Google Research

NYU

Carnegie Mellon University

A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning

Aishwarya Kamath*[1^], Peter Anderson*[2], Su Wang[2^], Jing Yu Koh[3^], Alex Ku[2], Austin Waters[2], Yinfei Yang[4^], Jason Baldridge[2], Zarana Parekh[2]

*Equal contributions ^Work done while at Google Research.
[1]New York University [2]Google Research [3]Carnegie Mellon University [4]Apple

PAPER ID: WED-AM-246

Contributions

- An automated instruction-trajectory augmentation pipeline, used to generate navigation graphs for Gibson environments
- A new synthetic dataset of 4.2M multilingual navigation instructions*
- Solely imitation learning agent trained without interaction with the environment
- New results demonstrating substantial gains in the state-of-the-art on the challenging RxR dataset

*Dataset available at:
github.com/google-research-datasets/RxR/tree/main/marky-mT5

Key takeaways

- Aspects of language understanding that are important for instruction-following agents appear to be hard to learn from static web data - e.g. actions/verbs (“climb the stairs”), imperatives and negations (“do not enter...”), spatial expressions (“behind you”) and temporal conditions (“walk until...”).
- Recipe that works well:
  - Large-scale imitation learning with generic architectures
  - Better and more diverse synthetic instructions
- However, aligning synthetic instructions to the target domain is essential
- Opens up doors to co-training with other image-text understanding tasks

Paper: A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning

What is Vision and Language Navigation?

- The agent is instantiated in an environment and must follow a natural language instruction *W*
- At time step *t*, the agent receives observation o_t and chooses action a_t that transitions it from state s_t to new state s_{t+1}
- Each observation is a photorealistic panoramic image (hereafter, pano) encoded as 36 image feature vectors (K=36)

$$o_t = \{I^o_{t,1}, I^o_{t,2}, ..., I^o_{t,K}\}$$

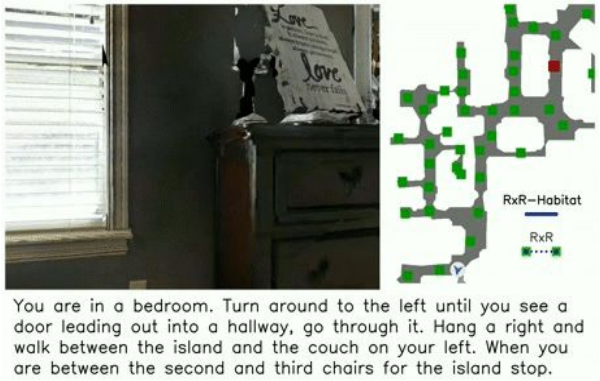
- The agent moves by choosing an action from a set of candidates (J is variable)

$$\mathcal{A}_t = \{I^a_{t,1}, I^a_{t,2}, ..., I^a_{t,J}\}$$

Language-Guided Navigation Agents

- Developing robots that follow human instructions is a long-term, formidable challenge.
- Vision-and-Language Navigation is an ideal test bed for studying instruction following:
  - Requires grounding of language in visual perceptions and actions
  - Navigation can be simulated photo-realistically at scale
  - Evaluation is straightforward - did the agent reach the goal?

RxR dataset / Habitat simulator



R2R dataset / Matterport3D simulator

Prior SOTA VLN agents

- Require interacting with the environment during training
  - Slows down training
  - Require complex training set-up
  - Difficult to scale up
- Focus on training on a small set of environments (61 from Matterport 3D)
  - low diversity
  - poor generalization

Can scaling up pre-training using image-text data solve the problem?

- Pretraining on large image-text and text-only datasets has been thoroughly explored, but improvements are limited.

