Shifting the Baseline: Single Modality Performance on Visual Navigation & QA

Jesse Thomason **Daniel Gordon** **Yonatan Bisk**
Paul G. Allen School of Computer Science and Engineering
jdtho@cs.washington.edu

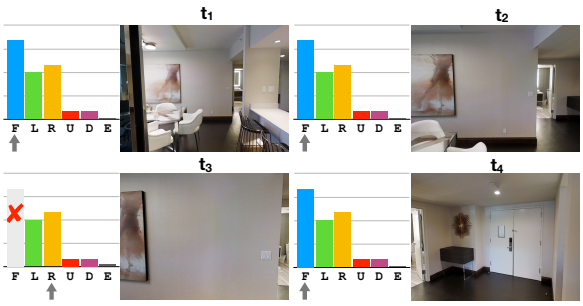
Abstract

We demonstrate the surprising strength of unimodal baselines in multimodal domains, and make concrete recommendations for best practices in future research. Where existing work often compares against *random* or *majority class* baselines, we argue that unimodal approaches better capture and reflect dataset biases and therefore provide an important comparison when assessing the performance of multimodal techniques. We present unimodal ablations on three recent datasets in visual navigation and QA, seeing an up to 29% absolute gain in performance over published baselines.

1 Introduction

All datasets have biases. Baselines should ca ture these regularities so that outperforming th indicates a model is actually solving a task. multimodal domains, bias can occur in any subset of the modalities. To address this, we argue it is not sufficient for researchers to provide random or majority class baselines; instead we recommend presenting results for unimodal models. We investigate visual navigation and question answering tasks, where agents move through simulated environments using egocentric (first person) vision. We find that unimodal ablations (e.g., language only) in these seemingly multimodal tasks can outperform corresponding full models (§4.1).

This work extends observations made in both the Computer Vision (Goyal et al., 2018; Cirik et al., 2018) and Natural Language (Mudrakarta et al., 2018; Glockner et al., 2018; Poliak et al., 2018; Gururangan et al., 2018; Kaushik and Lipton, 2018) communities that complex models often perform well by fitting to simple, unintended correlations in the data, bypassing the complex grounding and reasoning that experimenters hoped was necessary for their tasks.



Actions: **F**orward, turn **L**eft & **R**ight, tilt **U**p & **D**own, **E**nd

Figure 1: Navigating without vision leads to sensible navigation trajectories in response to commands like "walk past the bar and turn right". At t_3, "forward" is unavailable as the agent would collide with the wall.

We ablate models from three recent papers: (1) navigation (Figure 1) using images of real homes paired with crowdsourced language descriptions (Anderson et al., 2018); and (2, 3) navigation and egocentric question answering (Gordon et al., 2018; Das et al., 2018a) in simulation with synthetic questions. We find that unimodal ablations often outperform the baselines that accompany these tasks.

Recommendation for Best Practices: Our findings show that in the new space of visual navigation and egocentric QA, all modalities, even an agent's action history, are strongly informative. Therefore, while many papers ablate either language *or* vision, new results should ablate *both*. Such baselines expose possible gains from unimodal biases in multimodal datasets irrespective of training and architecture details.

2 Ablation Evaluation Framework

In the visual navigation and egocentric question answering tasks, at each timestep an agent receives an observation and produces an action. Actions can move the agent to a new location or heading

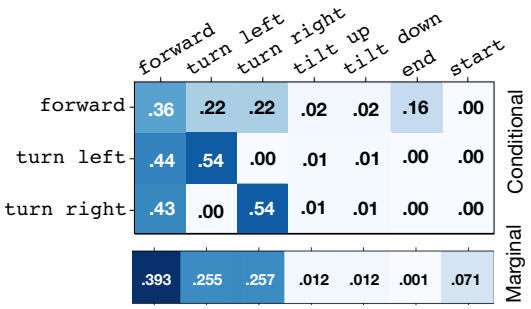


Figure 2: $P(act = col|prev = row)$ and marginal action distributions in Matterport training. Peaked distributions enable agents to memorize simple rules like not turning left immediately after turning right, or moving forward an average number of steps.

