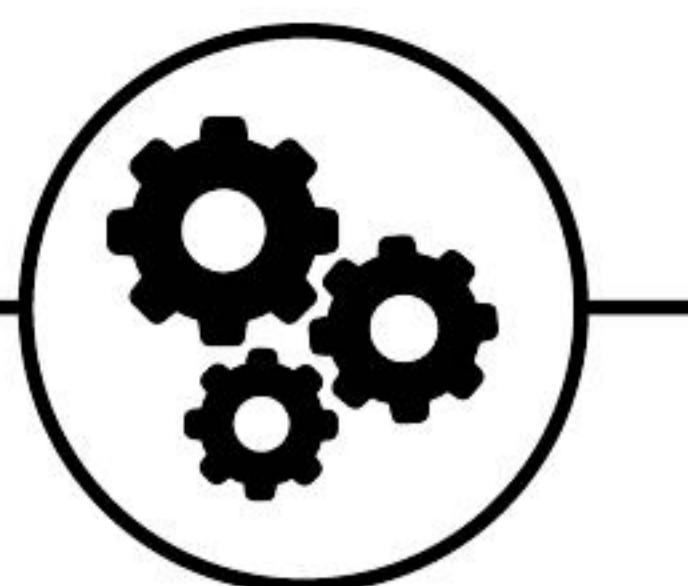


Program

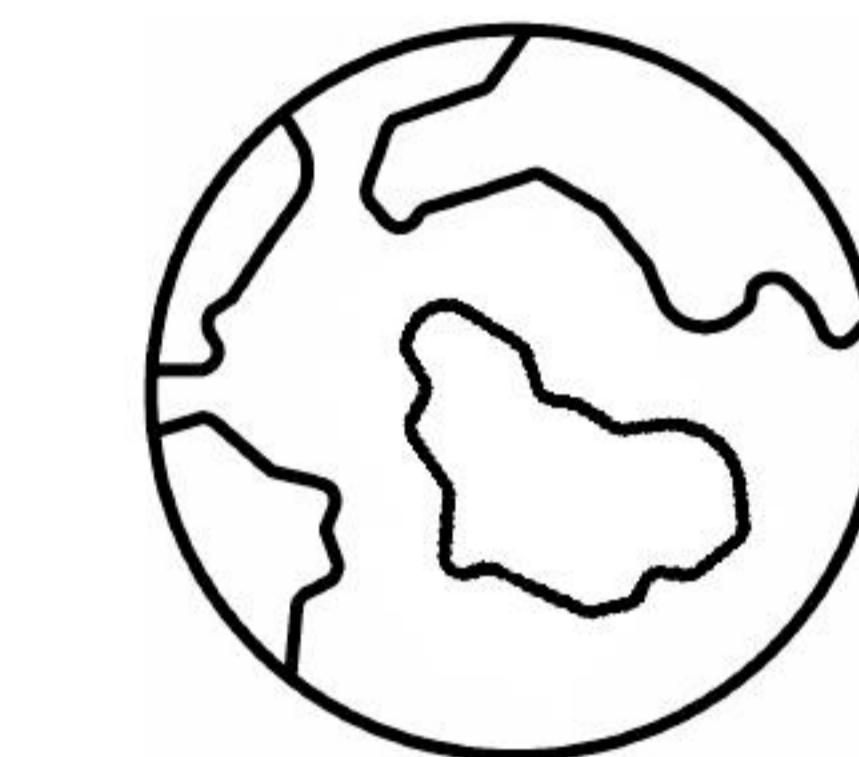
```
DEF run() m(
  WHILE c( markerPresent c) w(
    WHILE c( markerPresent c) w(
      pickMarker
      move w)
    turnRight
    move
    turnLeft
    WHILE c( markerPresent c) w(
      pickMarker
      move w)
    turnLeft
    move
    turnRight w) m)
```

Execute



# Learning to Synthesize Programs as Interpretable and Generalizable Reinforcement Learning Policies

Environment



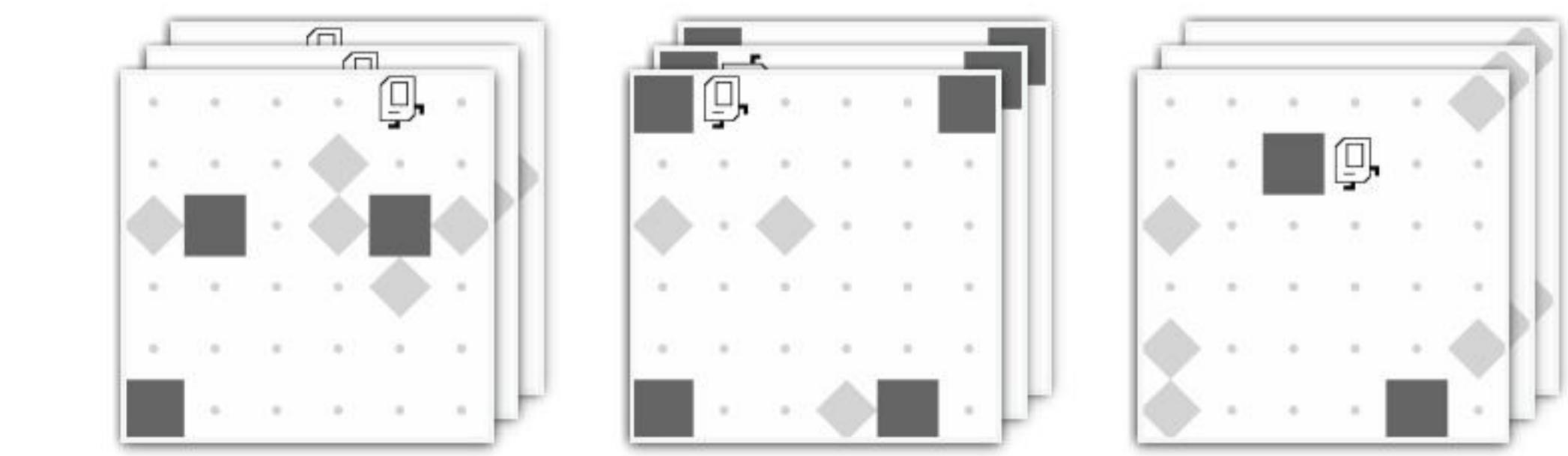
Shao-Hua Sun (孫紹華)

Assistant Professor

Dept. of Electrical Engineering (EE)

