Object Finding Query
(One-Hot Encoding)

512

Masking

MLP

$Q(s,a)$ → L_{Critic}

$\pi(a|s)$ → L_{actor}

512   512   256   5

VECA Visual Observation
(Binocular Eyesight)

Left-Eye(RGB channel)

Right-Eye(RGB channel)

CNNs

84   84   20   4   9   3   7
8   8   20   4   9   3   7
6   32   64   64

MLP

3136   256

VECA Audio Observation
(3D spatialized audio)

Left-Ear HRTF

Right-Ear HRTF

Frequency (samples)

250   1

250   1

Flatten
$R^{\in 500}$

1

MLP

500   256

512
Interaction
Feature Maps

(a) Training Process

Real-World Observation

6   84   84
8
32   20   4
64   9   3
64   7

CNNs

3136   MLP   256

Interaction
Feature Maps

Linear Regression

Image Classification

"Teddy Bear"

(b) Transfer Learning