

# Visual-Language Navigation

- Introduction
- Dataset & Platform
- Models
- Further Works

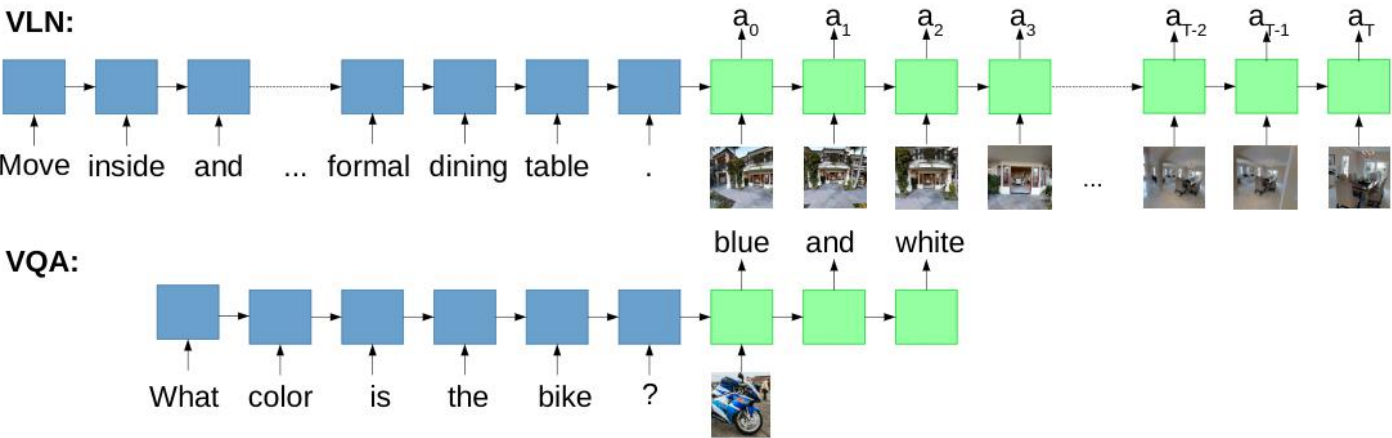
# Introduction



**Instruction:** Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

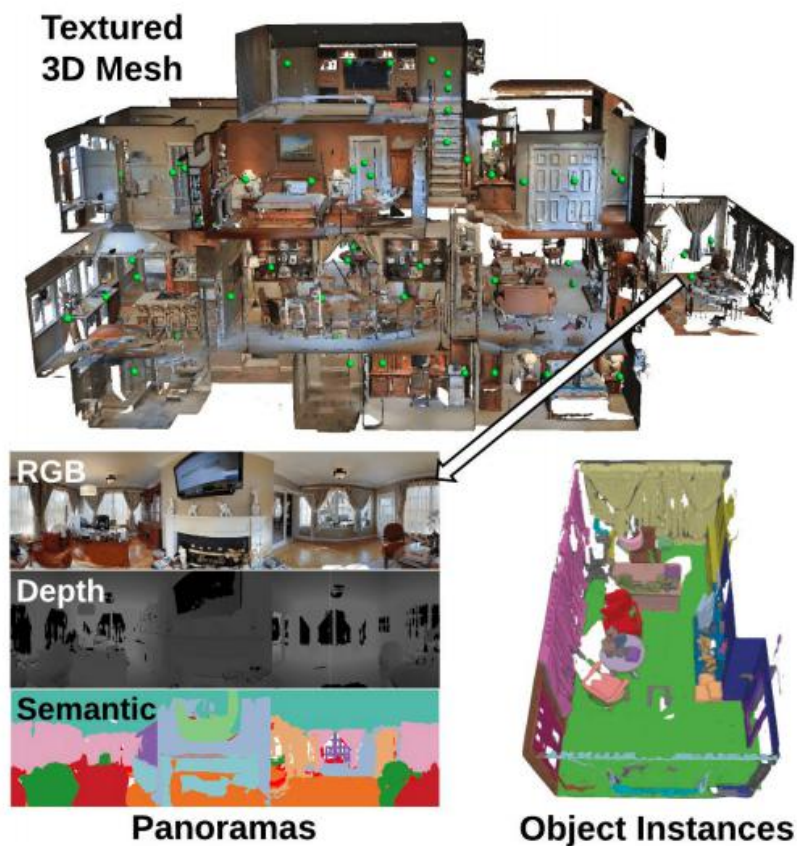


What color are her eyes?  
What is the mustache made of?



