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

Jesse Thomason, Daniel Gordon, and Yonatan Bisk
University of Washington

Unimodal Baselines

We demonstrate the strength of unimodal baselines in multimodal domains. We argue that unimodal approaches better capture and reflect dataset biases than random or majority class baselines, and therefore provide an important comparison when assessing the performance of multimodal techniques.

Evaluation Framework

We ablate benchmark models from:

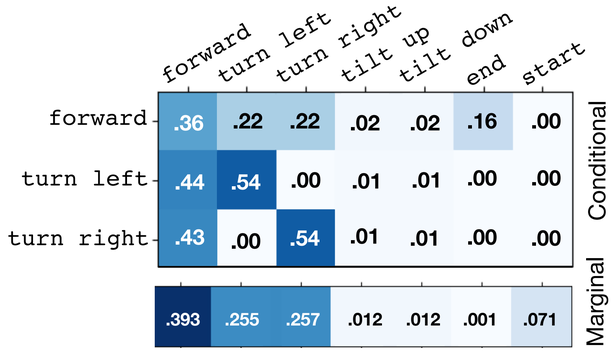
1. **Matterport** Room-2-Room Navigation – (*Anderson et al., CVPR'18*);
2. **THOR** Interactive Question Answering – (*Gordon et al., CVPR'18*);
3. **EQA** Embodied Question Answering – (*Das et al., CVPR'18*).

We define three ablations:

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

with $\mathcal{A}$ction inputs, $\mathcal{V}$ision inputs, and $\mathcal{L}$anguage inputs.
At each timestep an agent receives an observation and produces an action.

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



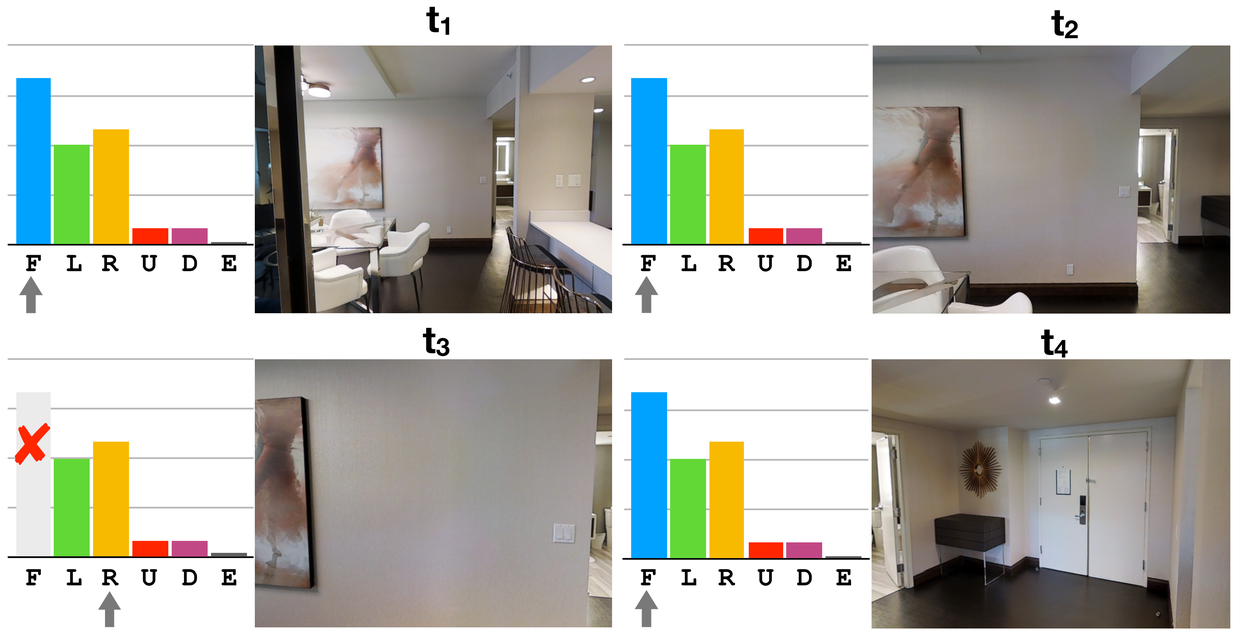
$P(act = col | prev = row)$ and marginal action distributions in Matterport data reveal memorizable peakiness.

Relevant visual observations are made after navigating to an implicit goal point in QA tasks.

