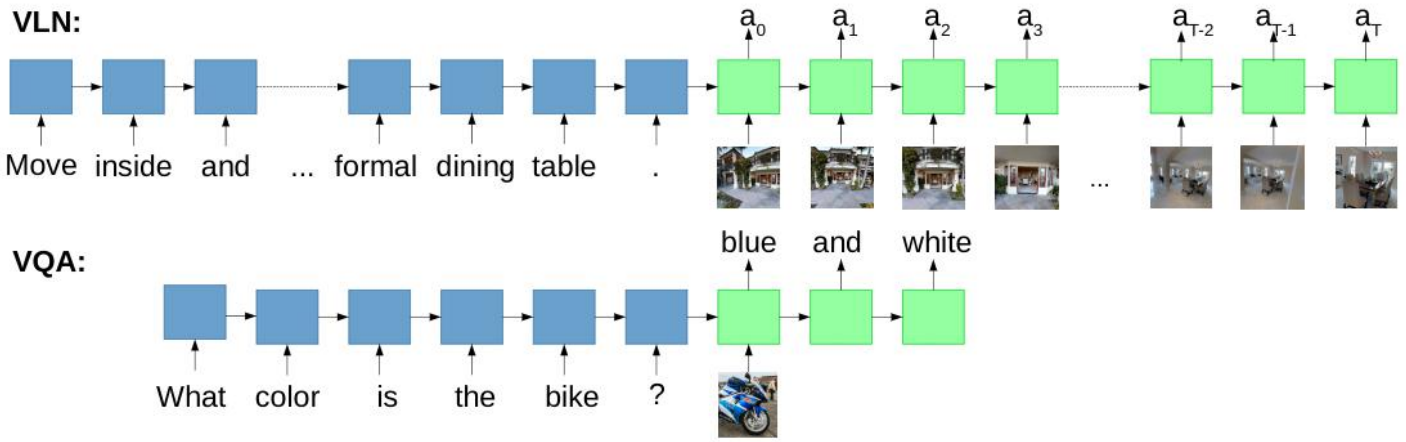
# Visual-Language Nivagation

- Introduction
- Dataset & Platform
- Models
- Futher Works

# Introduction

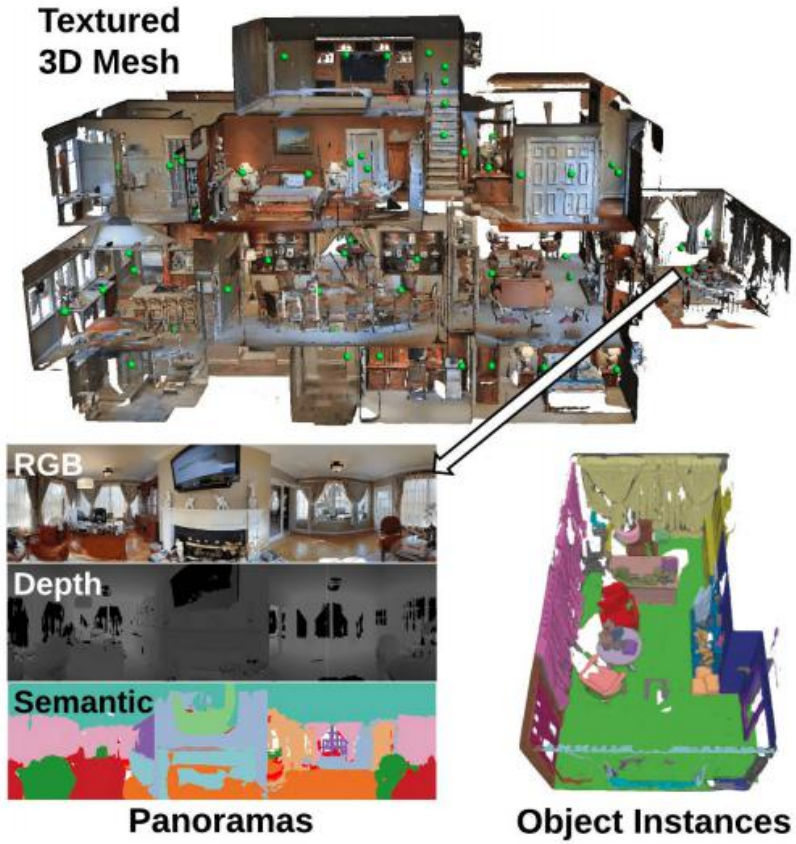
**Instruction:** Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