# Dataset & Platform

Matterport3D



A large RGB-D dataset of 90 building-scale scenes.



Panoramas are captured from viewpoints (green spheres) on average 2.25m apart.

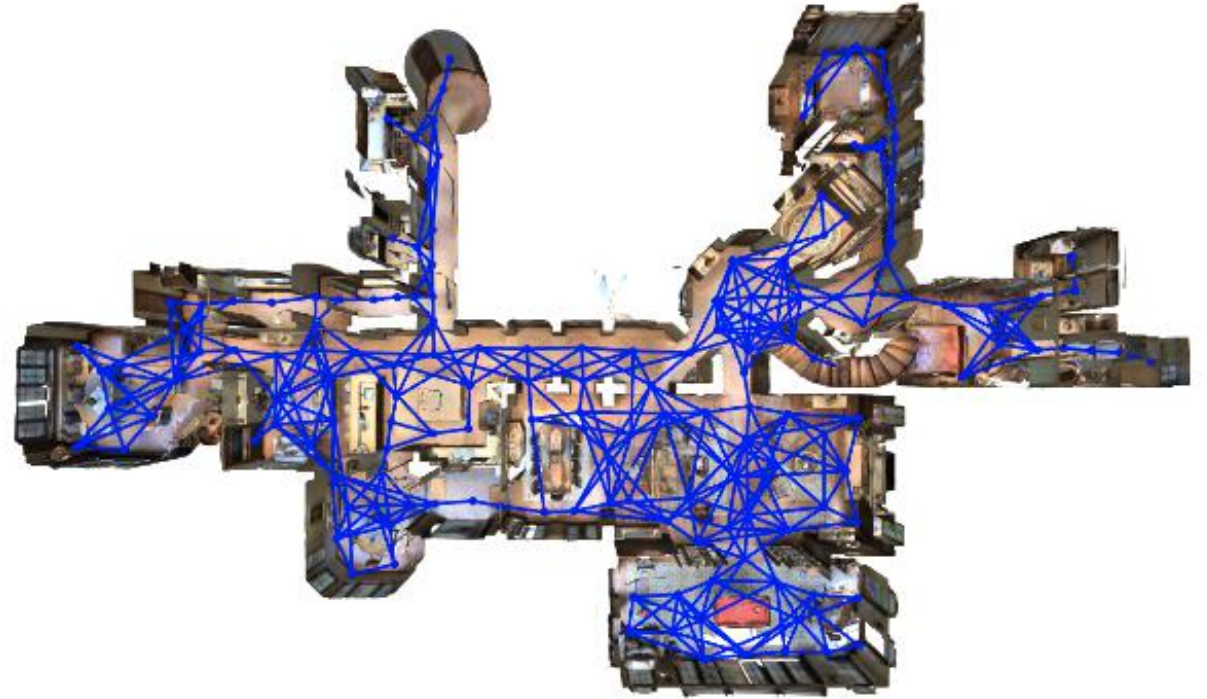