National Taiwan University

Demonstrations



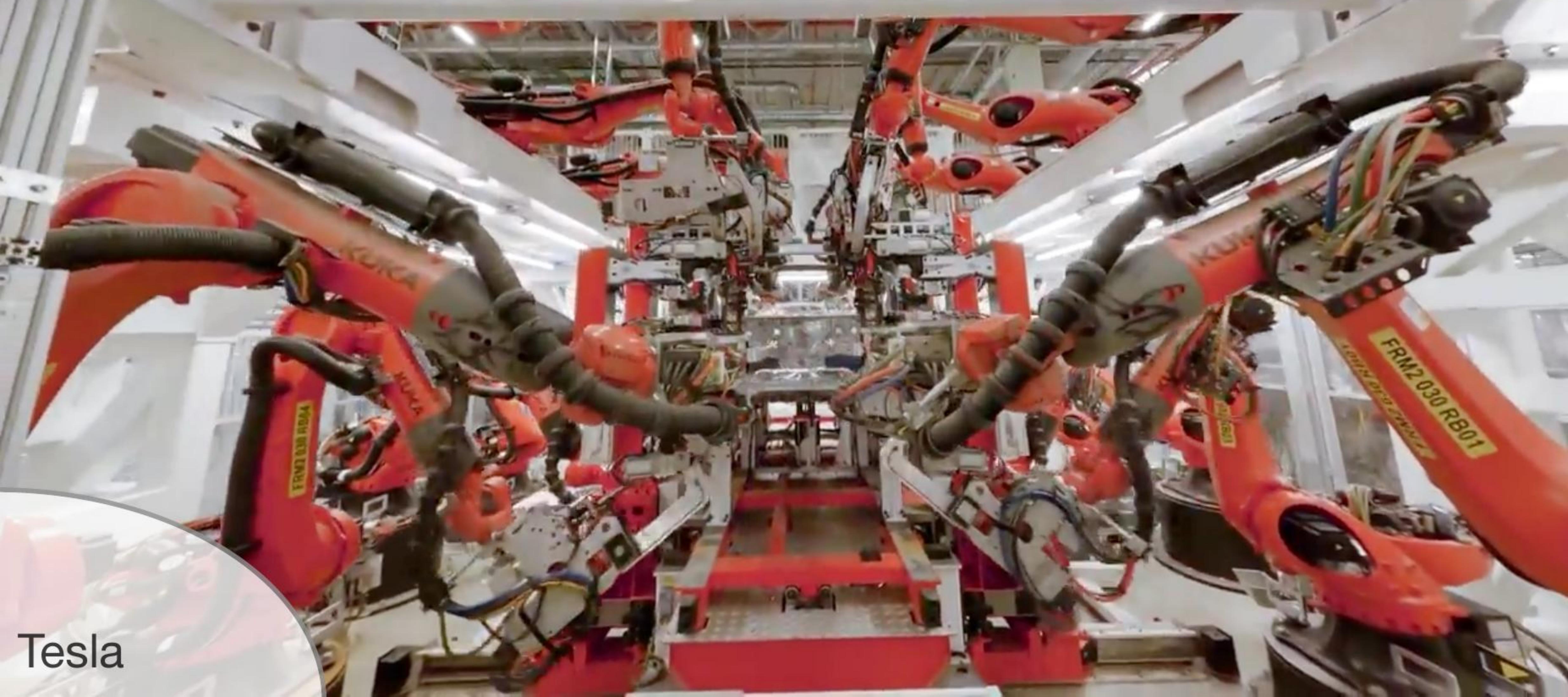
@ TAAI AI Forum 人工智慧論壇 2023

Reward

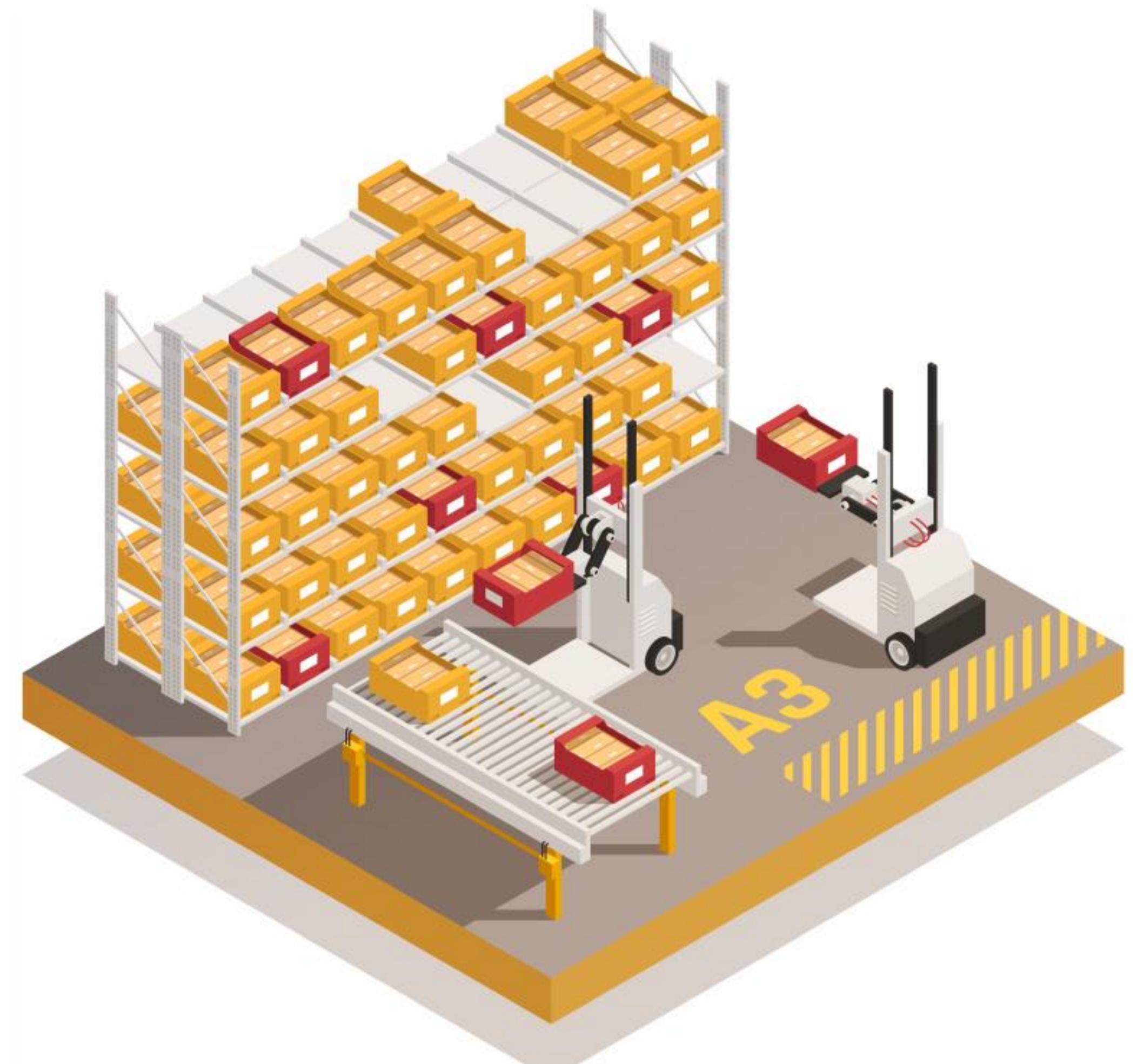




# Why Aren't Robots in Our Everyday Lives?



## Environment



Structured

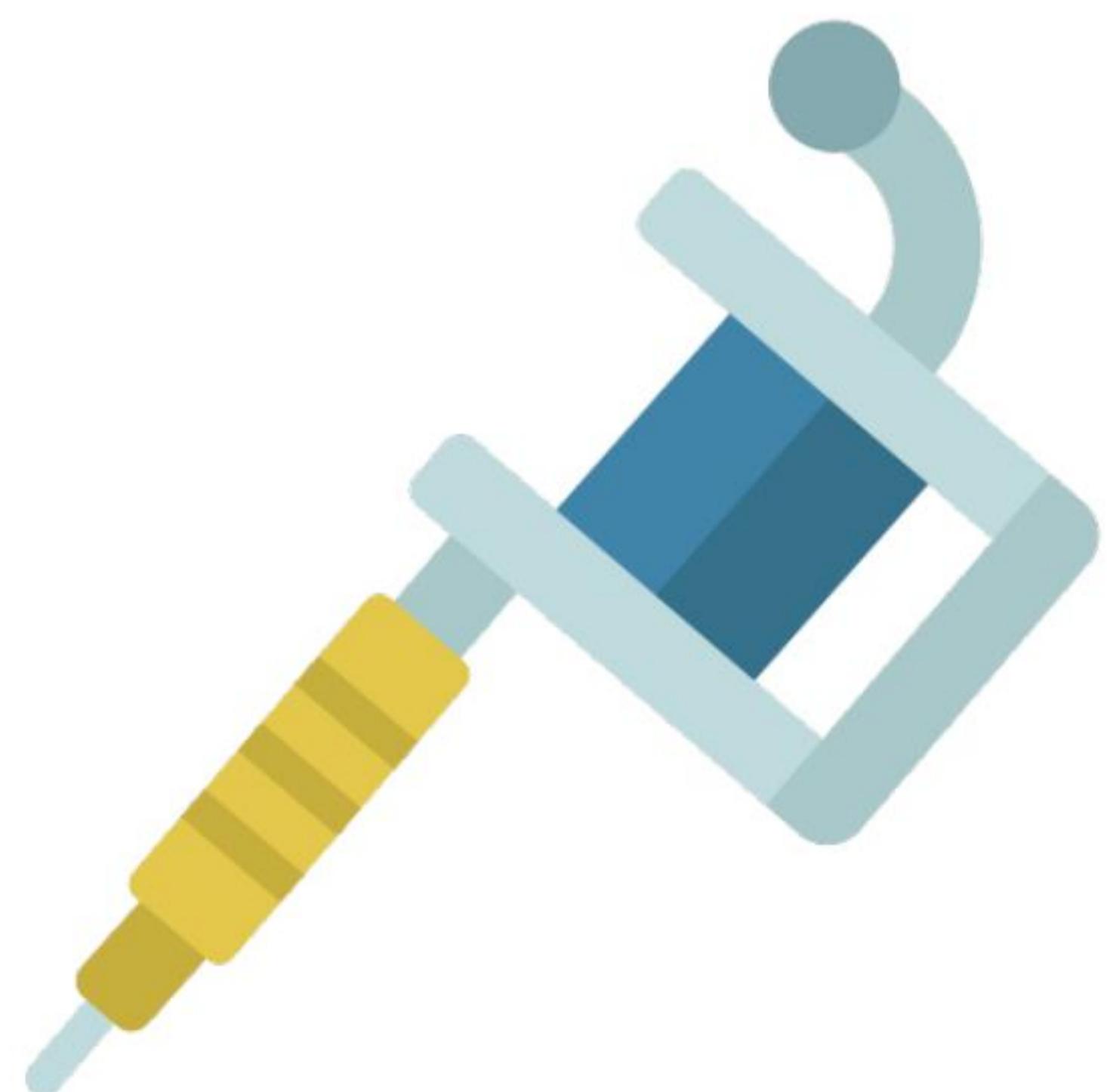


Unstructured

## Object



Known

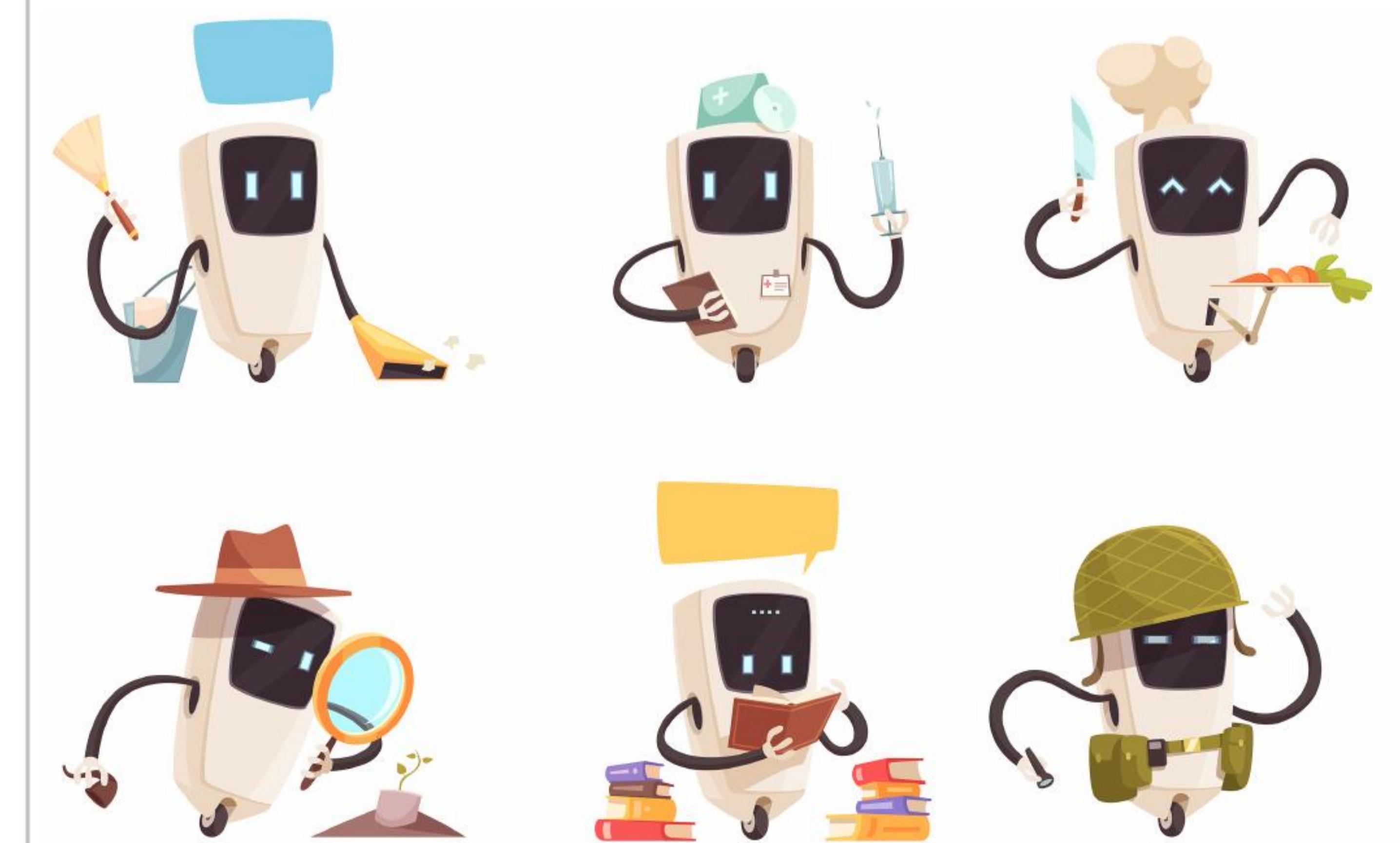


Unseen

## Task



Pre-defined / Pre-programmed



Diverse and Novel

A Venn diagram consisting of two overlapping circles. The left circle is colored light red and contains the text "Machine Learning". The right circle is colored light blue and contains the text "Robotics". The overlapping area between the two circles is shaded in a darker shade of both colors and contains the text "Robot Learning".

Machine  
Learning

Robot  
Learning

Robotics

# Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2M	1.94B	74.8	92.2
NASNet-A (5 @ 1538)	299×299	10.9M	2.35B	78.6	94.2
Inception V3 [60]	299×299	23.8M	5.72B	78.8	94.4
Xception [9]	299×299	22.8M	8.38B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8M	13.2B	80.1	95.1
NASNet-A (7 @ 1920)	299×299	22.6M	4.93B	80.8	95.3
ResNeXt-101 (64 × 4d) [68]	320×320	83.6M	31.5B	80.9	95.6
PolyNet [69]	331×331	92M	34.7B	81.3	95.8
DPN-131 [8]	320×320	79.5M	32.0B	81.5	95.8
SENet [25]	320×320	145.8M	42.3B	82.7	96.2