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



# Dataset & Platform

Matterport3D

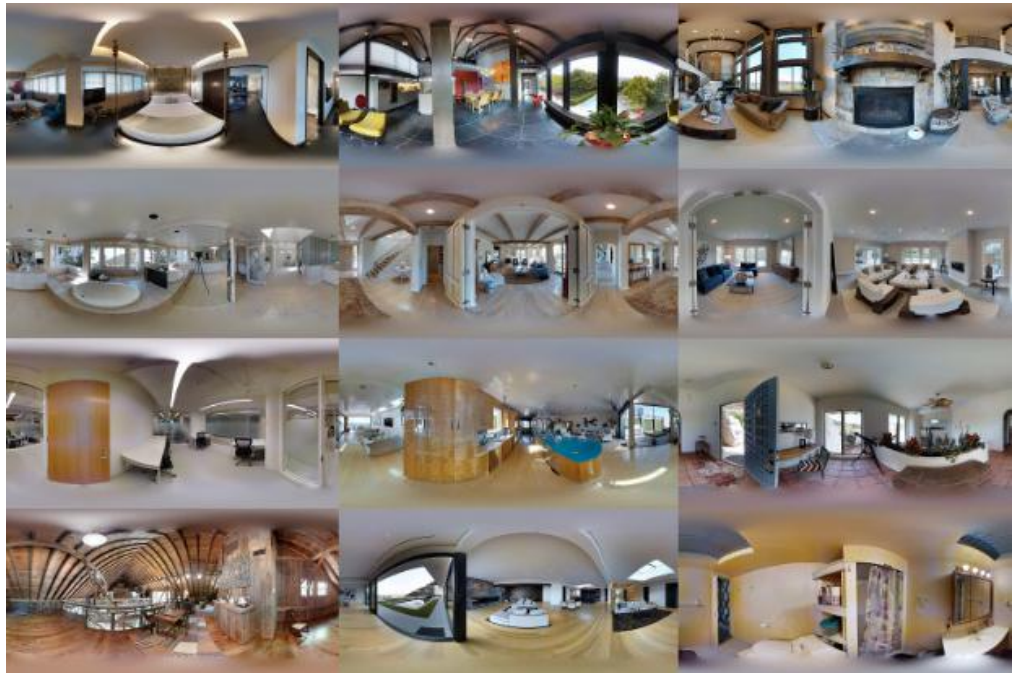



A large RGB-D dataset of 90 building-scale scenes.

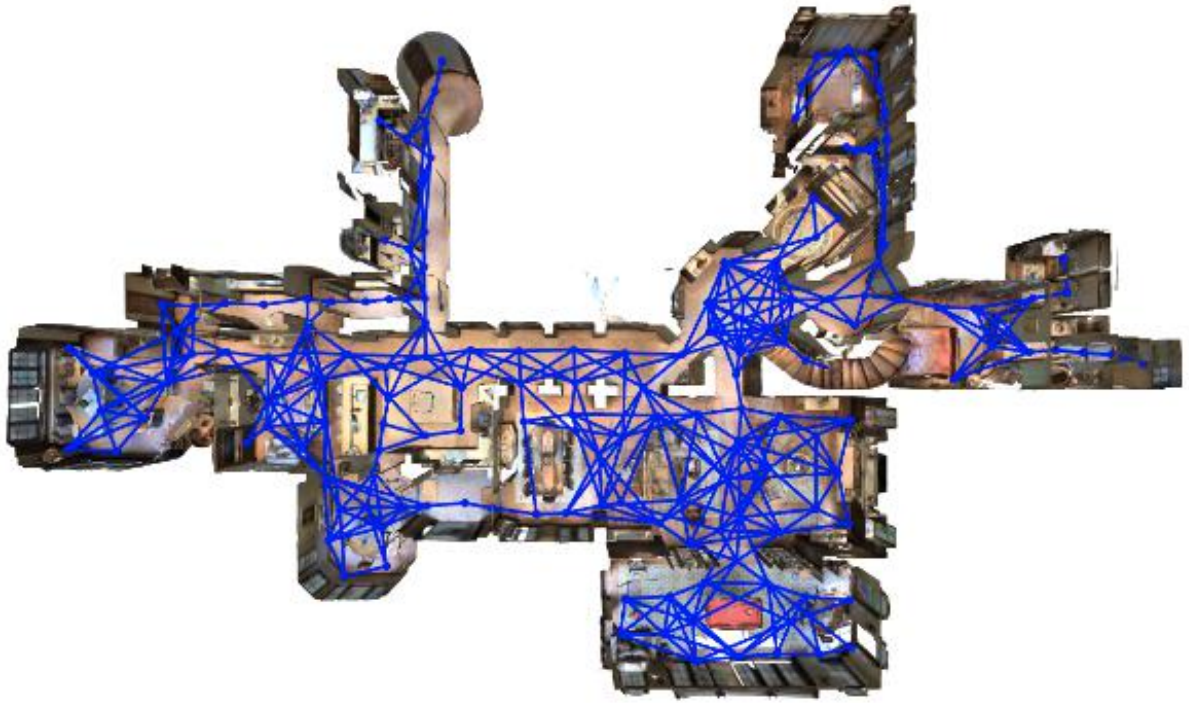
Panoramas are captured from viewpoints (green spheres) on average 2.25m apart.