# Dataset & Platform

Matterport3D



Observations



Nav-Graph

# Dataset & Platform

Task



**Instruction:** Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

language instruction

$$\bar{X} = \langle x_1, x_2, \dots, x_L \rangle$$

agent's pose

$$s_t = \langle v_t, \varphi_t, \theta_t \rangle$$

observation

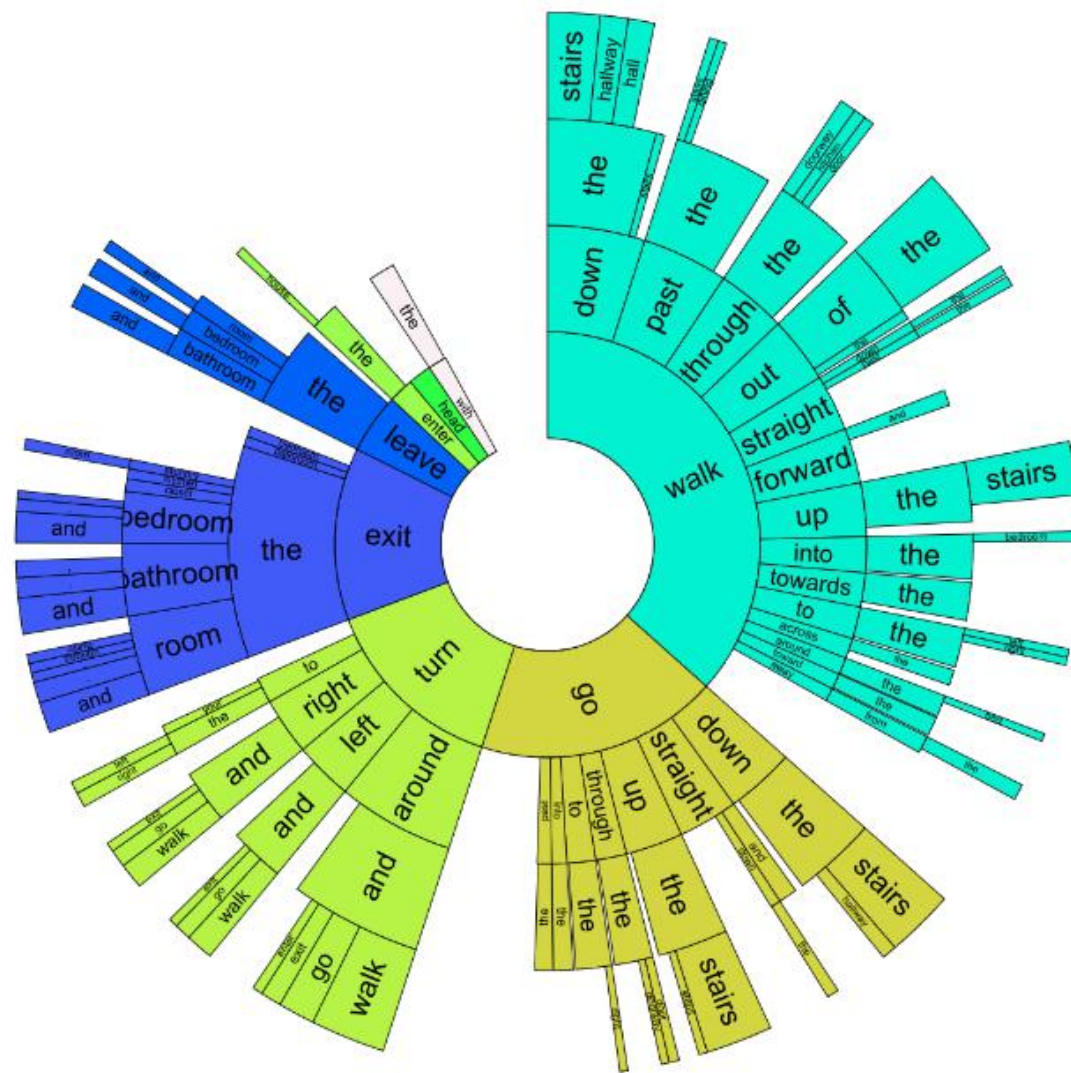
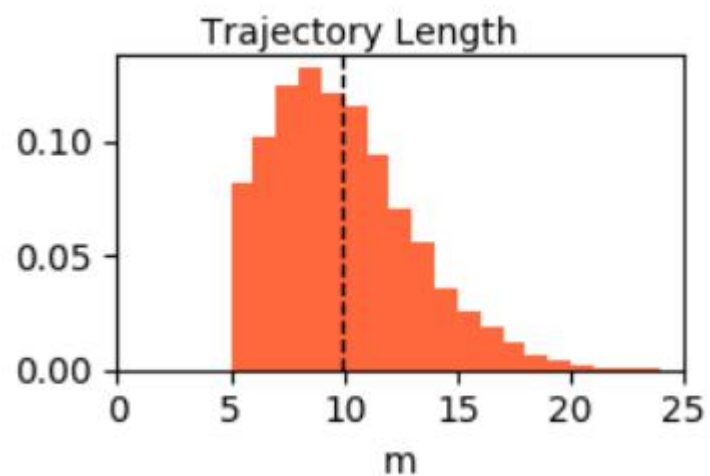
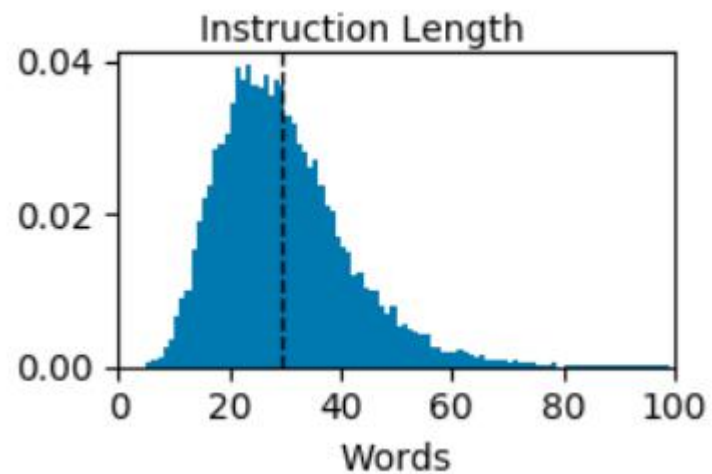
$$o_t$$