NASNet-A (6 @ 4032)	331×331	88.9M	23.8B	82.7	96.2

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image size reported in the table. Model size for [25] calculated from open-source implementation.

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Instance Segmentation

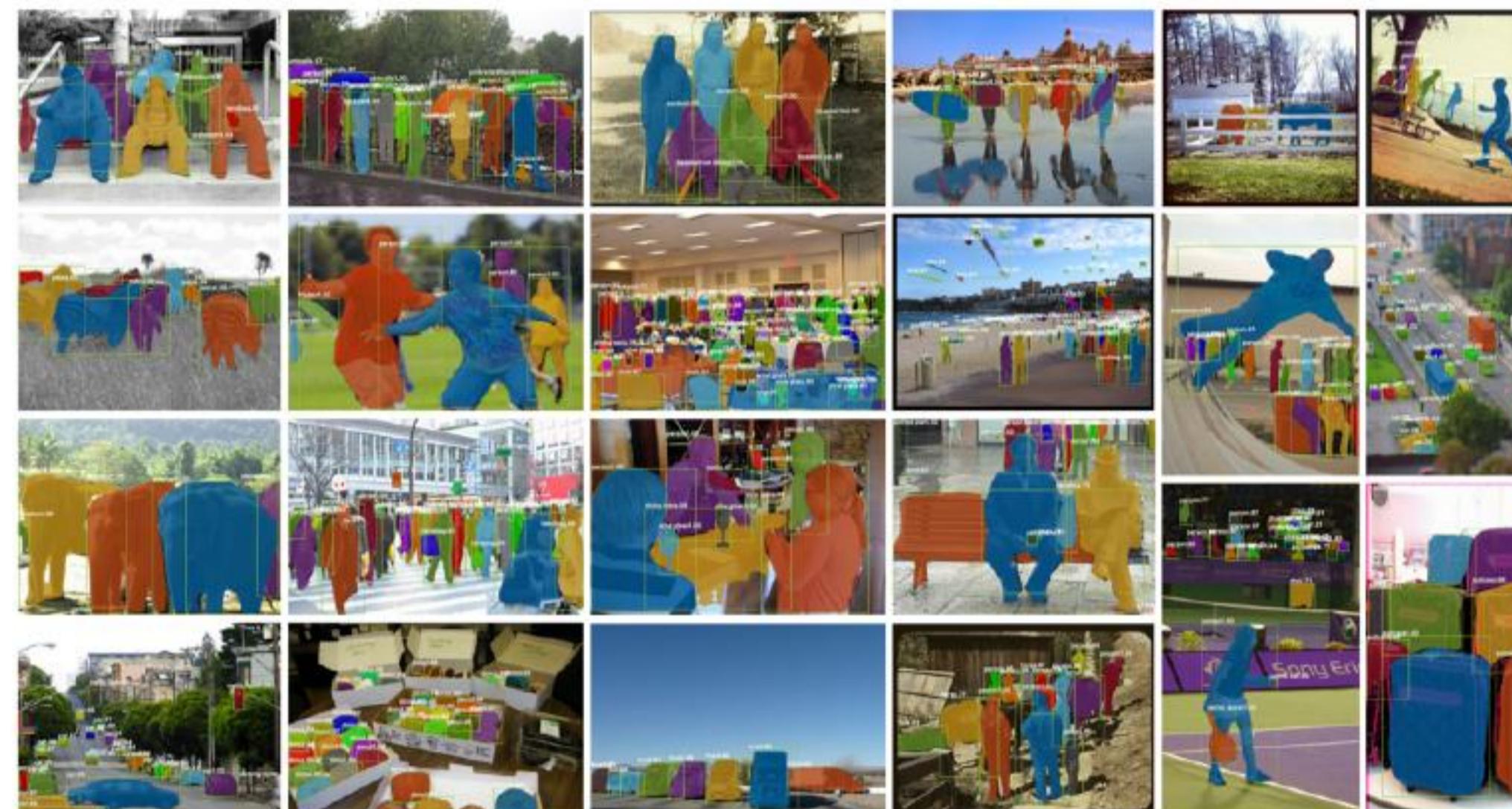


Figure 5. More results of Mask R-CNN on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).

He et al. Mask R-CNN

## Visual Question Answering

Method	VQA v2 test-dev			VQA v2 test-train		
	All	Value	Number	Other	All	Value
Prop (most common answer in training set) [1]	25.08	51.20	93.96	1.17	—	—
LSTM Language only (base model) [11]	44.26	67.01	71.85	27.37	—	—
Deeper LSTM Q (norm. 1) as reported in [1]	54.22	73.08	73.18	41.83	—	—
MCB [1] as reported in [1]	62.27	78.82	38.28	53.36	—	—
UNet+LSTM [1]	—	—	—	65.71	73.07	73.17
MoCo	67.50	82.50	44.16	59.97	—	—
Li+NUS	66.77	81.89	46.26	58.30	—	—
HDDU+SYD+UNCC	68.09	94.50	45.79	59.01	—	—
Proposed model	62.07	79.20	39.46	52.62	62.27	79.32
ResNet features 7x7, single network	65.32	81.82	44.22	56.06	65.27	81.76
Image features from bottom-up attention, adaptive $K$ , single network	66.31	83.38	43.17	51.10	66.23	83.71
ResNet features 7x7, ensemble	69.87	86.08	45.09	50.80	70.34	86.68
Image features from bottom-up attention, adaptive $K$ , ensemble	69.87	86.08	45.09	50.80	70.34	86.68