# 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

$$\overline{x} =< x_1, x_2, \ldots x_L >$$

agent's pose

$$s_t =< v_t, \varphi_t, \theta_t >$$

observation

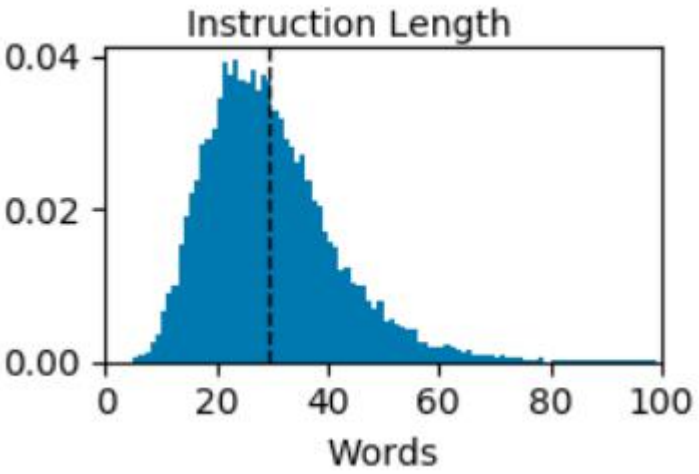
$$o_t$$

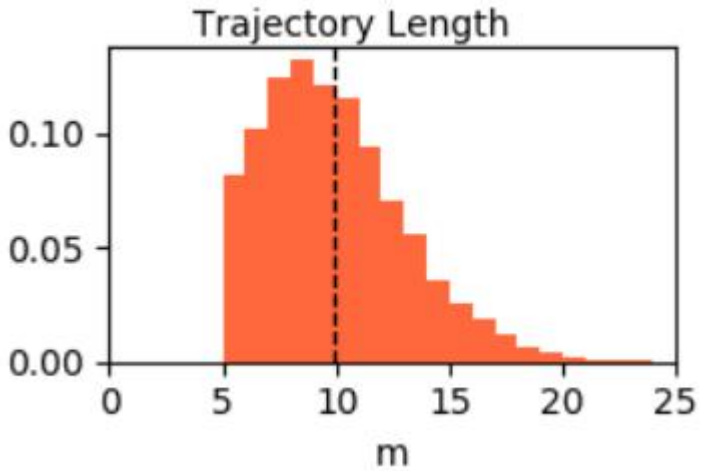
