   'question': {'question_text': 'what colour is the door
        ↪ in the kitchen?',
69    'answer_text': 'grey',
70    'question_tokens': [4, 5, 6, 7, 19, 9, 7, 10],
71    'answer_token': [0, 0, 0, 0],
72    'question_type': 'colour_room'}}
```