(e.g., *turn left*), or answer questions (e.g., *answer 'brown'*). At timestep t, a multimodal model $\mathcal{M}$ takes in a visual input $\mathcal{V}_t$ and language question or navigation command $\mathcal{L}$ to predict the next action a_t. The navigation models we examine also take in their action from the previous timestep, a_{t-1}, and 'minimally sensed' world information W specifying which actions are available (e.g., that *forward* is unavailable if the agent is facing a wall).

$$a_t \leftarrow \mathcal{M}(\mathcal{V}_t, \mathcal{L}, a_{t-1}; W) \quad (1)$$

In each benchmark, $\mathcal{M}$ corresponds to the author's released code and training paradigm. In addition to their full model, we evaluate the role of each input modality by removing those inputs and replacing them with zero vectors. Formally, we define the full model and three ablations:

$$\text{Full Model} \quad \text{is} \quad \mathcal{M}(\mathcal{V}_t, \mathcal{L}, a_{t-1}; W) \quad (2)$$
$$\mathcal{A} \quad \text{is} \quad \mathcal{M}(\vec{0}, \vec{0}, a_{t-1}; W) \quad (3)$$
$$\mathcal{A} + \mathcal{V} \quad \text{is} \quad \mathcal{M}(\mathcal{V}_t, \vec{0}, a_{t-1}; W) \quad (4)$$
$$\mathcal{A} + \mathcal{L} \quad \text{is} \quad \mathcal{M}(\vec{0}, \mathcal{L}, a_{t-1}; W) \quad (5)$$

corresponding to models with access to $\boldsymbol{\mathcal{A}}$ction inputs, $\boldsymbol{\mathcal{V}}$ision inputs, and $\boldsymbol{\mathcal{L}}$anguage inputs. These ablations preserve the architecture and number of parameters of $\mathcal{M}$ by changing only its inputs.

3 Benchmark Tasks

We evaluate on navigation and question answering tasks across three benchmark datasets: Matterport Room-to-Room (no question answering component), and IQUAD V1 and EQA (question answering that requires navigating to the relevant scene in the environment) (Anderson et al., 2018; Gordon et al., 2018; Das et al., 2018a). We divide the latter two into separate navigation and question answering components. We then train and evaluate models separately per subtask to analyze accuracy.

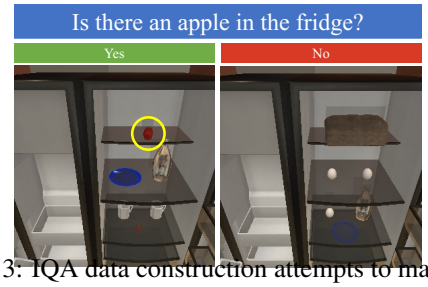


Figure 3: IQA data construction attempts to make both the question and image necessary for QA.

3.1 Matterport Room-to-Room

An agent is given a route in English and navigates through a discretized map to the specified destination (Anderson et al., 2018). This task includes high fidelity visual inputs and crowdsourced natural language routes.

Published Full Model: At each timestep an LSTM decoder uses a ResNet-encoded image $\mathcal{V}_t$ and previous action a_{t-1} to attend over the states of an LSTM language encoder ($\mathcal{L}$) to predict navigation action a_t (seen in Figure 2).

Published Baseline: The agent chooses a random direction and takes up to five forward actions, turning right when no forward action is available.

3.2 Interactive Question Answering

IQUAD V1 (Gordon et al., 2018) contains three question types: existence (e.g., *Is there a ...?*), counting (e.g., *How many ...?*) where the answer ranges from 0 to 3, and spatial relation: (e.g., *Is there a ... in the ...?*). The data was constructed via randomly generated configurations to weaken majority class baselines (Figure 3). To evaluate the navigation subtask, we introduce a new THOR-Nav benchmark.[1] The agent is placed in a random location in the room and must approach one of *fridge, garbage can,* or *microwave* in response to a natural language question.