agent execute a sequence of actions

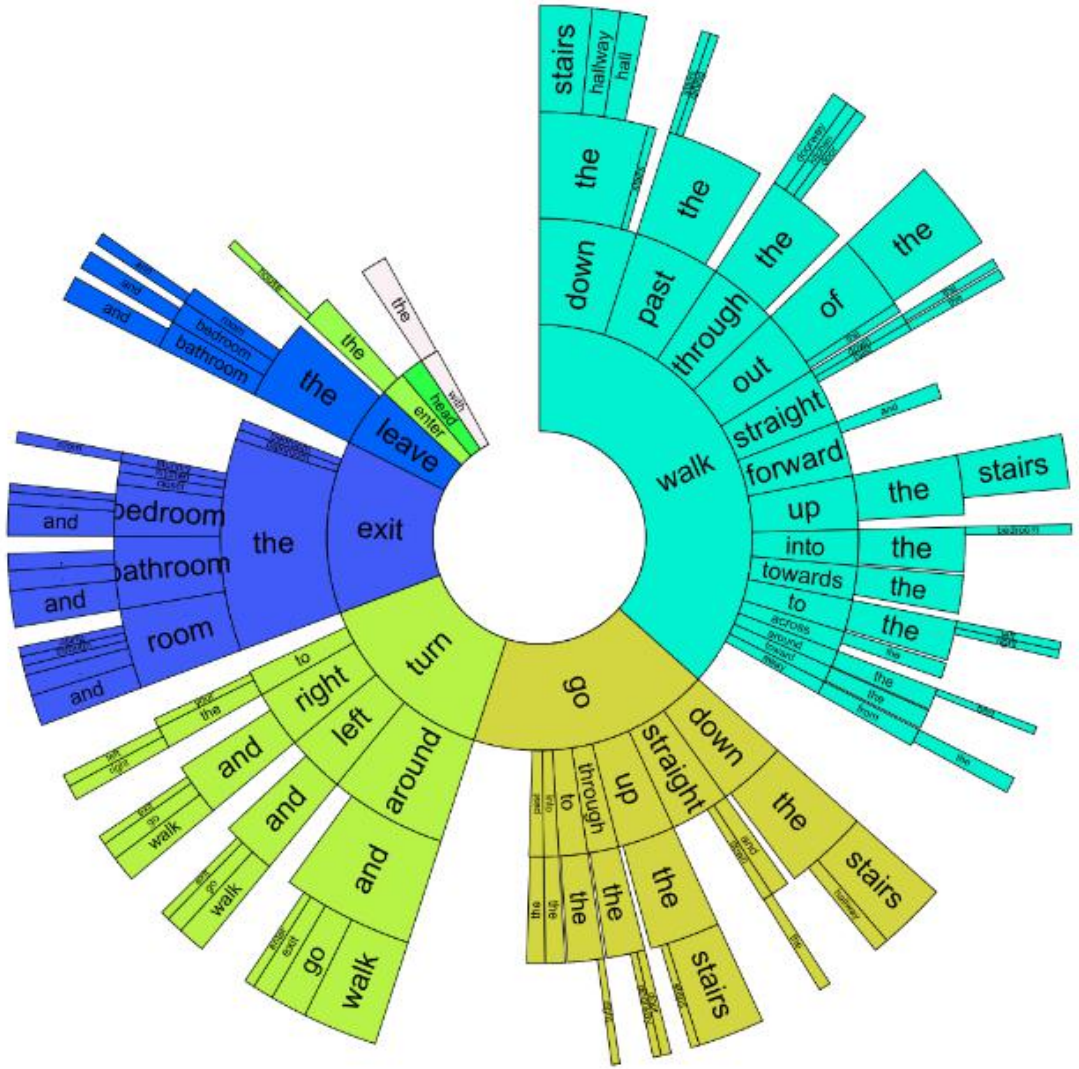
$$< s_0, a_0, s_1, a_1, \ldots >$$

# Dataset & Platform

Room-to-Room

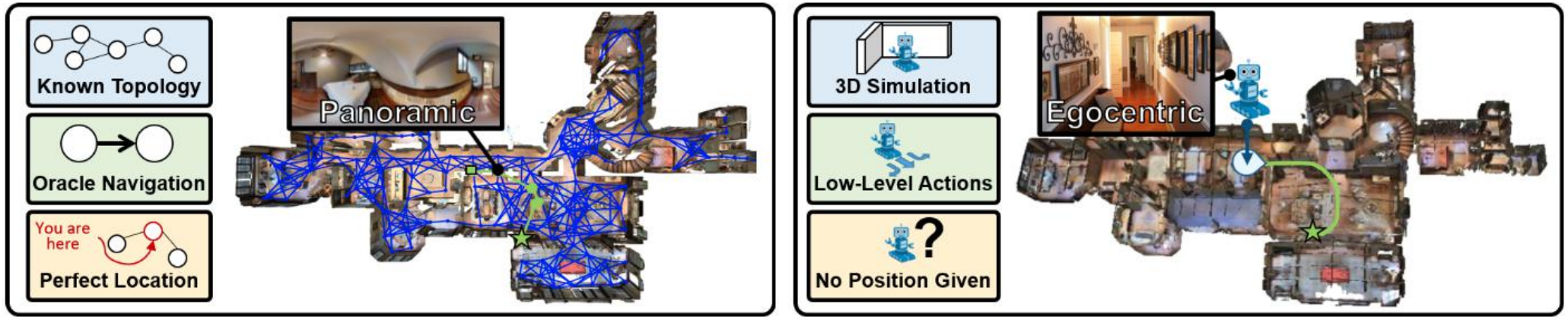


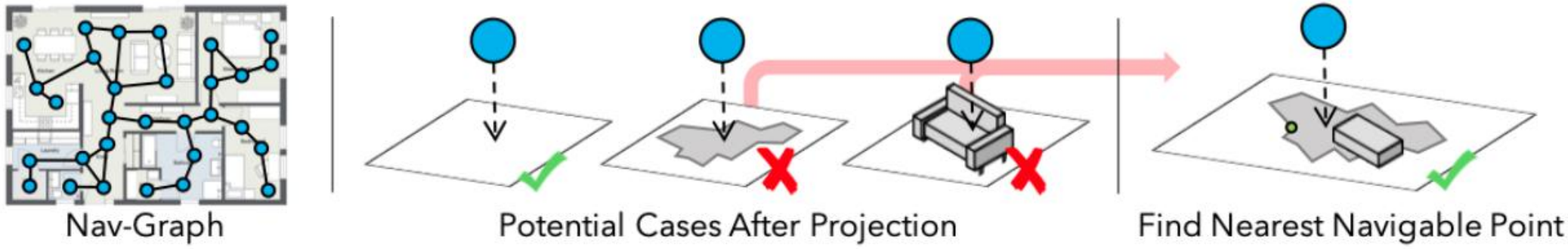