Table 3. Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

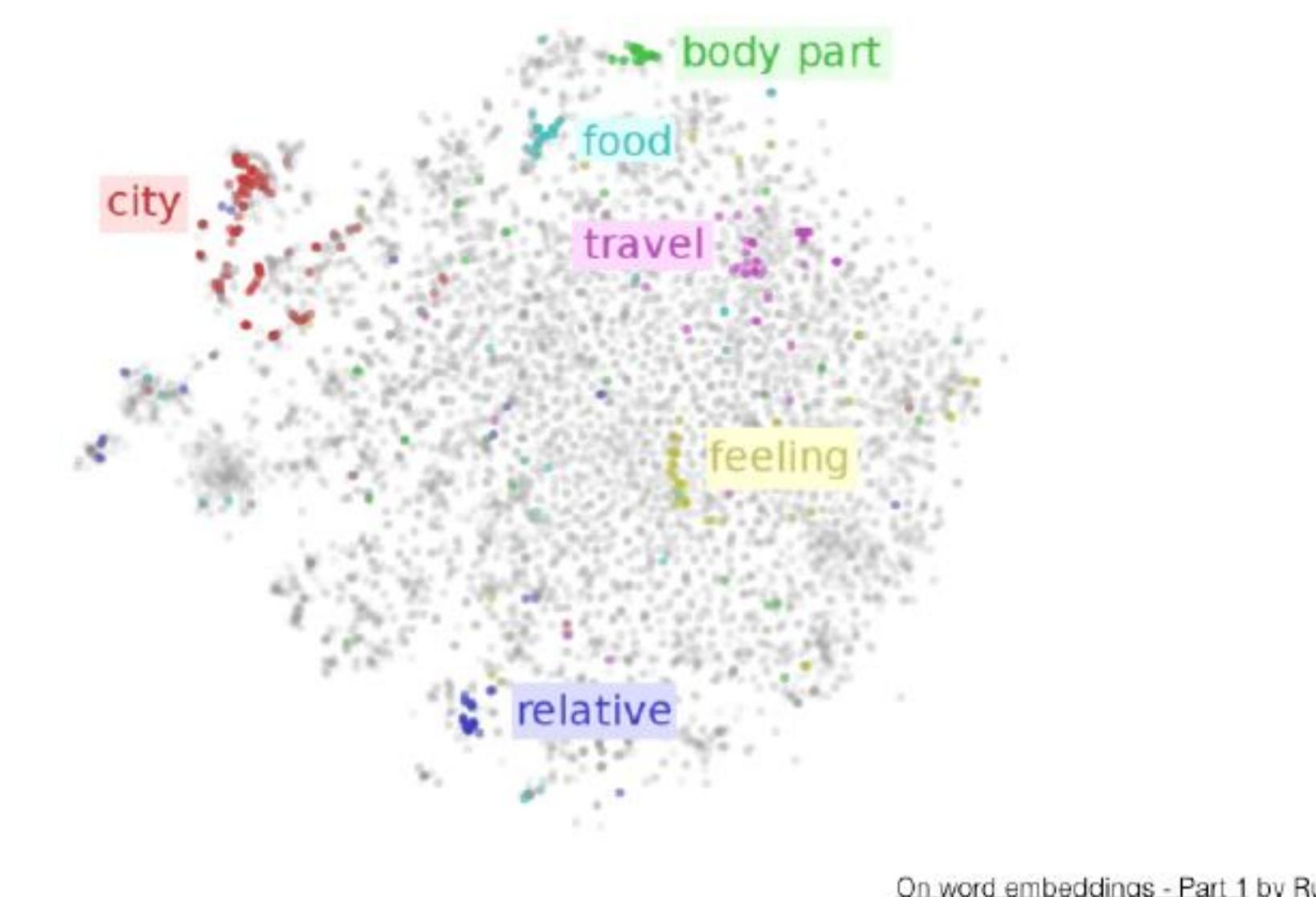
Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

## Machine Translation

Source	"The reason Boeing is doing this is to cram more seats in to make their plane more competitive with our products," said Kevin Keniston, head of passenger comfort at Europe's Airbus.
PBMT	"La raison pour laquelle Boeing sout en train de faire, c'est de concentrer davantage de sièges pour prendre leur avion plus compétitive avec nos produits", a déclaré Kevin M. Keniston, chef du confort des passagers de l'Airbus de l'Europe.
GNMT	"La raison pour laquelle Boeing fait cela est de créer plus de sièges pour rendre son avion plus compétitif avec nos produits," a déclaré Kevin Keniston, chef du confort des passagers chez Airbus.
Human	"Boeing fait ça pour pouvoir éaser plus de sièges et rendre ses avions plus compétitifs par rapport à nos produits," a déclaré Kevin Keniston, directeur du Confort Passager chez l'avionneur européen Airbus.
Source	When asked about this, an official of the American administration replied: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington."
PBMT	Interrogé à ce sujet, un responsable de l'administration américaine a répondu : "Les Etats-Unis n'est pas effectuer une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington".
GNMT	Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

Wu et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

## Word Embeddings



On word embeddings – Part\_1 by Ruder

## Named Entity Recognition

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF B I Agent Peter Strzok PERSON . Who Criticized Trump PERSON in Texts, Is FiredimagePeter Strzok, a top F B I GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. Credit J. Kirkpatrick PERSON for The New York Times Adam Goldman ORG and Michael S. SchmidtLog PERSON 13 CARDINAL — Peter Strzok PERSON , the F B I GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE . Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F B I GPE lawyer, Lea Raig — IN PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F B I GPE, to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry. Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account. The F B I GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel. Robert S. Mueller III PERSON , the president has repeatedly denounced Mr. Strzok PERSON in posts on

Named Entity Recognition and Classification with Scikit-Learn by Susan Li  
Esteves et al. Named Entity Recognition in Twitter using Images and Text

## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - ninet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - ninet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BIDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT <sub>BASE</sub> (Single)	80.8	88.5	-	-
BERT <sub>LARGE</sub> (Single)	84.1	90.9	-	-
BERT <sub>LARGE</sub> (Ensemble)	85.8	91.8	-	-
BERT <sub>LARGE</sub> (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
BERT <sub>LARGE</sub> (Ens.+TriviaQA)	86.2	92.2	87.4	93.2

Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

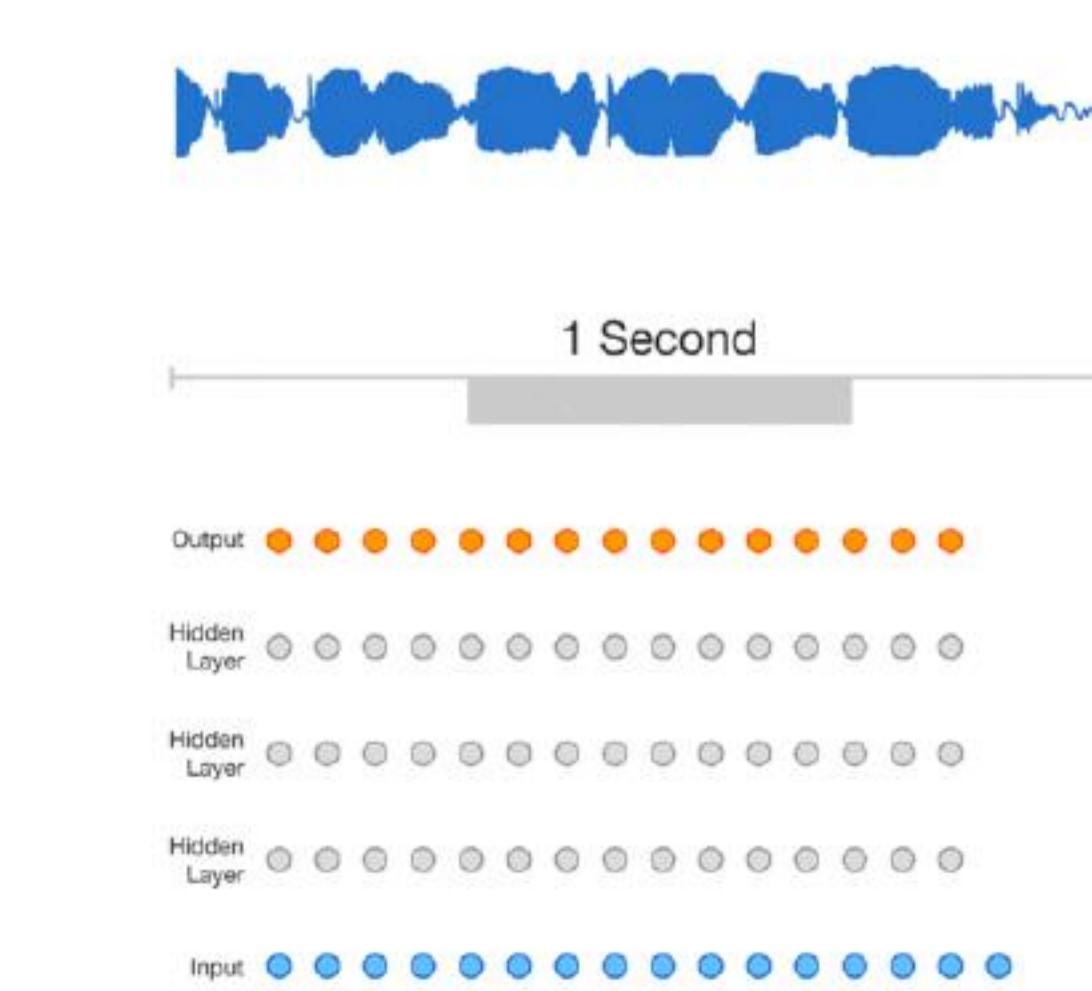
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## Speech Recognition

Senone set	Model/combination step	WER devset	WER test	WER devset	WER test
		ngram-LM	LSTM-LMs		
9k	BLSTM	11.5	8.3	9.2	6.3
27k-puhpuh	BLSTM	11.4	8.0	9.3	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	11.3	8.0	9.2	6.3
9k-puhpuh	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpuh	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8
Confusion network combination					
+ LSTM rescoring					
+ ngram rescoring					
+ backchannel penalty					

Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

## Speech Synthesis (text-to-speech)



Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

# Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2M	1.94B	74.8	92.2
NASNet-A (5 @ 1538)	299×299	10.9M	2.35B	78.6	94.2
Inception V3 [60]	299×299	23.8M	5.72B	78.8	94.4
Xception [9]	299×299	22.8M	8.38B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8M	13.2B	80.1	95.1
NASNet-A (7 @ 1920)	299×299	22.6M	4.93B	80.8	95.3
ResNeXt-101 (64 x 4d) [68]	320×320	83.6M	31.5B	80.9	95.6
PolyNet [69]	331×331	92M	34.7B	81.3	95.8
DPN-131 [8]	320×320	79.5M	32.0B	81.5	95.8
SENet [25]	320×320	145.8M	42.3B	82.7	96.2
NASNet-A (6 @ 4032)	331×331	88.9M	23.8B	82.7	96.2

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image size reported in the table. Model size for [25] calculated from open-source implementation.