Although we use the same full model as Gordon et al. (2018), our QA results are not directly comparable. In particular, Gordon et al. (2018) do not quantify the effectiveness of the QA component independent of the scene exploration (i.e. navigation and interaction). To remove the scene explo-

[1]Formed from a subset of IQUAD V1 questions.

ration steps of Gordon et al. (2018), we provide a complete ground-truth view of the environment.[2] We use ground-truth rather than YOLO (Redmon et al., 2016) due to speed constraints.

Nav Full Model: The image and ground-truth semantic segmentation mask $\mathcal{V}_t$, tiled question $\mathcal{L}$, and previous action a_{t-1} are encoded via a CNN which outputs a distribution over actions. Optimal actions are learned via teacher forcing.

Nav Baseline: The agent executes 100 randomly chosen navigation actions then terminates. In AI2THOR (Kolve et al., 2017), none of the kitchens span more than 5 meters. With a step-size of 0.25 meters, we observed that 100 actions was significantly shorter than the shortest path length.

Published QA Full Model: The question encoding $\mathcal{L}$ is tiled and concatenated with a top-down view of the ground truth location of all objects in the scene $\mathcal{V}$. This is fed into several convolutions, a spatial sum, and a final fully connected layer which outputs a likelihood for each answer.

Published QA Baseline: We include the majority class baseline from Gordon et al. (2018).

3.3 Embodied Question Answering

EQA (Das et al., 2018a) questions are programmatically generated to refer to a single, unambiguous object for a specific environment, and are filtered to avoid easy questions (e.g., *What room is the bathtub in?*). At evaluation, an agent is placed a fixed number of actions away from the object.

Published Nav Full Model: At each timestep, a planner LSTM takes in a CNN encoded image $\mathcal{V}_t$, LSTM encoded question $\mathcal{L}$, and the previous action a_{t-1} and emits an action a_t. The action is executed in the environment, and then a lower-level controller LSTM continues to take in new vision observations and a_t, either repeating a_t again or returning control to the planner.

Published Nav Baseline: This baseline model is trained and evaluated with the same inputs as the full model, but does not pass control to a lower-level controller, instead predicting a new action using the planner LSTM at each timestep (i.e., no hierarchical control). Das et al. (2018a) name this baseline *LSTM+Question*.

[2]This approximates the agent having visited every possible location, interacted with all possible objects, and looked in all possible directions before answering.

		Matterport↑ (%)		THOR-Nav↑ (%)		EQA↓ (m)
	Model	**Seen**	**Un**	**Seen**	**Un**	**Un**
Pub.	Full Model	27.1	19.6	77.7	18.08	4.17
	Baseline	15.9	16.3	2.18	1.54	4.21
Uni	$\mathcal{A}$	18.5	17.1	4.53	2.88	4.53
	$\mathcal{A}+\mathcal{V}$	21.2	16.6	**35.6**	**7.50**	*__4.11__
	$\mathcal{A}+\mathcal{L}$	**23.0**	*__22.1__	4.03	3.46	4.64
Δ	Uni – Base	**+7.1**	**+5.8**	**+33.4**	**+5.96**	**-0.10**

Table 1: Navigation success (Matterport, THOR-Nav) *(%)* and distance to target (EQA) *(m)*. **Best** unimodal: **better** than reported baseline; *better than full model.

Published QA Full Model: Given the last five image encodings along the gold standard navigation trajectory, $\mathcal{V}_{t-4}\ldots\mathcal{V}_t$, and the question encoding $\mathcal{L}$, image-question similarities are calculated via a dot product and converted via attention weights to a summary weight $\bar{\mathcal{V}}$, which is concatenated with $\mathcal{L}$ and used to predict the answer. Das et al. (2018a) name this oracle-navigation model *ShortestPath+VQA*.

