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 $\mathcal{M}(V, L, a_{t-1}; W)$
- **A** is $\mathcal{M}(\theta, a_{t-1}; W)$
- **A + V** is $\mathcal{M}(\theta, \hat{V}, a_{t-1}; W)$
- **A + L** is $\mathcal{M}(\hat{L}, \hat{L}, a_{t-1}; W)$

with action inputs, vision inputs, and language inputs.

At each timestep an agent receives an observation and produces an action.

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

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.

<table>
<thead>
<tr>
<th>Action</th>
<th>Forward</th>
<th>Turn left</th>
<th>Turn right</th>
<th>Tilt up</th>
<th>Tilt down</th>
<th>Start</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(col)</td>
<td>0.30</td>
<td>0.22</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>P(row)</td>
<td>0.30</td>
<td>0.22</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
</tr>
</tbody>
</table>

**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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Seen</th>
<th>Un Seen</th>
<th>Un Un</th>
<th>Un Un</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pub.</td>
<td>Full Model</td>
<td>27.1</td>
<td>19.6</td>
<td>77.7</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>15.9</td>
<td>16.3</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{A}$</td>
<td>18.5</td>
<td>17.1</td>
<td>4.53</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{A} + V$</td>
<td>21.2</td>
<td>16.6</td>
<td>35.6</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{A} + L$</td>
<td>23.0</td>
<td>22.1</td>
<td>4.03</td>
</tr>
<tr>
<td>Delta</td>
<td>$\Delta$ Uni - Base</td>
<td>$+7.1$</td>
<td>$+5.8$</td>
<td>$+33.4$</td>
</tr>
</tbody>
</table>

Training via behavior cloning.

In QA tasks.

Question: What room is the iron located in?

**Answer:**

- **V Only**: Brown
- **L Only**: Bathroom
- **Majority Class**: Brown
- **Full Model**: Brown

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.

**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.

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.

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.

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.