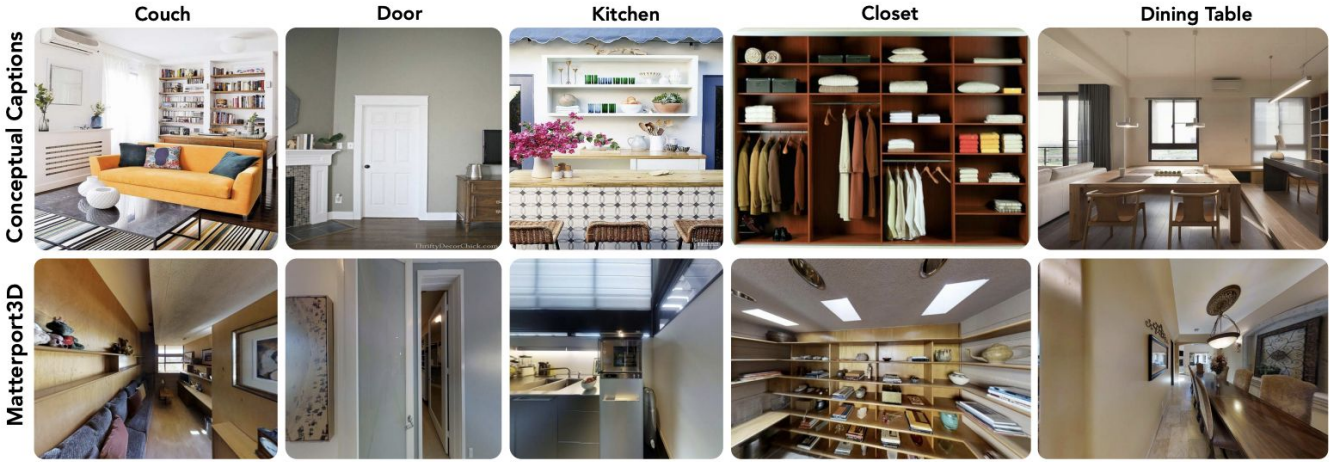


Figure from: Majumdar et al, “Improving Vision-and-Language Navigation with Image-Text Pairs from the Web”

Can scaling up pre-training using image-text data solve the problem?

- Pretraining on large image-text and text-only datasets has been thoroughly explored, but improvements are limited.
- Generic image-text data doesn't usually have
  - *Allocentric and egocentric spatial expressions*

near a grey console table **behind you**

Can scaling up pre-training using image-text data solve the problem?

- Pretraining on large image-text and text-only datasets has been thoroughly explored, but improvements are limited.
- Generic image-text data doesn't usually have
  - Allocentric and egocentric spatial expressions
  - *Imperatives and negations*

do not enter the room in front

Can scaling up pre-training using image-text data solve the problem?

- Pretraining on large image-text and text-only datasets has been thoroughly explored, but improvements are limited.
- Generic image-text data doesn't usually have
  - Allocentric and egocentric spatial expressions
  - Imperatives and negations
  - *Temporal conditions*

walk until you see an entrance on your left

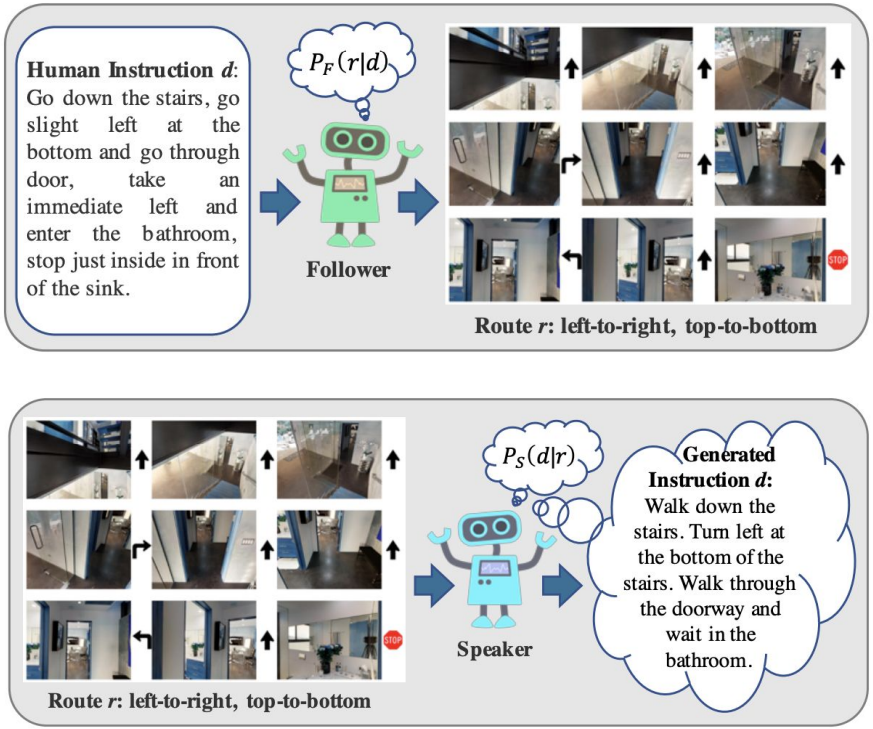
Can scaling up pre-training using image-text data solve the problem?