Punchline

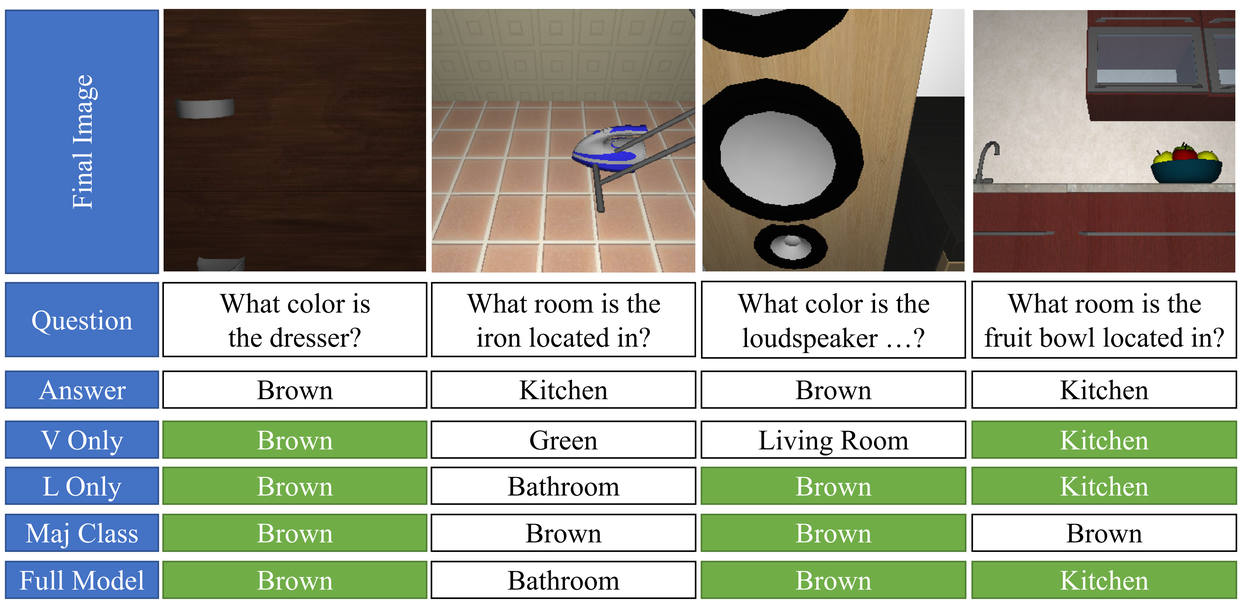
Language-only and vision-only VLN and QA models outperform published baselines and even beat their multi-modal counterparts!

Recommendation for Best Practices: While many papers ablate either language *or* vision, researchers should ablate *both*. These unimodal baselines expose possible gains from single-modality biases in multimodal datasets irrespective of training and architecture details.



Actions: Forward, turn Left & Right, tilt Up & Down, End

In the Matterport Room-2-Room task, navigating without vision can lead 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, rendering the visual context for the command unnecessary.



Qualitative results on the EQA task illuminate some unimodal biases in the data. The language only model can pick out the most likely answer for a given question without visual context. The vision only model finds and reports salient color and room feature as answers without being aware of the question.

Unimodal Evaluation

Best unimodal in **bold**, blue indicates better than baseline; and * indicates better than full model.

Visual Navigation.

		Matterport↑ (%)		IQA↑ (%)		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**

Training via behavior cloning.

		Matterport↑ (%)	
	Model	Seen	Un
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

Training via student forcing.

		$\mathbf{d_T} \downarrow$ (*m*)		
	Model	T_{-10}	T_{-30}	T_{-50}
Pub.	Full	0.971	4.17	8.83
	Baseline	1.020	4.21	8.73
Uni	$\mathcal{A}$	***0.893**	4.53	9.56
	$\mathcal{A} + \mathcal{V}$	*0.951	***4.11**	†**8.83**
	$\mathcal{A} + \mathcal{L}$	0.987	4.64	9.51
Δ Uni – Base		**-0.127**	**-0.10**	+0.10

EQA navigation final distance.

		$\mathbf{d_{min}} \downarrow$ (*m*)		
	Model	T_{-10}	T_{-30}	T_{-50}
Pub.	Full	0.291	2.43	6.45
	Baseline	0.293	2.45	6.38
Uni	$\mathcal{A}$	*0.242	3.16	7.99
	$\mathcal{A} + \mathcal{V}$	*0.287	**2.51**	***6.44**
	$\mathcal{A} + \mathcal{L}$	***0.240**	3.19	7.96
Δ Uni – Base		**-0.053**	+0.06	+0.06

EQA navigation closest distance.

Question Answering.

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

Question answering accuracy.

jdtho@cs.washington.edu
https://arxiv.org/abs/1811.00613