agent execute a sequence of actions

$$\langle s_0, a_0, s_1, a_1, \dots \rangle$$

## Dataset & Platform

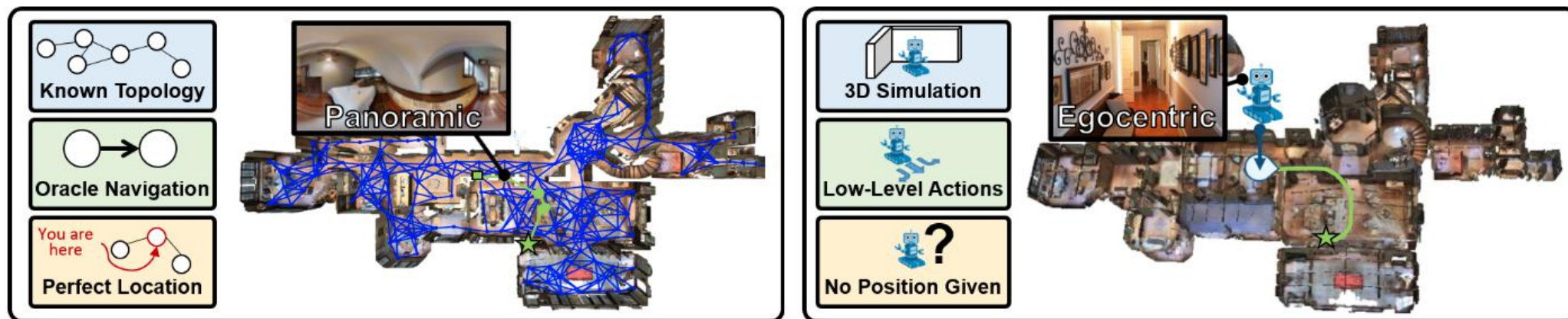
## Room-to-Room





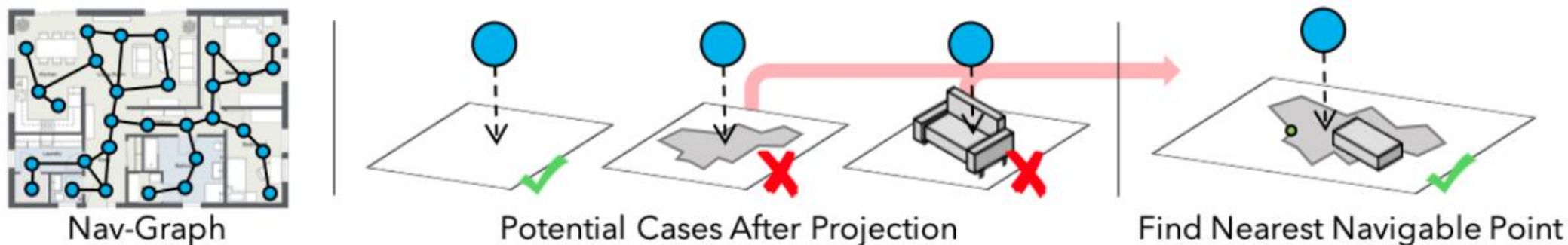
# Dataset & Platform

## Vision-and-Language Navigation in Continuous Environments



(a) Vision-and-Language Navigation (VLN)    (b) VLN in Continuous Environments (VLN-CE)

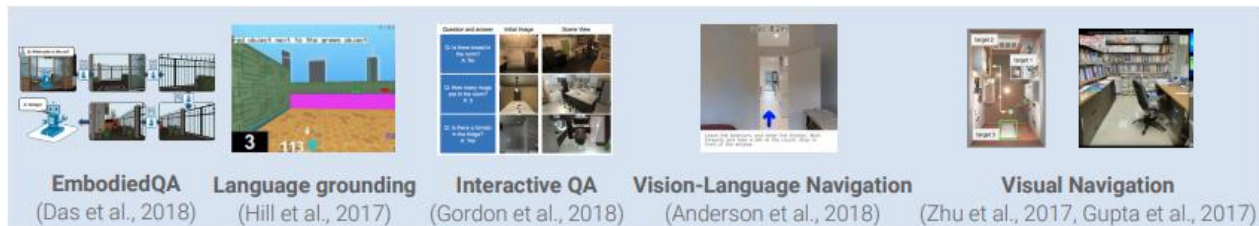
Converting Room-to-Room Trajectories to Habitat.