# Dataset & Platform

Vision-and-Language Navigation in Continuous Environments



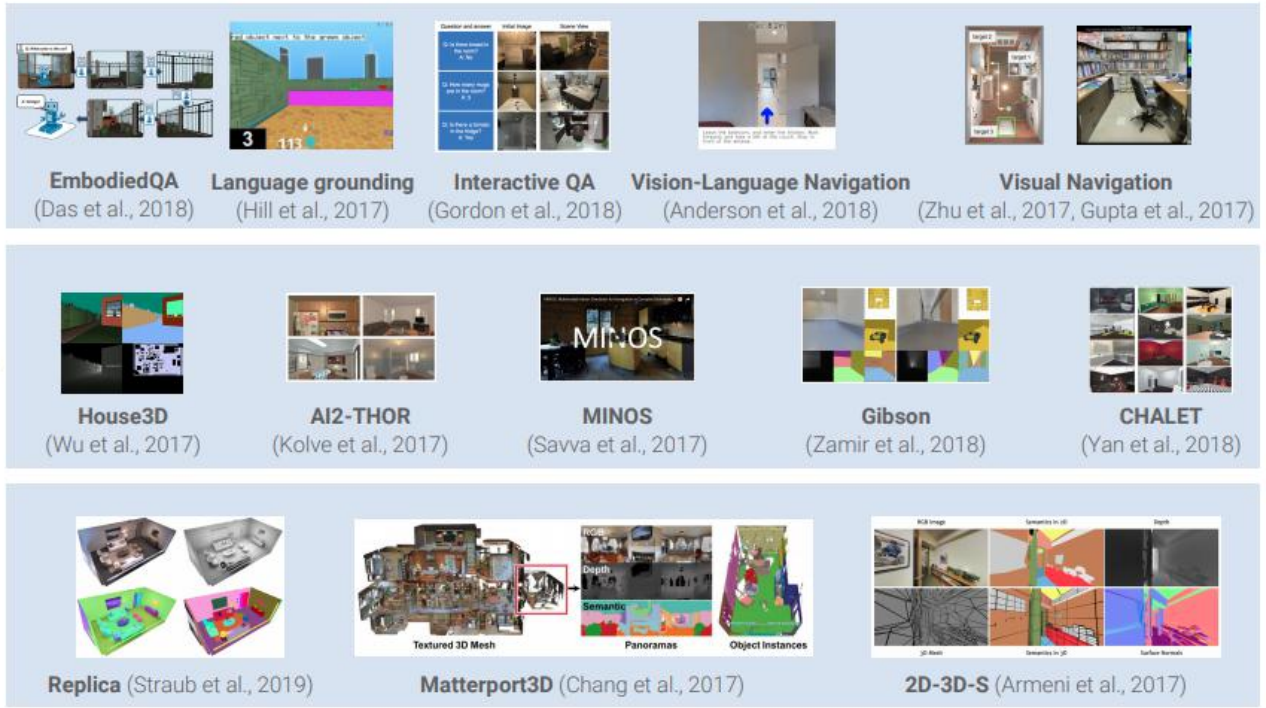

(a) Vision-and-Language Navigation (VLN) (b) VLN in Continuous Environments (VLN-CE)

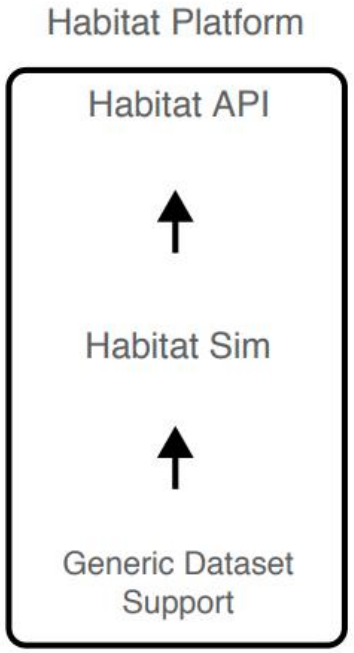
Converting Room-to-Room Trajectories to Habitat.