- Pretraining on large image-text and text-only datasets has been thoroughly explored, but improvements are limited.
- Generic image-text data doesn't usually have
  - Allocentric and egocentric spatial expressions
  - Imperatives and negations
  - Temporal conditions
- Text only datasets contain such language, but it is meaningless without sensorimotor context

Can scaling up pre-training using image-text data solve the problem?

- Pretraining on large image-text and text-only datasets has been thoroughly explored, but improvements are limited.
- Generic image-text data doesn't usually have
  - Allocentric and egocentric spatial expressions
  - Imperatives and negations
  - Temporal conditions
- Text only datasets contain such language, but is meaningless without sensorimotor context

(Answer: Not if you do it with generic image-text data)



Synthetic instruction generation

Speaker-follower model:

Data augmentation where speaker helps follower by synthesizing additional route-instruction pairs to expand the limited training data.

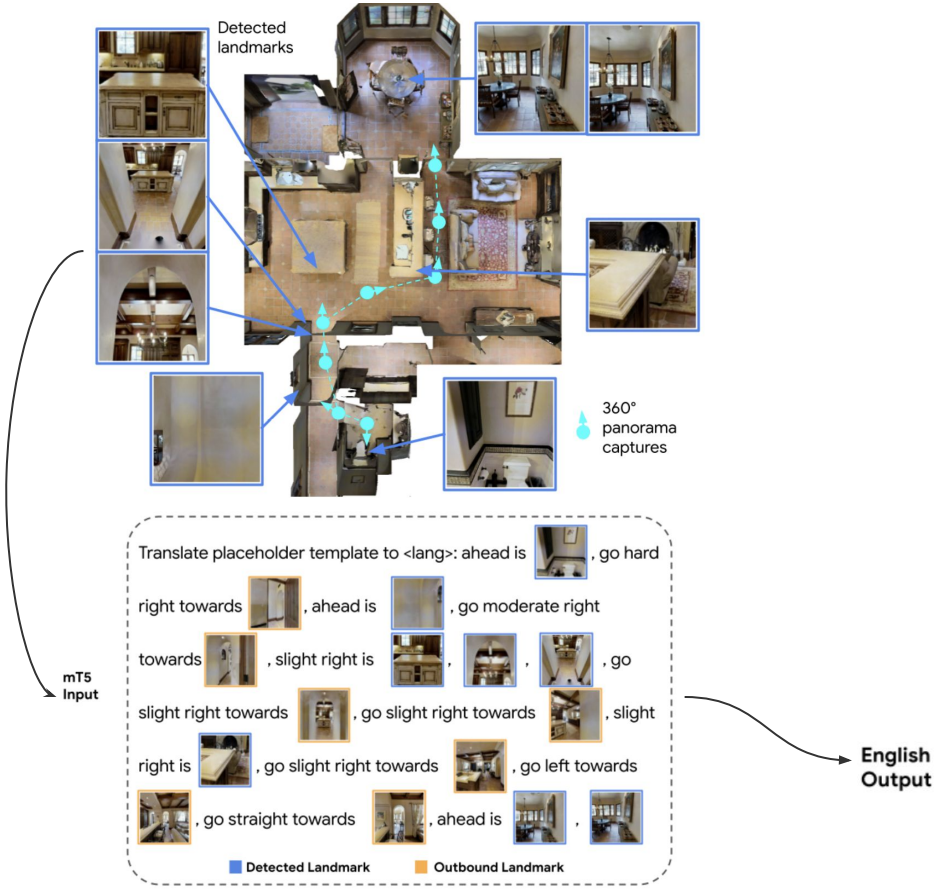
Figure from Fried, Daniel, et al. "Speaker-follower models for vision-and-language navigation." NeurIPS 2018.