# Dataset & Platform

Habitat

Tasks



Simulators



Datasets



Habitat Platform

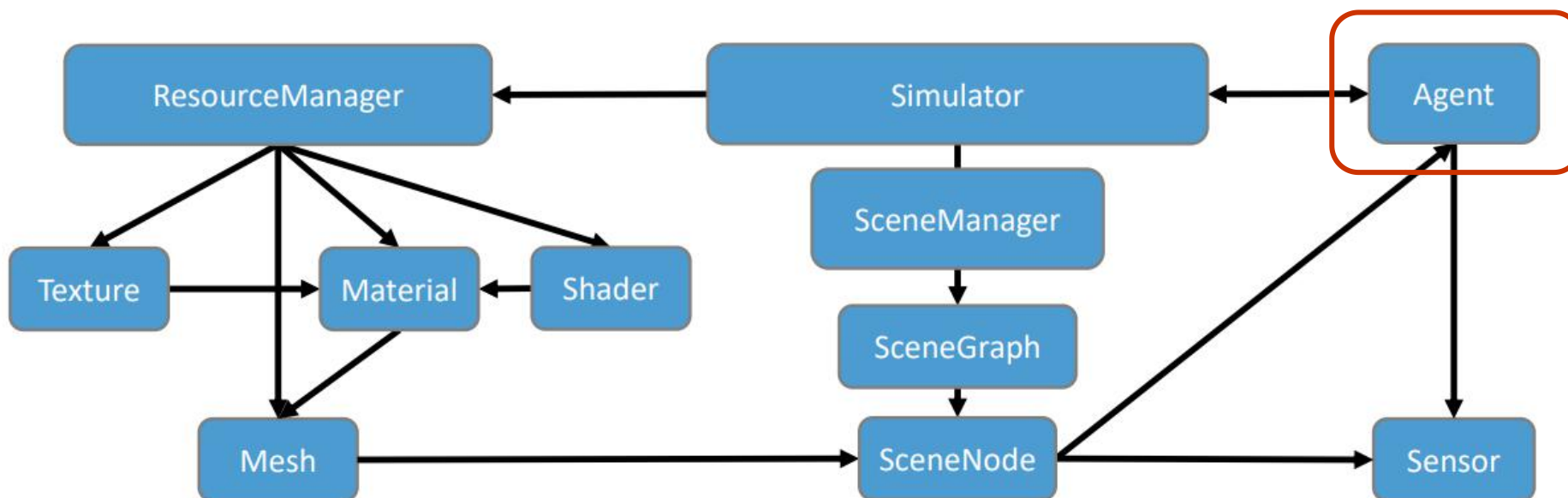
Habitat API



Habitat Sim



Generic Dataset Support



<https://aihabitat.org/>



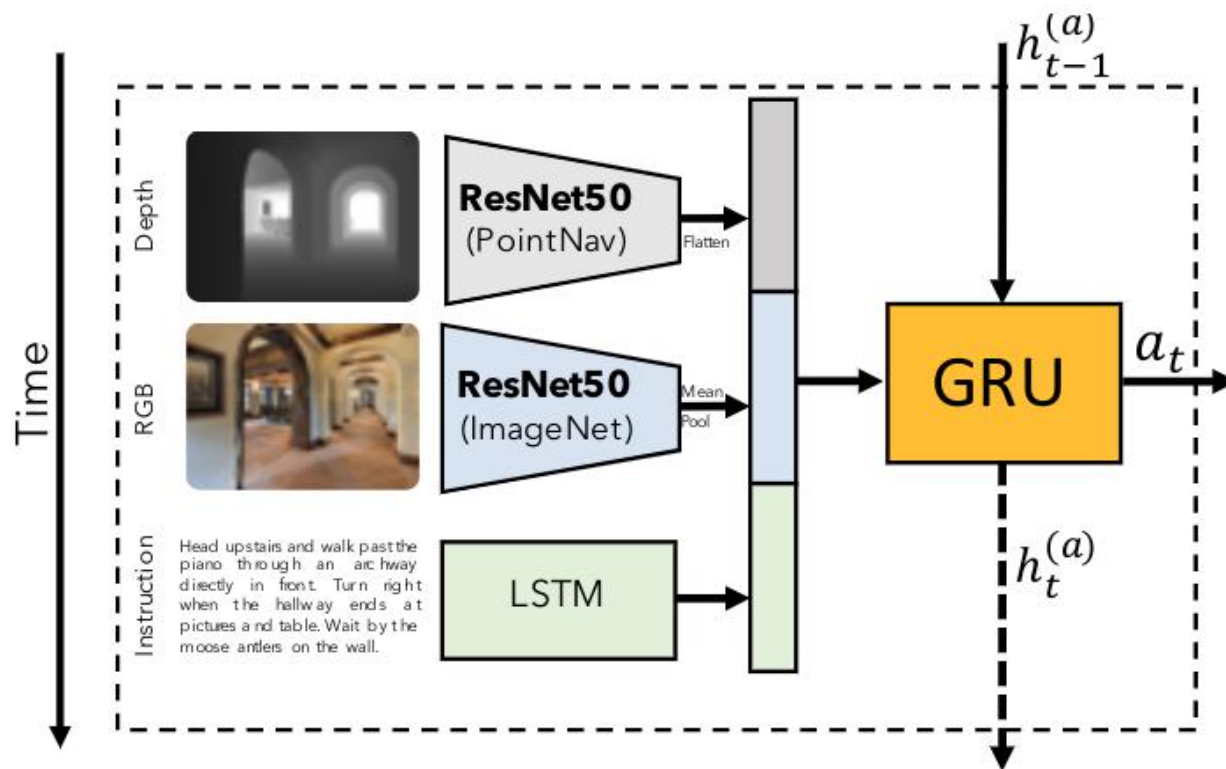
# Models

## Sequence-to-Sequence Model

$$\bar{\mathbf{v}}_t = \text{mean-pool}(\mathcal{V}_t), \quad \bar{\mathbf{d}}_t = [\mathbf{d}_1, \dots, \mathbf{d}_{wh}], \quad \mathbf{s} = \text{LSTM}(\mathbf{w}_1, \dots, \mathbf{w}_T)$$

$$\mathbf{h}_t^{(a)} = \text{GRU}\left([\bar{\mathbf{v}}_t, \bar{\mathbf{d}}_t, \mathbf{s}], \mathbf{h}_{t-1}^{(a)}\right)$$

$$a_t = \underset{a}{\operatorname{argmax}} \quad \operatorname{softmax}\left(W_a \mathbf{h}_t^{(a)} + \mathbf{b}_a\right)$$



# Models

## Cross-Modal Attention

$$\mathbf{h}_t^{(attn)} = \text{GRU} \left( [\bar{\mathbf{v}}_t, \bar{\mathbf{d}}_t, \mathbf{a}_{t-1}], \mathbf{h}_{t-1}^{(attn)} \right)$$

$$\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_T\} = \text{BiLSTM}(\mathbf{w}_1, \dots, \mathbf{w}_T)$$