QA Baseline: Das et al. (2018a) provide no explicit baseline for the VQA component alone. We use a majority class baseline inspired by the data's entropy based filtering.

4 Experiments

Across all benchmarks, unimodal baselines outperform baseline models used in or derived from the original works. Navigating unseen environments, these unimodal ablations outperform their corresponding full models on the Matterport (absolute $\uparrow$ 2.5% success rate) and EQA ($\downarrow$ $0.06m$ distance to target).

4.1 Navigation

We evaluate our ablation baselines on Matterport,[3] THOR-Nav, and EQA (Table 1),[4] and discover that some unimodal ablations outperform their corresponding full models. For Matterport and THOR-Nav, success rate is defined by proximity to the target. For EQA, we measure absolute distance from the target in meters.

Unimodal Performance: Across Matterport, THOR-Nav, and EQA, either $\mathcal{A}+\mathcal{V}$ or $\mathcal{A}+\mathcal{L}$

[3]We report on Matterport-validation since this allows comparing Seen versus Unseen house performance.

[4]For consistency with THOR-Nav and EQA, we here evaluate Matterport using *teacher forcing*.

	Model	Matterport↑ (%) Seen	Unseen
Pub.	Full Model	38.6	21.8
	Baseline	15.9	16.3
Uni	$\mathcal{A}$	4.1	3.2
	$\mathcal{A}+\mathcal{V}$	**30.6**	13.3
	$\mathcal{A}+\mathcal{L}$	15.4	**13.9**
Δ	Uni – Base	**+14.7**	-2.4

Table 2: Navigation results for Matterport when trained using student forcing. **Best** unimodal: **better** than reported baseline.

	Model	$\mathbf{d_T}\downarrow$ (*m*) T_{-10}	T_{-30}	T_{-50}	$\mathbf{d_{min}}\downarrow$ (*m*) T_{-10}	T_{-30}	T_{-50}
Pub.	Full	0.971	4.17	8.83	0.291	2.43	6.45
	Baseline	1.020	4.21	8.73	0.293	2.45	6.38
Uni	$\mathcal{A}$	*__0.893__	4.53	9.56	*0.242	3.16	7.99
	$\mathcal{A}+\mathcal{V}$	*0.951	*__4.11__	†__8.83__	*0.287	__2.51__	*__6.44__
	$\mathcal{A}+\mathcal{L}$	0.987	4.64	9.51	*__0.240__	3.19	7.96
Δ	U – B	**-0.127**	**-0.10**	+0.10	**-0.053**	+0.06	+0.06

Table 3: Final distances to targets ($\mathbf{d_T}$) and minimum distance from target achieved along paths ($\mathbf{d_{min}}$) in EQA navigation. **Best** unimodal: **better** than reported baseline; *better than full model; †tied with full model.

achieves better performance than existing baselines. In Matterport, the $\mathcal{A}+\mathcal{L}$ ablation performs *better* than the Full Model in unseen environments. The diverse scenes in this simulator may render the vision signal more noisy than helpful in previously unseen environments. The $\mathcal{A}+\mathcal{V}$ model in THOR-Nav and EQA is able to latch onto dataset biases in scene structure to navigate better than chance (for IQA), and the non-hierarchical baseline (in EQA). In EQA, $\mathcal{A}+\mathcal{V}$ also outperforms the Full Model;[5] the latent information about navigation from questions may be too distant for the model to infer.

The agent with access only to its action history ($\mathcal{A}$) outperforms the baseline agent in Matterport and THOR-Nav environments, suggesting it learns navigation correlations that are not captured by simple random actions (THOR-Nav) or programmatic walks away from the starting position (Matterport). Minimal sensing (which actions are available, W) coupled with the *topological* biases in trajectories (Figure 2), help this nearly zero-input agent outperform existing baselines.[6]

Matterport Teacher vs Student forcing With teacher forcing, at each timestep the navigation agent takes the gold-standard action regardless of what action it predicted, meaning it only sees steps along gold-standard trajectories. This paradigm is used to train the navigation agent in THOR-Nav and EQA. Under student forcing, the agent samples the action to take from its predictions, and loss is computed at each time step against the action that would have put the agent on the shortest path to the goal. Thus, the agent sees more of the scene, but can take more training iterations to learn to move to the goal.