Not much progress on using data augmentation since 2018…

Possible reasons:

- Issues with quality of instructions (non-grounded and/or bad followability)
  - Human wayfinding success rates on R2R are 36% for Speaker-Follower generated instructions*
- Still limited in size (sticks to Matterport3D environments)
- RL based training makes it unclear how to efficiently use more data, even if available

*Zhao et al., 'On the Evaluation of Vision-and-Language Navigation Instructions', EACL 2021.



Building on top of MARKY

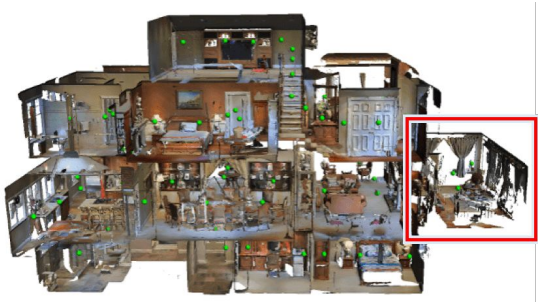
- MARKY-mT5*: A first stage detects landmarks and a second stage generates instructions conditioned on these landmarks.
- Instructions are more *fine-grained and grounded*
- On R2R *71% success rate (SR) vs. 75% for human instructions*, 42% for previous

*Wang et al., 'Less is More: Generating Grounded Navigation Instructions from Landmarks', CVPR 2022.

Training Environments

Matterport3D (Chang et al. 3DV 2017)

- 90 buildings
- We sample 330K paths, and use Marky to annotate each path with instructions in en/hi/te (1M total)



Gibson (Xia et al. CVPR 2018)

- 572 buildings
- We create navigation graphs to indicate navigable paths between panoramas
- Using Marky we annotate 1.1M paths with 3.2M instructions in en/hi/te

MARVAL - Maximum Augmentation Regime for Vision And Language navigation

- Issues with quality of instructions (non-grounded, bad followability, etc)

MARVAL: Marky based instruction generation

Better instructions = Better follower ?

MARVAL - Maximum Augmentation Regime for Vision And Language navigation

- Need a training pipeline that can be easily scaled up - existing RL based approaches are too slow to train at scale.

MARVAL: Can we get rid of RL and only use imitation learning to train efficiently at scale?

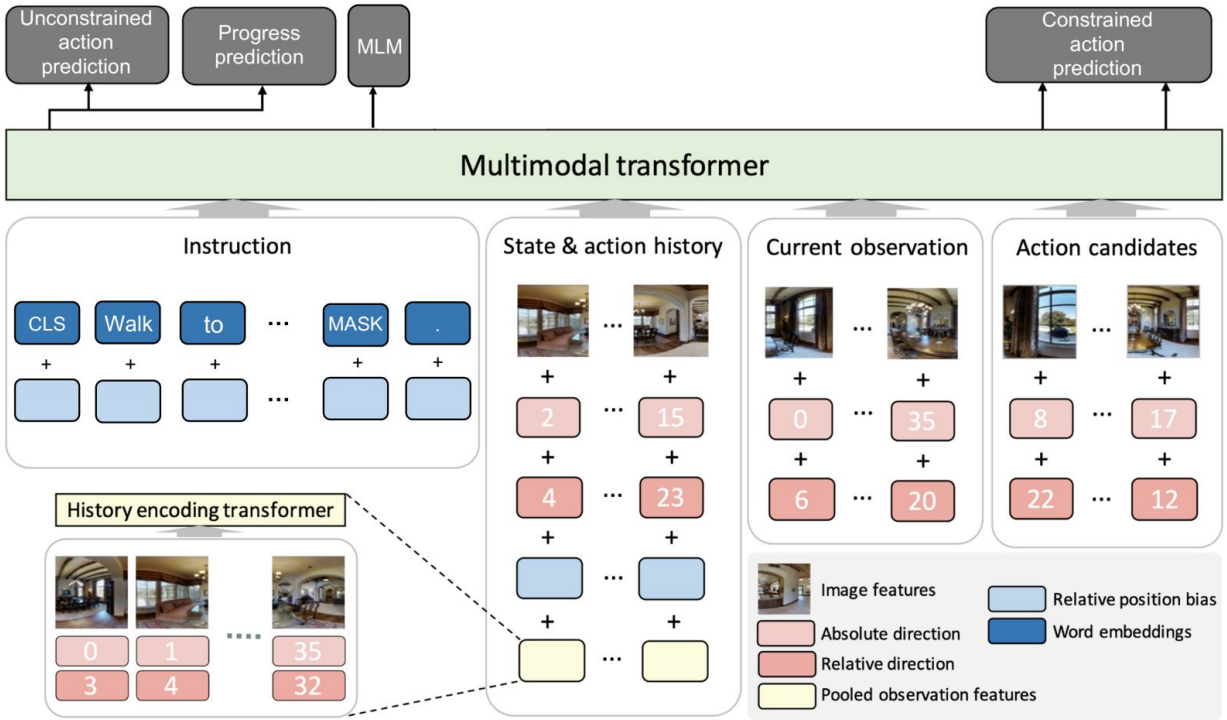
MARVAL - Maximum Augmentation Regime for Vision And Language navigation