$$\hat{\mathbf{s}}_t = \text{Attn}(\mathcal{S}, \mathbf{h}_t^{(attn)}), \quad \hat{\mathbf{v}}_t = \text{Attn}(\mathcal{V}_t, \hat{\mathbf{s}}_t), \quad \hat{\mathbf{d}}_t = \text{Attn}(\mathcal{D}_t, \hat{\mathbf{s}}_t)$$

$$\mathbf{h}_t^{(a)} = \text{GRU} \left( [\hat{\mathbf{s}}_t, \hat{\mathbf{v}}_t, \hat{\mathbf{d}}_t, \mathbf{a}_{t-1}, \mathbf{h}_t^{(attn)}], \mathbf{h}_{t-1}^{(a)} \right)$$

$$a_t = \underset{a}{\operatorname{argmax}} \quad \operatorname{softmax} \left( W_a \mathbf{h}_t^{(a)} + \mathbf{b}_a \right)$$

turn completely around until you face an open door with a window to the left and a patio to the right, walk forward though the door and into a dinning room, ... ..

Language Encoder

$\{\mathbf{w}_i\}_{i=1}^n$

Attention

$\mathbf{c}_t^{text}$

Action Predictor

$\mathbf{a}_t$

Attention

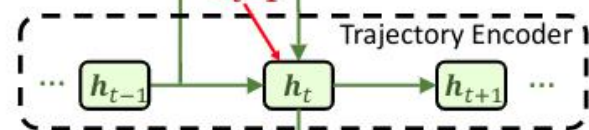


Panoramic Features

$\{\mathbf{v}_{t,j}\}_{j=1}^m$

Attention

$\mathbf{a}_{t-1}$



$\mathbf{h}_{t-1}$ ,  $\mathbf{h}_t$ ,  $\mathbf{h}_{t+1}$

$\mathbf{c}_t^{text}$

$\mathbf{c}_t^{visual}$

## Futher Works

- Instance-level information
- Structural memory
- Dynamic Nav-Graph
- ...