Table 2 gives the highest validation success rates across all epochs achieved in Matterport by models trained using student forcing. The unimodal ablations show that the Full Model, possibly because with more exploration and more training episodes, is better able to align the vision and language signals, enabling generalization in unseen environments that fails with teacher forcing.

EQA Navigation Variants Table 3 gives the average final distance from the target ($\mathbf{d_T}$, used as the metric in Table 1) and average minimum distance from target achieved along the path ($\mathbf{d_{min}}$) during EQA episodes for agents starting 10, 30, and 50 actions away from the target in the EQA navigation task. At 10 actions away, the unimodal ablations tend to outperform the full model on both metrics, possibly due to the shorter length of the episodes (less data to train the joint parameters). The $\mathcal{A}+\mathcal{V}$ ablation performs best among the ablations, and ties with or outperforms the Full Model in all but one setting, suggesting that the EQA Full Model is not taking advantage of language information under any variant.

4.2 Question Answering

We evaluate our ablation baselines on IQUAD V1 and EQA, reporting top-1 QA accuracy (Table 4) given gold standard navigation information as $\mathcal{V}$. These decoupled QA models do not take in a previous action, so we do not consider $\mathcal{A}$ ONLY ablations for this task.

Unimodal Performance: On IQUAD V1, due to randomization in its construction, model ab-

[5]EQA full & baseline model performances do not exactly match those in Das et al. (2018a) because we use the expanded data updated by the authors https://github.com/facebookresearch/EmbodiedQA/.

[6]This learned agent begins to relate to work in minimally sensing robotics (O'Kane and LaValle, 2006).

	Model	IQUAD V1↑ Unseen	IQUAD V1↑ Seen	EQA ↑ Unseen
Pub.	Full Model	88.3	89.3	64.0
Pub.	Baseline	41.7	41.7	19.8
Uni	$\mathcal{V}$ ONLY	**43.5**	**42.8**	44.2
Uni	$\mathcal{L}$ ONLY	41.7	41.7	**48.8**
Δ	Uni – Base	**+1.8**	**+1.1**	**+29.0**

Table 4: Top-1 QA accuracy. **Best** unimodal: **better** than reported baseline.

lations perform nearly at chance.[7] The $\mathcal{V}$ ONLY model with access to the locations of all scene objects only improves by 2% over random guessing.

For EQA, single modality models perform significantly better than the majority class baseline. The vision-only model is able to identify salient colors and basic room features that allow it to reduce the likely set of answers to an unknown question. The language only models achieve nearly 50%, suggesting that despite the entropy filtering in Das et al. (2018a) each question has one answer that is as likely as all other answers combined (e.g. 50% of the answers for *What color is the bathtub?* are *grey*, and other examples in Figure 4).

5 Related Work

Historically, semantic parsing was used to map natural language instructions to visual navigation in simulation environments (Chen and Mooney, 2011; MacMahon et al., 2006). Modern approaches use neural architectures to map natural language to the (simulated) world and execute actions (Paxton et al., 2019; Chen et al., 2018; Nguyen et al., 2018; Blukis et al., 2018; Fried et al., 2018; Mei et al., 2016). In visual question answering (VQA) (Antol et al., 2015; Hudson and Manning, 2019) and visual commonsense reasoning (VCR) (Zellers et al., 2019), input images are accompanied with natural language questions. Given the question, egocentric QA requires an agent to navigate and interact with the world to gather the relevant information to answer the question. In both cases, end-to-end neural architectures make progress on these tasks.