## Instance Segmentation

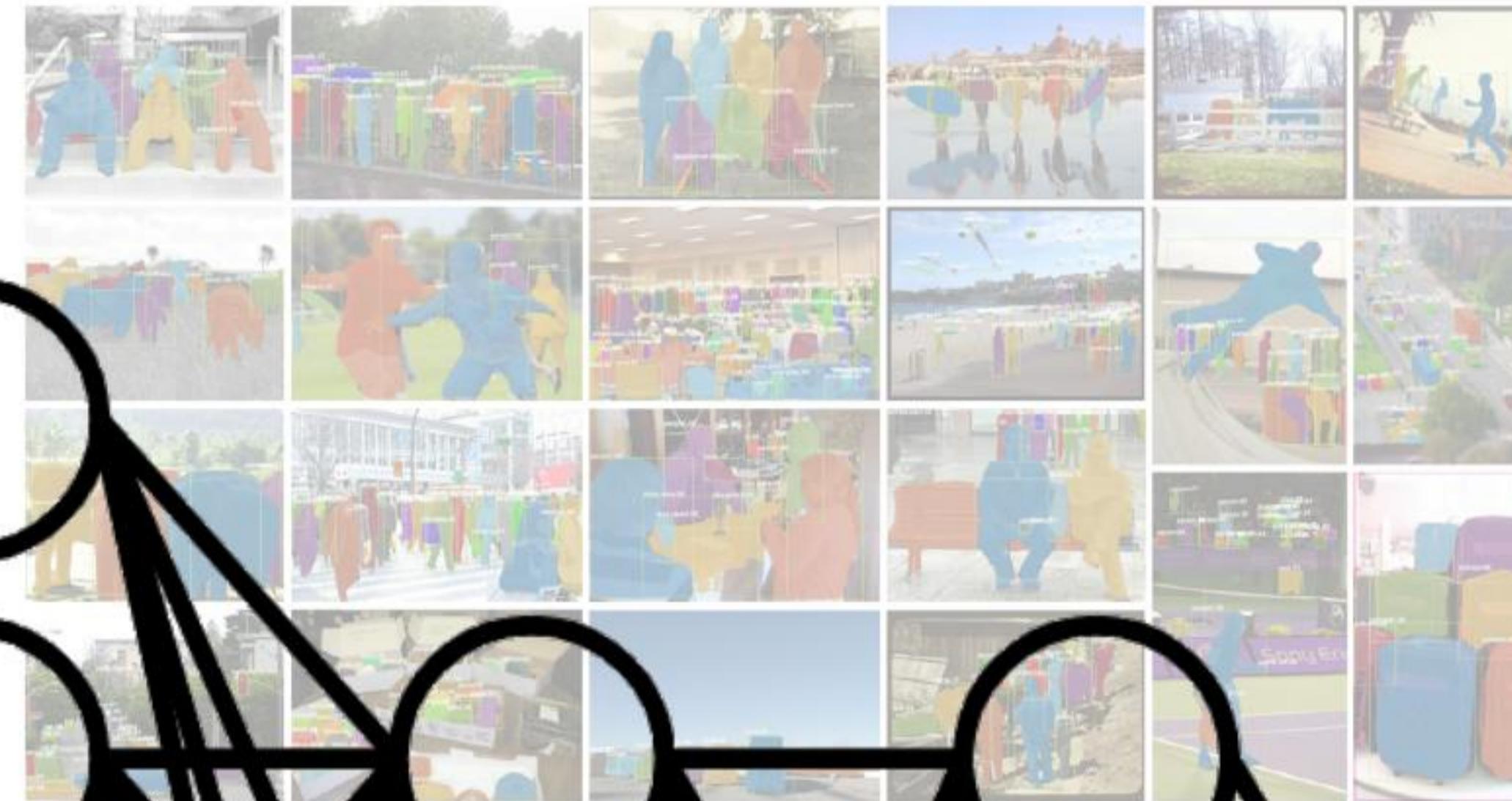


Figure 1. Qualitative results of Mask R-CNN on COCO test set, trained using ResNet-101 [28] and running at 5 FPS with 35.7 mask AP (Table 1).

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Method	VQA v2 test dev			VQA v2 test		
	All	Youth	Adults	All	Youth	Adults
prior (most common answer in training set) [1]	29.08	84.20	95.98	3.14	—	—
LSTM Language only (base model) [1]	—	—	—	41.26	81.03	91.85
Deep LSTM Q (ours, 127) as reported in [1]	—	—	—	54.22	73.06	85.18
MCR [1] (as reported in [1])	—	—	—	62.27	78.82	82.28
UPMC-LDNN [1]	—	—	—	63.71	81.37	87.17
LM-NUS	—	—	—	67.70	82.50	84.49
HDU-SYD-UNCC	—	—	—	66.77	81.89	82.89
Proposed model	68.09	94.50	95.91	59.01	—	—
ResNet features 7x7, single network	62.07	79.20	89.96	93.62	62.27	79.71
Image features from bottom-up attention, adaptive K, single network	65.52	81.62	84.27	85.07	70.30	85.00
ResNet features 7x7, ensemble	66.31	81.39	83.17	81.60	66.23	82.87
Image features from bottom-up attention, adaptive K, ensemble	68.87	86.08	89.99	90.80	70.34	86.60

Table 3. Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

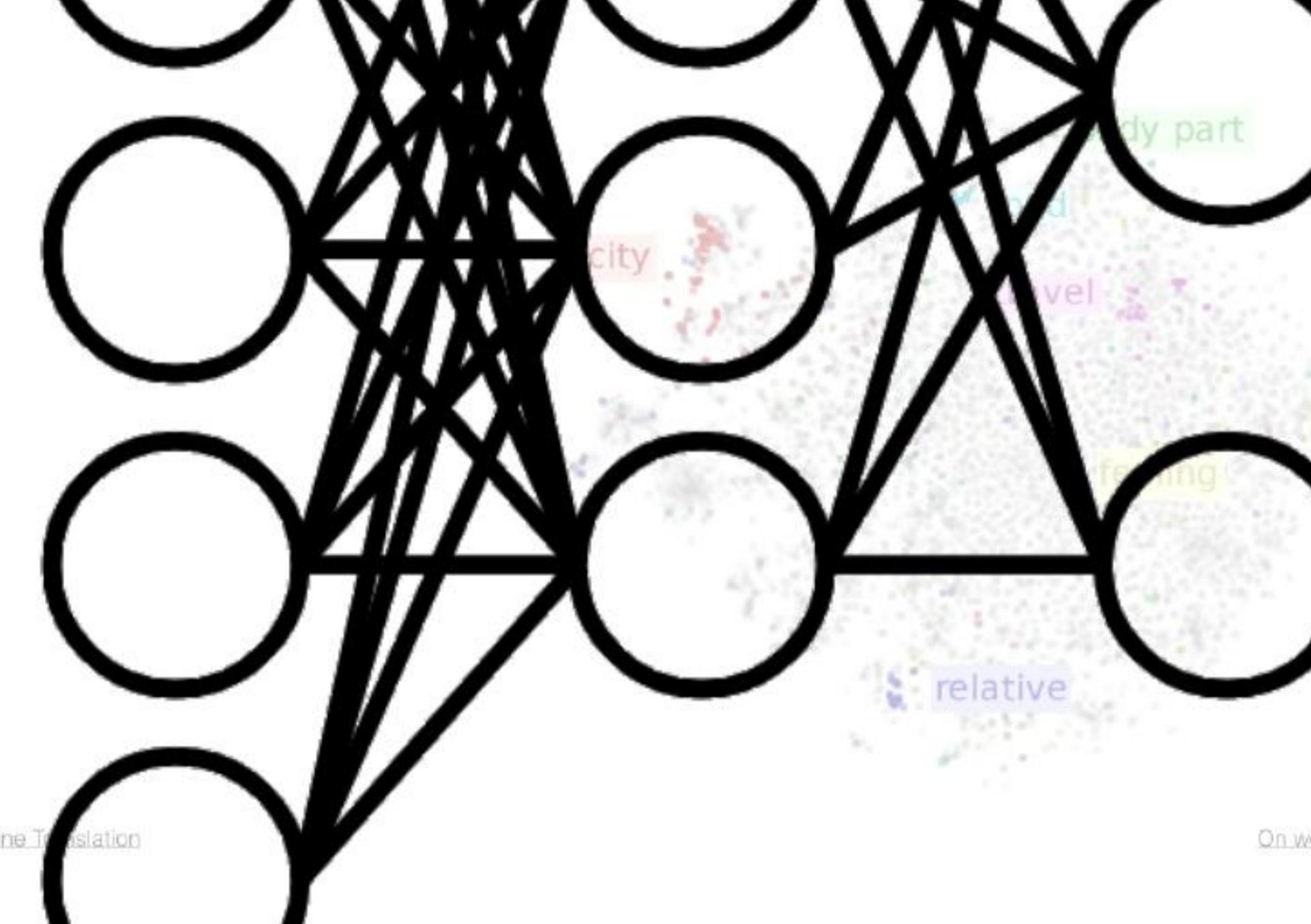
Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