- Generic large scale image-text data don’t contain navigation specific language & only exists on small number of environments from Matterport3D.

MARVAL: Can we use a grounded instruction generator **and** utilize novel environments to get a more diverse dataset?

MARVAL agent

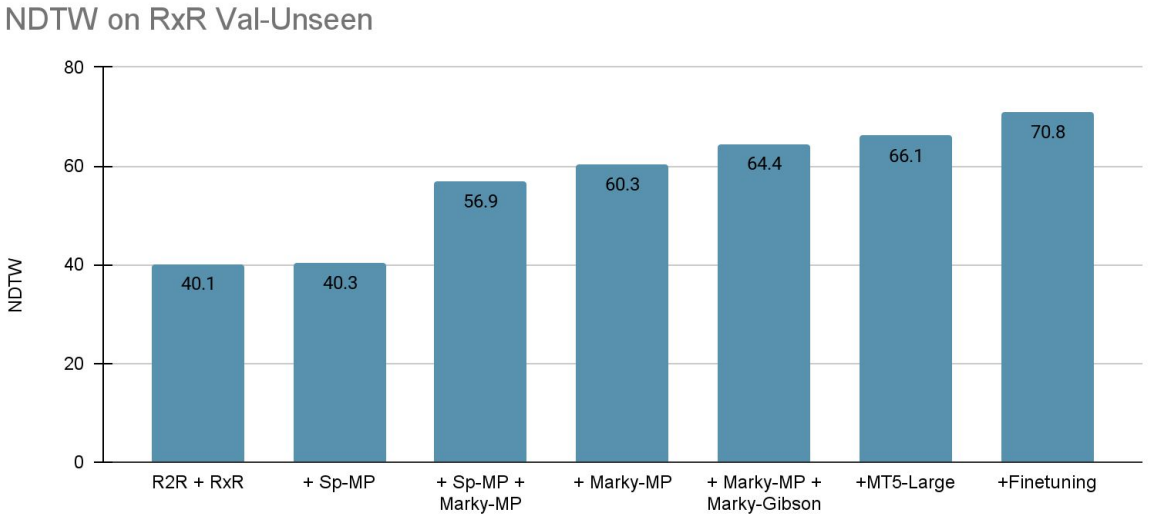
- Generic agent based on an mT5 (multilingual T5) encoder
- For high throughput, we train *only with Imitation Learning* (BC and DAGGER)
- Pretraining is performed on all available data.
- Finetuning is on human-annotated data only and the MLM objective is dropped.



What works?

On RxR Val-Unseen, pretraining with synthetic instructions provides large gains over pretraining with just the human-annotated RxR and R2R datasets.