For language annotations, task design, difficulty, and annotator pay can introduce unintended artifacts which can be exploited by models to "cheat" on otherwise complex tasks (Glockner et al., 2018; Poliak et al., 2018). Such issues also occur in multimodal data like VQA (Goyal et al., 2018), where models can answer correctly without looking at the image. In image captioning, work has shown competitive models relying only on nearest-neighbor lookups (Devlin et al., 2015) as well as exposed misalignment between caption relevance and text-based metrics (Rohrbach et al., 2018). Our unimodal ablations of visual navigation and QA benchmarks uncover similar biases, which deep architectures are quick to exploit.

[7] Majority class and chance for IQUAD V1 both achieve 50%, 50%, 25% when conditioned on question type; our Baseline model achieves the average of these.

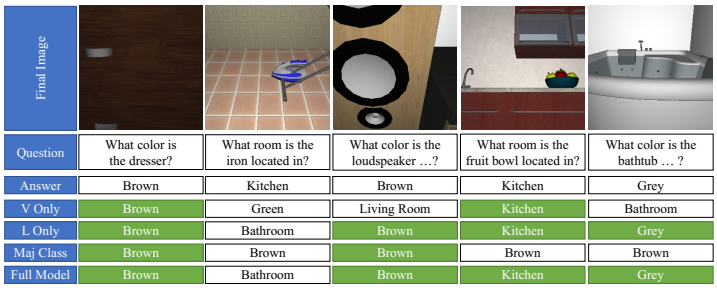


Figure 4: Qualitative results on the EQA task. The language only model can pick out the most likely answer for a question. The vision only model finds salient color and room features, but is unaware of the question.

6 Conclusions

In this work, we introduce an evaluation framework and perform the missing analysis from several new datasets. While new state-of-the-art models are being introduced for several of these domains (e.g., Matterport: (Ma et al., 2019a; Ke et al., 2019; Wang et al., 2019; Ma et al., 2019b; Tan et al., 2019; Fried et al., 2018), and EQA: (Das et al., 2018b)), they lack informative, individual unimodal ablations (i.e., ablating *both* language and vision) of the proposed models.

We find a performance gap between baselines used in or derived from the benchmarks examined in this paper and unimodal models, with unimodal models outperforming those baselines across all benchmarks. These unimodal models can even outperform their multimodal counterparts. In light of this, we recommend all future work include unimodal ablations of proposed multimodal models to vet and highlight their learned representations.

Acknowledgements

This work was supported by NSF IIS-1524371, 1703166, NRI-1637479, IIS-1338054, 1652052, ONR N00014-13-1-0720, and the DARPA CwC program through ARO (W911NF-15-1-0543).

References

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Valts Blukis, Dipendra Misra, Ross A. Knepper, and Yoav Artzi. 2018. Mapping navigation instructions to continuous control actions with position visitation prediction. In *Proceedings of the Conference on Robot Learning*.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Howard Chen, Alane Shur, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2018. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *NeurIPS Workshop on Visually Grounded Interaction and Language (ViGIL)*.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018a. Embodied Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Neural Information Processing Systems (NeurIPS)*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proc. of the Association for Computational Linguistics (ACL)*.

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

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2018. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision (IJCV)*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: a new dataset for compositional question answering over real-world images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha Srinivasa. 2019. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019a. Self-aware visual-textual co-grounded navigation agent. In *International Conference on Learning Representations (ICLR)*.

Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. 2019b. The regretful agent: Heuristic-aided navigation through progress estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proc. of the Association for Computational Linguistics (ACL)*.

Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. 2018. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jason M O'Kane and Steven M. LaValle. 2006. On comparing the power of mobile robots. In *Robotics: Science and Systems (RSS)*.

Chris Paxton, Yonatan Bisk, Jesse Thomason, Arunkumar Byravan, and Dieter Fox. 2019. Prospection: Interpretable plans from language by predicting the future. In *International Conference on Robotics and Automation (ICRA)*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Joint Conference on Lexical and Computational Semantics (StarSem)*.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.