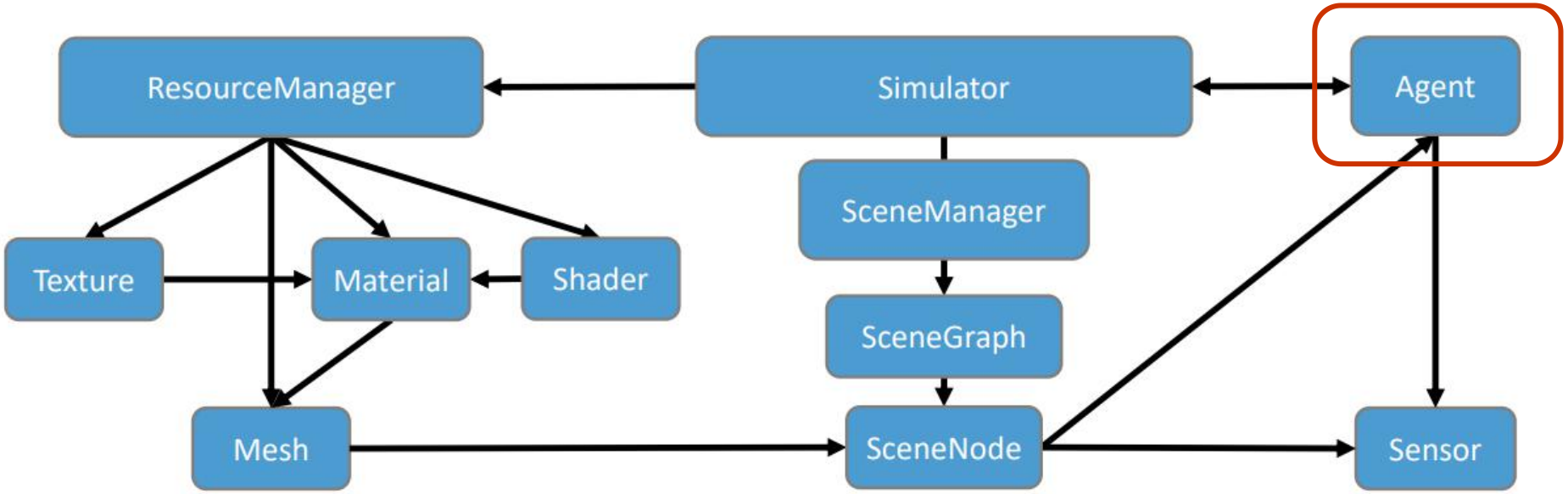


# Dataset & Platform

Habitat




https://aihabitat.org/

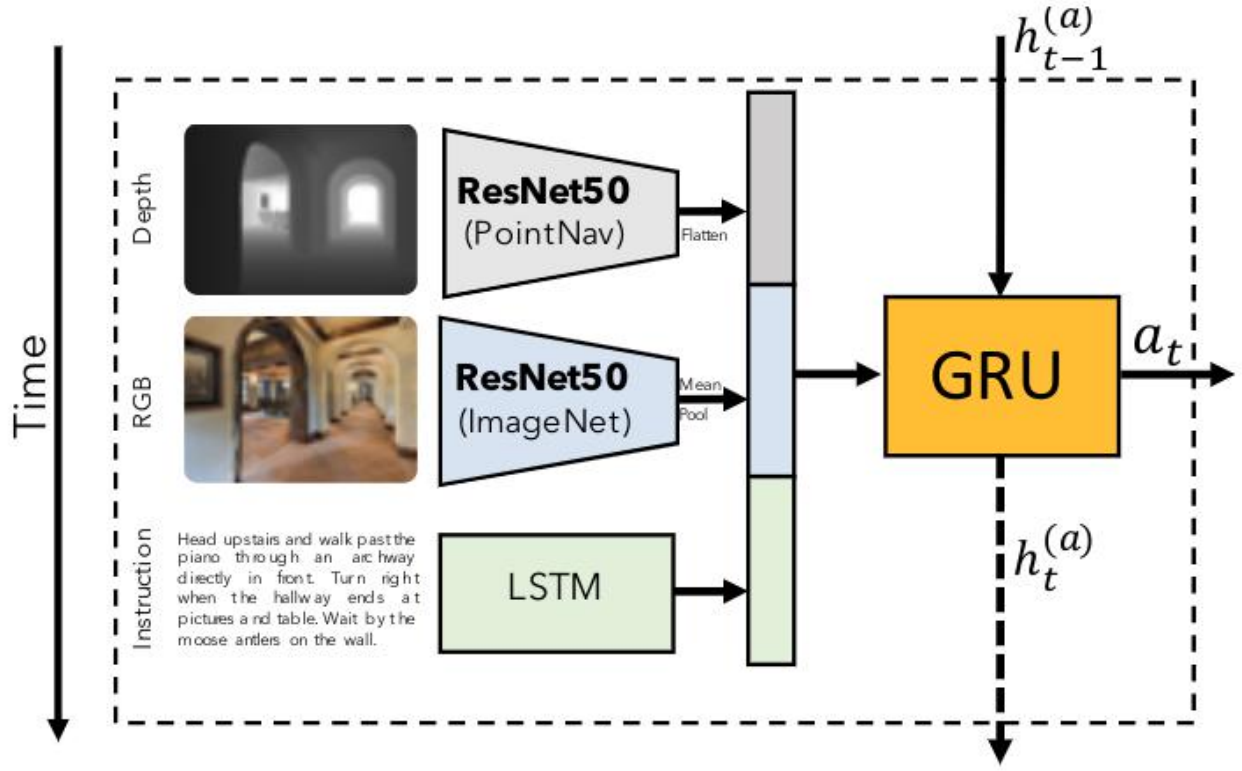
# Models

Sequence-to-Sequence Model

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

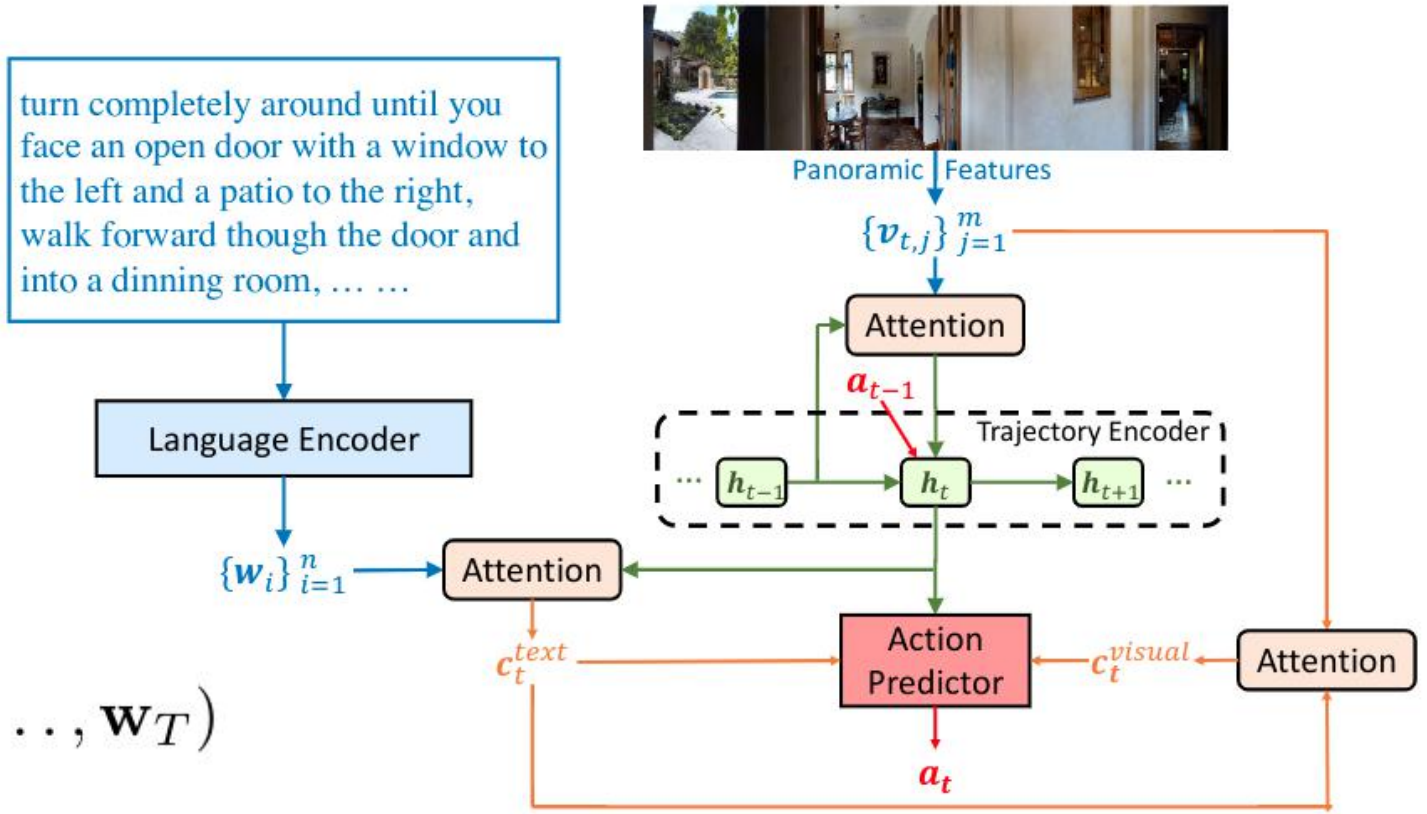
$$\mathbf{h}_t^{(a)} = \text{GRU}\left(\left[\bar{\mathbf{v}}_t, \bar{\mathbf{d}}_t, \mathbf{s}\right], \mathbf{h}_{t-1}^{(a)}\right)$$

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



# Models

Cross-Modal Attention




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

$$\mathcal{S} = \{\mathbf{s_1}, \ldots, \mathbf{s_T}\} = \mathrm{BiLSTM}\left(\mathbf{w}_1, \ldots, \mathbf{w}_T\right)$$

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

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

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

# Futher Works

- Instance-level information

- Structural memory

- Dynamic Nav-Graph

- ...