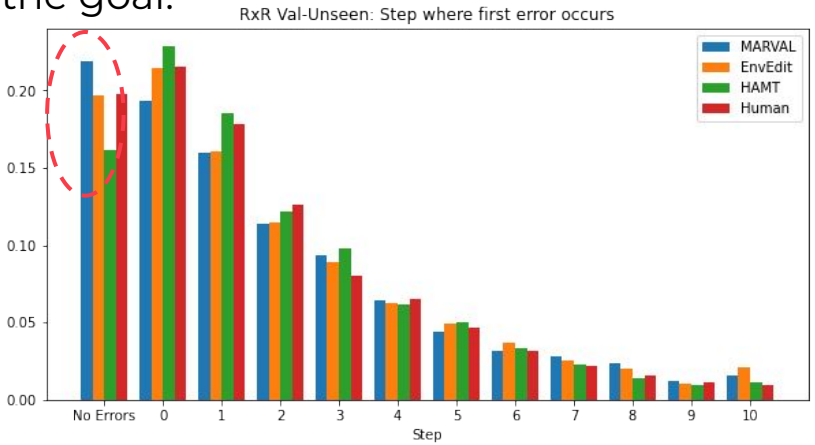
Switching from mT5-base to mT5-large and finetuning is also beneficial.

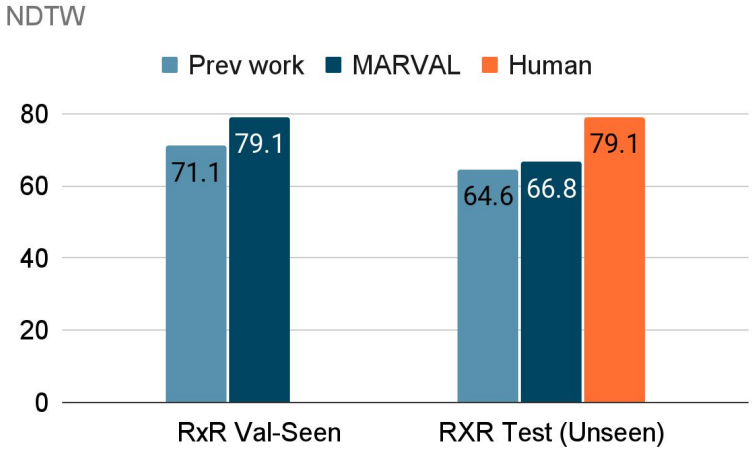


Helping the model recover from mistakes

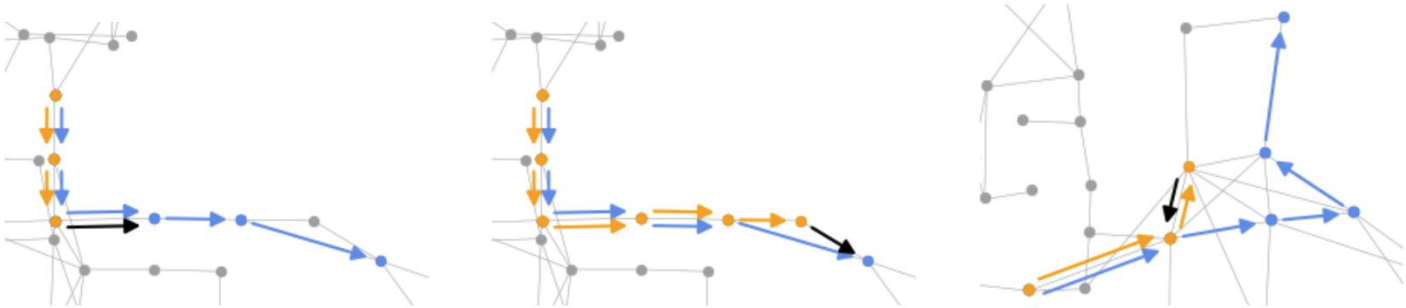
Surprisingly, MARVAL makes less mistakes (produces more perfect trajectories) than human wayfinders... but human Success Rate at reaching the goal is much higher.

Error recovery is the difference, and a possible focus for future improvement. Humans make mistakes along the way but almost always recover to still reach the goal.





Helping the model recover from mistakes



(a) Agent is on the GT trajectory: Expert selects the next action in the GT trajectory.

(b) Agent is off the GT trajectory; GT trajectory is the shortest-path from start to goal: Expert action is the first step in the recalculated shortest-path to the goal.

(c) Agent is off the GT trajectory; GT trajectory is not a shortest-path: Expert takes the shortest path back to the closest node in the GT trajectory.