## Machine Translation

Source	Input
PBMT	"The rose competitive Europe's "La raison sièges pour Keniston, "La raison avion plus passagers chez Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs Human Source When asked about this, un officiel de l'American administration replied: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington." PBMT Etats-Unis n'est pas effectuer une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington". Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington". Human Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".
GNMT	leur plane more ger confort at davantage de déclaré Kevin M. 3.0 leur rendre son avion plus pas de confort des Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs par rapport à nos produits", a déclaré Kevin Keniston, directeur de Confort Passager 6.0 chez l'avionneur Airbus. Human Source When asked about this, un officiel de l'American administration replied: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington." PBMT Etats-Unis n'est pas effectuer une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington". Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington". Human Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

Wu et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

## Word Embeddings



On word embeddings...Part\_1 by Ruder

## Named Entity Recognition

Output  
"I dropped by site index/politics/subscribe  
who criticized Trump PERSON in Text  
investigation after his disparaging texts against  
Timothy Adam Goldman ORG and  
ARDIN, CARDINAL — Peter Strzok PERSON  
PERSON, the F.B.I. GPE senior counterintelligence agent Peter Strzok PERSON.  
oversed the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON's lawyer said Monday DATE Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer.  
Lisa Page — IN PERSON assisting the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry. Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account. The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON, who was removed last summer DATE from the staff of the special counsel. Robert D. Mueller III PERSON The present has repeatedly denounced Mr. Strzok PERSON in posts on

Named Entity Recognition and Classification with Scikit-Learn by Susan Li  
Esteves et al. Named Entity Recognition in Twitter using Images and Text

## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - qnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - qnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BIDAF+ELMo (Single)	-	-	85.8	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT <sub>BASE</sub> (Single)	80.8	88.5	-	-
BERT <sub>LARGE</sub> (Single)	84.1	90.9	-	-
BERT <sub>LARGE</sub> (Ensemble)	85.8	91.8	-	-
BERT <sub>LARGE</sub> (Sgl+TriviaQA)	84.2	91.1	85.1	91.8
BERT <sub>LARGE</sub> (Ens.+TriviaQA)	86.2	92.2	87.4	93.2

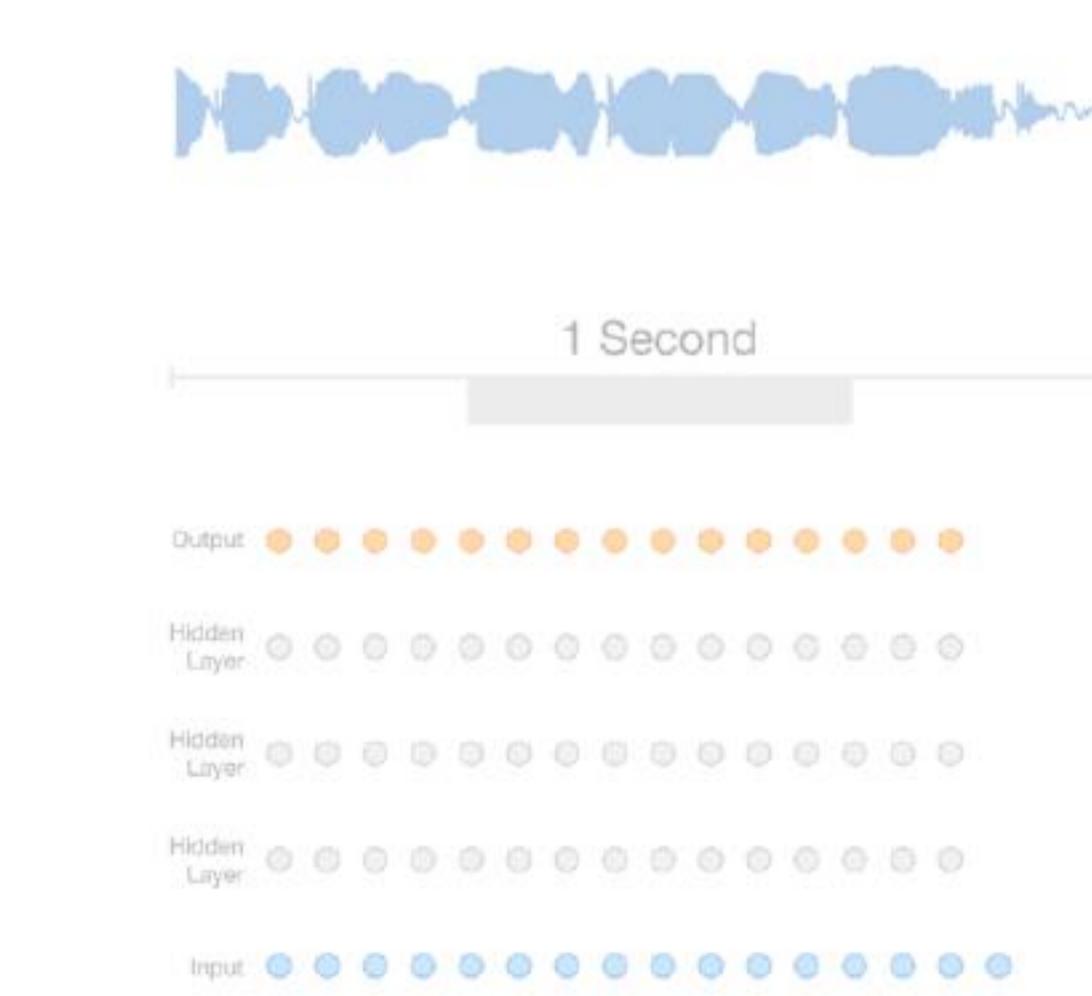
Table 2: SQuAD results. The BERT ensemble is 7x faster than the baseline systems which use different pre-training checkpoints and fine-tuning seeds.

## Speech Recognition

Senone set	Model/combination step	WER	WER	WER	WER
		devset	test	devset	test
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
27k-puhpum	BLSTM	11.3	8.0	9.2	6.3
9k	BLSTM+ResNet+LACE+CNN+BLSTM	9.6	7.2	7.7	5.4
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpum	BLSTM+ResNet+LACE+CNN+BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8
Confusion network combination		-	-	7.4	5.2
+ LSTM rescoring		-	-	7.3	5.2
+ ngram rescoring		-	-	7.2	5.2
+ backchannel penalty		-	-	7.2	5.1

Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

## Speech Synthesis (text-to-speech)



Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

# Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224x224	11.2M	1.94B	74.8	92.2
NASNet-A (5 @ 1538)	299x299	10.9 M	2.35 B	78.6	94.2
Inception V3 [60]	299x299	23.8 M	5.72 B	78.8	94.4
Xception [9]	299x299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299x299	55.8 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	299x299	22.6 M	4.93 B	80.8	95.3
ResNeXt-101 (64 x 4d) [68]	320x320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331x331	97 M	34.7 B	81.3	95.8
DPN-131 [8]	331x331	10.4 M	32.0 B	81.5	95.8
SENet [25]	331x331	145 M	42.3 B	82.7	96.2
NASNet-A (6 @ 4032)	331x331	86.9 M	23.8 B	82.7	96.2



Wu et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

## Instance Segmentation



Figure 1. Qualitative results of Mask R-CNN on COCO test set, using ResNet-101 [2] and running at 5 FPS with 35.7 mask AP (Table 1).

He et al. Mask R-CNN

## Visual Question Answering

Method	VQA v2 test dev			VQA v2 test				
	All	Youth	Adults	All	Youth	Adults		
prior (most common answer in training set) [1]	29.08	84.20	95.98	3.14	—	—		
LSTM Language only (base model) [1]	—	—	—	41.26	81.03	91.85	27.37	
Deep LSTM Q (ours, 127) as reported in [1]	—	—	—	54.22	73.06	85.18	41.83	
MCR [1] (as reported in [1])	—	—	—	62.27	78.82	82.28	52.36	
UPMC-LDNN [1]	—	—	—	63.71	81.17	87.17	47.17	
LM-NUS	—	—	—	67.79	82.50	84.49	59.07	
HDU-SYD-UNCC	—	—	—	66.77	81.89	82.89	58.30	
Proposed model	68.09	84.50	85.91	59.01	—	—	—	
ResNet features 7x7, single network	62.07	79.20	89.46	93.62	62.27	79.32	89.77	82.39
Image features from bottom-up attention, adaptive K, single network	65.52	81.62	84.23	87.07	65.52	81.62	84.23	86.26
ResNet features 7x7, ensemble	66.31	81.29	83.17	81.60	66.23	82.04	84.20	87.20
Image features from bottom-up attention, adaptive K, ensemble	68.87	86.08	88.99	90.80	70.34	86.68	88.64	83.43

Tanay et al. Tips and tricks for Visual Question Answering: Learnings from the 2017 Challenge

## Class

Tanay et al. Tips and tricks for Visual Question Answering: Learnings from the 2017 Challenge

## Question Answering

System	Dev EM	Test F1	Dev EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BIDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours	80.8	88.5	-	-
BERT <sub>BASE</sub> (Single)	84.1	90.9	-	-
BERT <sub>LARGE</sub> (Single)	85.8	91.8	-	-
BERT <sub>LARGE</sub> (Ensemble)	84.2	91.1	85.1	91.8
BERT <sub>LARGE</sub> (Sgl.+TriviaQA)	86.2	92.2	87.4	93.2

Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

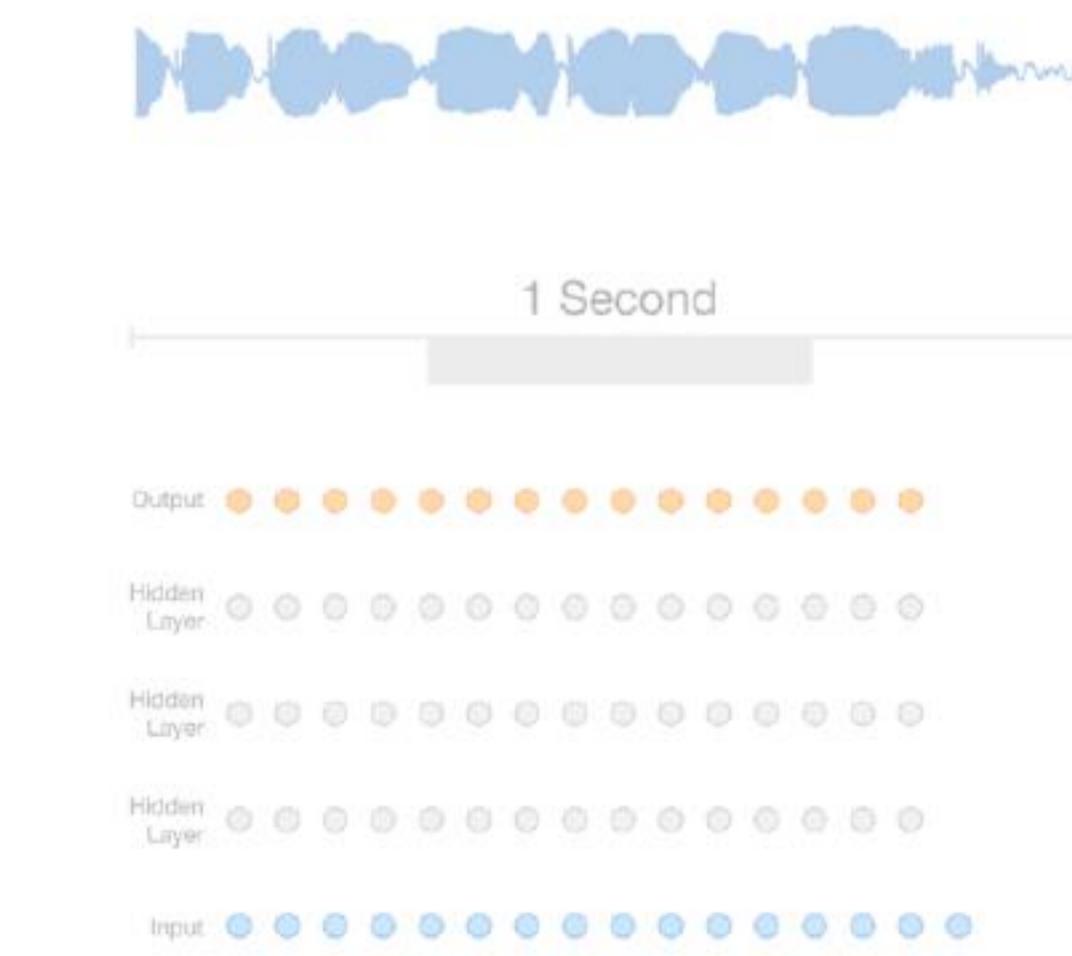
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## Speech Recognition

Senone set	Model/combination step	WER	WER	WER	WER
		devset	test	devset	test
9k	BLSTM	11.5	8.3	9.2	6.3
27k-puhpum	BLSTM	11.4	8.0	9.3	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8
-	Confusion network combination	-	7.4	5.2	-
-	+ LSTM rescoring	-	7.3	5.2	-
-	+ ngram rescoring	-	7.2	5.2	-
-	+ backchannel penalty	-	7.2	5.1	-

Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

## Speech Synthesis (text-to-speech)



Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

# Supervised Learning

# Image Classification

Model	image size	# parameters	Mult-Add	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	299×299	10.9 M	2.35 B	78.6	94.2
Inception V3 [60]	299×299	23.8 M	5.72 B	78.8	94.4
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	299×299	22.6 M	4.93 B	80.8	95.3
ResNeXt-101 (64 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	320×320	145.8 M	42.3 B	82.7	96.2
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	96.2

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operation is calculated for the image size reported in the table. Model size for [25] calculated from open-source implementation.

# English sentence

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Instance Segmentation

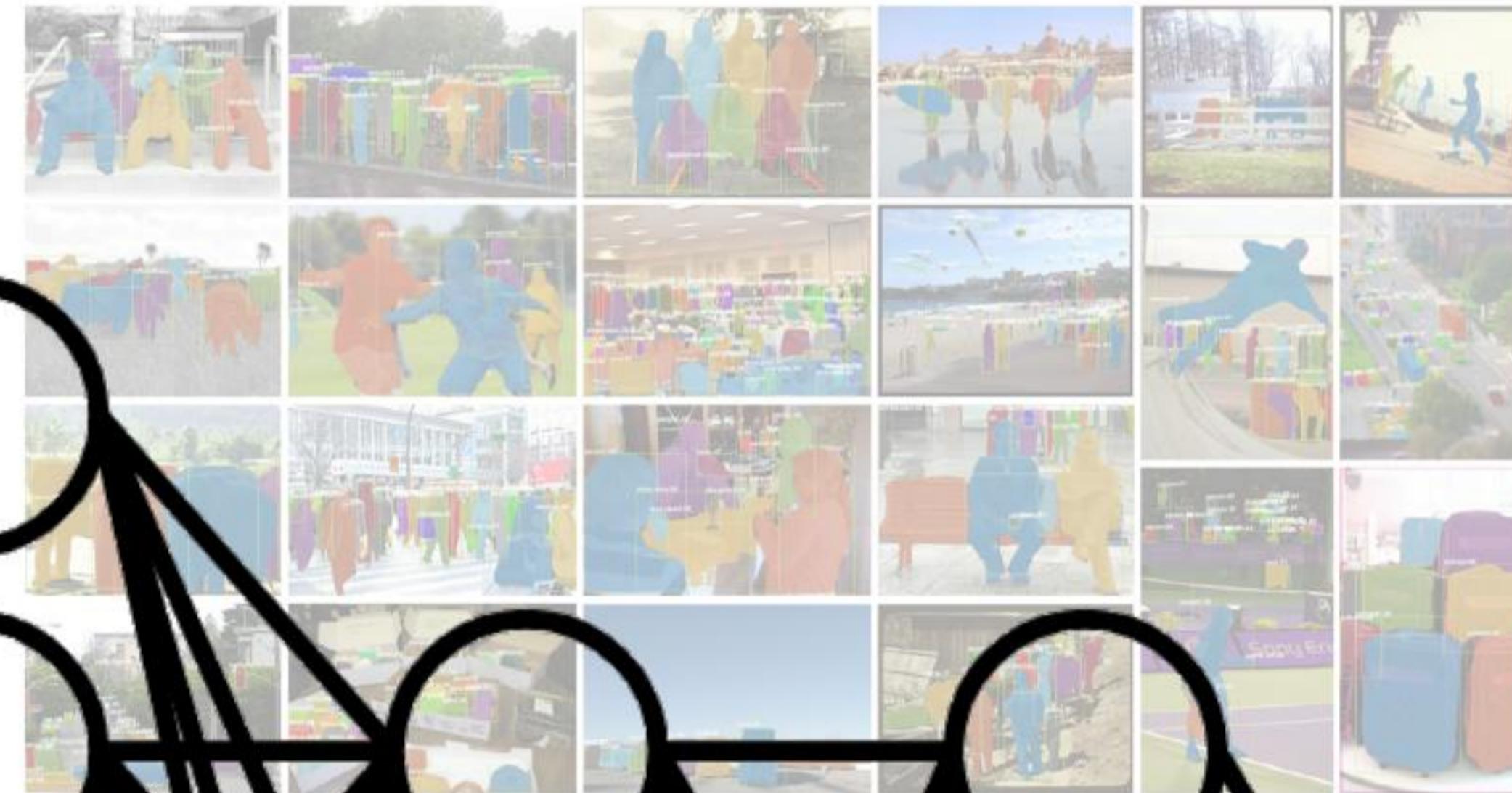


Fig. 5. Qualitative results of Mask R-CNN on COCO test images using ResNet-101 [2] and running at 5 fps with 35.7 mask AP (Table 1).