Figure 3. Calculation of the DAGGER expert action (black) given the ground-truth (GT) trajectory (blue) and an agent trajectory (yellow).

Final results

	Val-Seen				Val-Unseen				Test (Unseen)			
Agent	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**NDTW**	**SDTW**
LSTM [30]	10.7	25.2	42.2	20.7	10.9	22.8	38.9	18.2	12.0	21.0	36.8	16.9
EnvDrop+ [58]	-	-	-	-	-	42.6	55.7	-	-	38.3	51.1	32.4
CLEAR-C [33]	-	-	-	-	-	-	-	-	-	40.3	53.7	34.9
HAMT [10]	-	59.4	65.3	50.9	-	56.5	63.1	48.3	6.2	53.1	59.9	45.2
EnvEdit* [34]	-	67.2	71.1	58.5	-	62.8	68.5	54.6	**5.1**	60.4	64.6	51.8
MARVAL (Pretrained)	3.62	72.7	77.0	65.9	5.56	59.4	67.0	52.7	-	-	-	-
MARVAL (Finetuned-BC)	3.25	75.4	79.0	68.7	4.80	63.7	70.6	56.9	-	-	-	-
MARVAL (DAGGER)	**3.01**	**75.9**	**79.1**	**68.8**	**4.49**	**64.8**	**70.8**	**57.5**	5.5	**60.7**	**66.8**	**53.5**
MARVAL (Pre-Explore)†	3.33	73.7	77.6	66.6	4.19	66.5	72.2	59.1	5.2	61.8	68.6	54.8
Human [30]	-	-	-	-	-	-	-	-	0.9	93.9	79.5	76.9

*Results from an ensemble of three agents.

Table 3. Results on RxR. Our MARVAL agent trained with imitation learning – behavioral cloning (BC) or DAGGER – outperforms all existing RL agents. Pre-Exploration in the eval environments († a form of privileged access, but still without human annotations) can provide a further boost.

Final results

	Val-Seen				Val-Unseen				Test (Unseen)			
Agent	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**NDTW**	**SDTW**
LSTM [30]	10.7	25.2	42.2	20.7	10.9	22.8	38.9	18.2	12.0	21.0	36.8	16.9
EnvDrop+ [58]	-	-	-	-	-	42.6	55.7	-	-	38.3	51.1	32.4
CLEAR-C [33]	-	-	-	-	-	-	-	-	-	40.3	53.7	34.9
HAMT [10]	-	59.4	65.3	50.9	-	56.5	63.1	48.3	6.2	53.1	59.9	45.2
EnvEdit* [34]	-	67.2	71.1	58.5	-	62.8	68.5	54.6	**5.1**	60.4	64.6	51.8
MARVAL (Pretrained)	3.62	72.7	77.0	65.9	5.56	59.4	67.0	52.7	-	-	-	-
MARVAL (Finetuned-BC)	3.25	75.4	79.0	68.7	4.80	63.7	70.6	56.9	-	-	-	-
MARVAL (DAGGER)	**3.01**	**75.9**	**79.1**	**68.8**	**4.49**	**64.8**	**70.8**	**57.5**	5.5	**60.7**	**66.8**	**53.5**
MARVAL (Pre-Explore)†	3.33	73.7	77.6	66.6	4.19	66.5	72.2	59.1	5.2	61.8	68.6	54.8
Human [30]	-	-	-	-	-	-	-	-	0.9	93.9	79.5	76.9

*Results from an ensemble of three agents.

Table 3. Results on RxR. Our MARVAL agent trained with imitation learning – behavioral cloning (BC) or DAGGER – outperforms all existing RL agents. Pre-Exploration in the eval environments († a form of privileged access, but still without human annotations) can provide a further boost.

Final results

	Val-Seen				Val-Unseen				Test (Unseen)			
Agent	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**NDTW**	**SDTW**
LSTM [30]	10.7	25.2	42.2	20.7	10.9	22.8	38.9	18.2	12.0	21.0	36.8	16.9
EnvDrop+ [58]	-	-	-	-	-	42.6	55.7	-	-	38.3	51.1	32.4
CLEAR-C [33]	-	-	-	-	-	-	-	-	-	40.3	53.7	34.9
HAMT [10]	-	59.4	65.3	50.9	-	56.5	63.1	48.3	6.2	53.1	59.9	45.2
EnvEdit* [34]	-	67.2	71.1	58.5	-	62.8	68.5	54.6	**5.1**	60.4	64.6	51.8
MARVAL (Pretrained)	3.62	72.7	77.0	65.9	5.56	59.4	67.0	52.7	-	-	-	-
MARVAL (Finetuned-BC)	3.25	75.4	79.0	68.7	4.80	63.7	70.6	56.9	-	-	-	-
MARVAL (DAGGER)	**3.01**	**75.9**	**79.1**	**68.8**	**4.49**	**64.8**	**70.8**	**57.5**	5.5	**60.7**	**66.8**	**53.5**
MARVAL (Pre-Explore)†	3.33	73.7	77.6	66.6	4.19	66.5	72.2	59.1	5.2	61.8	68.6	54.8
Human [30]	-	-	-	-	-	-	-	-	0.9	93.9	79.5	76.9

*Results from an ensemble of three agents.

Table 3. Results on RxR. Our MARVAL agent trained with imitation learning – behavioral cloning (BC) or DAGGER – outperforms all existing RL agents. Pre-Exploration in the eval environments († a form of privileged access, but still without human annotations) can provide a further boost.

Final results

	Val-Seen				Val-Unseen				Test (Unseen)			
Agent	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**NDTW**	**SDTW**	**NE**	**SR**	**NDTW**	**SDTW**
LSTM [30]	10.7	25.2	42.2	20.7	10.9	22.8	38.9	18.2	12.0	21.0	36.8	16.9
EnvDrop+ [58]	-	-	-	-	-	42.6	55.7	-	-	38.3	51.1	32.4
CLEAR-C [33]	-	-	-	-	-	-	-	-	-	40.3	53.7	34.9
HAMT [10]	-	59.4	65.3	50.9	-	56.5	63.1	48.3	6.2	53.1	59.9	45.2
EnvEdit* [34]	-	67.2	71.1	58.5	-	62.8	68.5	54.6	**5.1**	60.4	64.6	51.8
MARVAL (Pretrained)	3.62	72.7	77.0	65.9	5.56	59.4	67.0	52.7	-	-	-	-
MARVAL (Finetuned-BC)	3.25	75.4	79.0	68.7	4.80	63.7	70.6	56.9	-	-	-	-
MARVAL (DAGGER)	**3.01**	**75.9**	**79.1**	**68.8**	**4.49**	**64.8**	**70.8**	**57.5**	5.5	**60.7**	**66.8**	**53.5**
MARVAL (Pre-Explore)†	3.33	73.7	77.6	66.6	4.19	66.5	72.2	59.1	5.2	61.8	68.6	54.8
Human [30]	-	-	-	-	-	-	-	-	0.9	93.9	79.5	76.9

*Results from an ensemble of three agents.

Table 3. Results on RxR. Our MARVAL agent trained with imitation learning – behavioral cloning (BC) or DAGGER – outperforms all existing RL agents. Pre-Exploration in the eval environments († a form of privileged access, but still without human annotations) can provide a further boost.

Contributions

- An automated instruction-trajectory augmentation pipeline, used to generate navigation graphs for Gibson environments
- A new synthetic dataset of 4.2M multilingual navigation instructions*
- Solely imitation learning agent trained without interaction with the environment
- New results demonstrating substantial gains in the state-of-the-art on the challenging RxR dataset

*Dataset available at:
github.com/google-research-datasets/RxR/tree/main/marky-mT5

Key takeaways

- Aspects of language understanding that are important for instruction-following agents appear to be hard to learn from static web data - e.g. actions/verbs (“climb the stairs”), imperatives and negations (“do not enter...”), spatial expressions (“behind you”) and temporal conditions (“walk until...”).
- Recipe that works well:
  - Large-scale imitation learning with generic architectures
  - Better and more diverse synthetic instructions
- However, aligning synthetic instructions to the target domain is essential
- Opens up doors to co-training with other image-text understanding tasks

Paper: A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning

Thank you!