# Visual Question Answering

Method	VQA v2 test-dev				VQA v2 test-std			
	All	Yes/no	Num/b	Other	All	Yes/no	Num/b	Other
Prior (most common answer in training set) [11]	-	-	-	-	25.98	61.20	0.56	1.17
LSTM Language only (blind model) [44]	-	-	-	-	44.26	67.01	31.55	21.37
Deeper LSTM Q norm. 1 [7] as reported in [33]	-	-	-	-	54.22	73.36	35.18	41.83
MCB [13] as reported in [33]	-	-	-	-	62.27	78.82	38.28	53.36
UPMC-LIP6 [1]	-	-	-	-	65.71	82.07	41.06	51.12
Athena	-	-	-	-	67.59	82.50	44.19	59.97
LV-NUS	-	-	-	-	66.77	81.89	46.29	58.30
HDIU-USYD-UNOC	-	-	-	-	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive $K$ , single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive $K$ , ensemble	69.87	86.08	48.99	60.80	70.34	86.60	48.64	61.15

Table 3. Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

### **Facebook contacts**

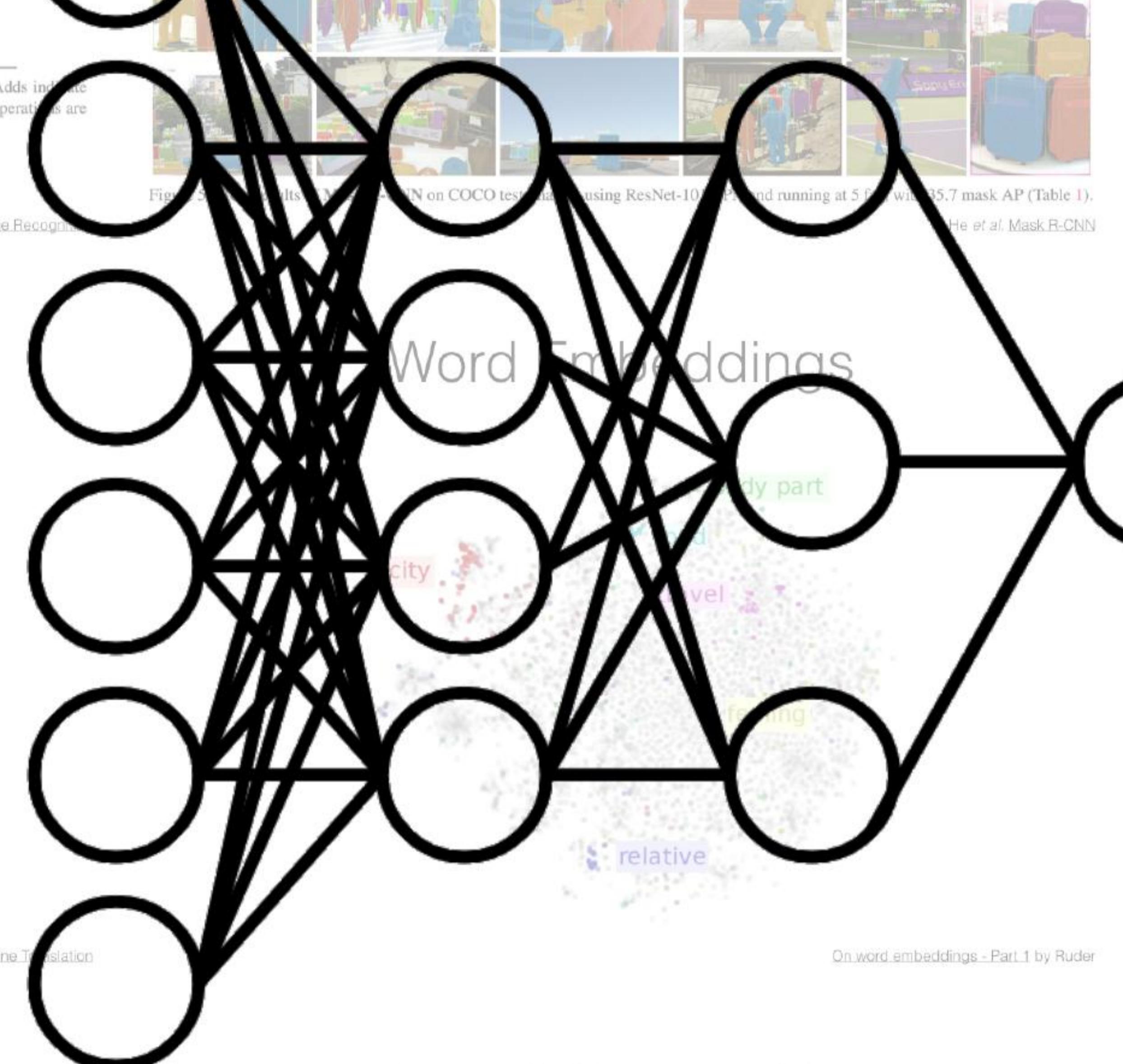
Teney et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

Teney et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

France is never cold  
in September

		pour rendre son chef du confort des	6.0
Human	"Boeing fait ça pour pouvoir caser plus de sièges et rendre ses avions plus compétitifs par rapports à nos produits", a déclaré Kevin Keniston, directeur de Confort Passager chez l'avionneur européen Airbus.		6.0
Source	When asked about this, an official of the American administration replied: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington."		
PBMT	Interrogé à ce sujet, un responsable de l'administration américaine a répondu : "Les Etats-Unis n'est pas effectuer une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington".	3.0	
GNMT	Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les États-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".	6.0	
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".	6.0	

Wu et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation



On word embeddings - Part 1 by Ruder

## Question Answering

	System	Dev		Test	
		EM	Fl	EM	Fl
Leaderboard (Oct 8th, 2018)					
Human		-	-	82.3	90.0
#1 Ensemble - nlnet		-	-	86.0	90.0
#2 Ensemble - QANet		-	-	84.5	90.0
#1 Single - nlnet		-	-	83.5	90.0
#2 Single - QANet		-	-	82.5	88.0
Published					
BiDAF+ELMo (Single)		-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	88.0	-
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.0	-
Ours					
BERT <sub>BASE</sub> (Single)	80.8	88.5	-	-	-
BERT <sub>LARGE</sub> (Single)	84.1	90.9	-	-	-
BERT <sub>LARGE</sub> (Ensemble)	85.8	91.8	-	-	-
BERT <sub>LARGE</sub> (Sgl.+TriviaQA)	84.2	91.1	85.1	90.0	-
BERT <sub>LARGE</sub> (Ens.+TriviaQA)	86.2	92.2	87.4	91.0	-

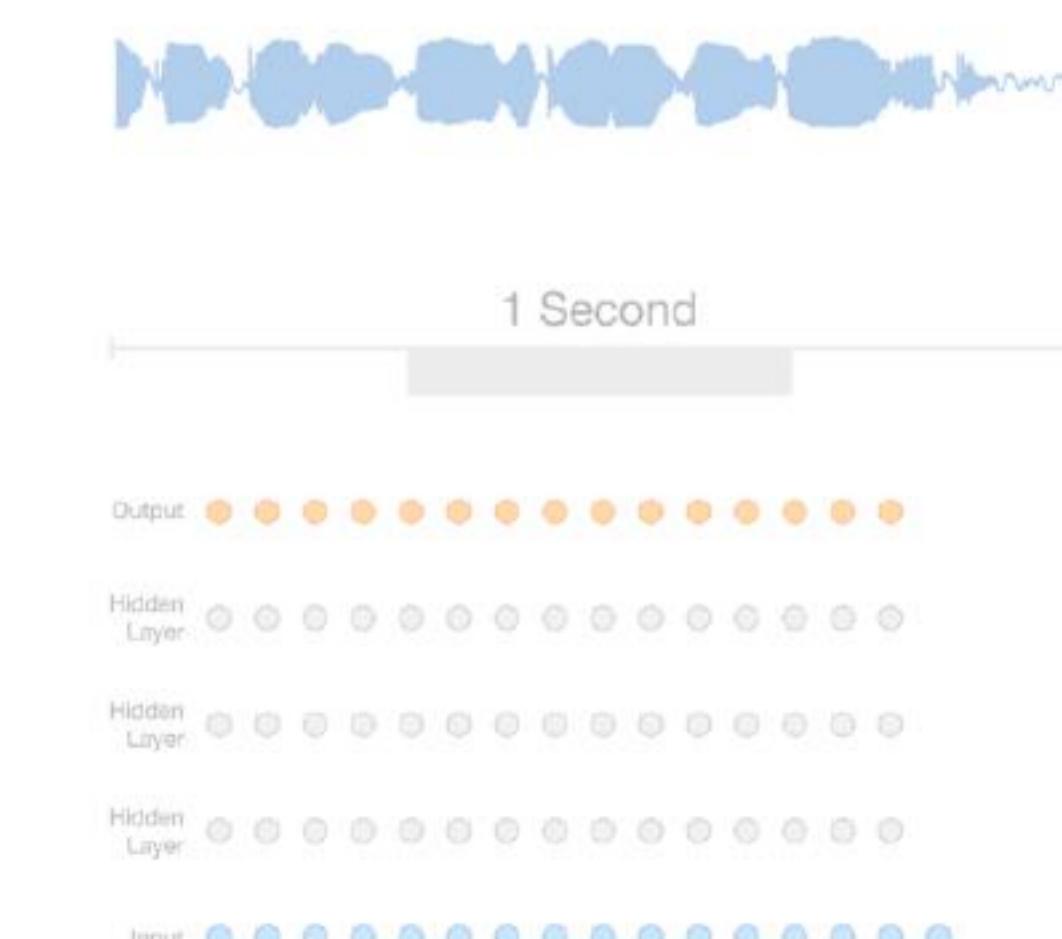
Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

# Speech Recognition

Senone set	Model/combination step	Word Error Rate			
		WER devset	WER test	WER devset	WER test
	ngram-LM	LSTM-LM			
9k	BLSTM	11.5	8.3	9.2	6.5
27k	BLSTM	11.4	8.0	9.3	6.6
27k-puhpum	BLSTM	11.3	8.0	9.2	6.6
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.5
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.5
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.5
-	Confusion network combination			7.4	5.5
-	+ LSTM rescoring			7.3	5.5
-	+ ngram rescoring			7.2	5.5
-	+ backchannel penalty			7.2	5.5

Xiong et al.: The Microsoft 2017 Conversational Speech Recognition System

# Speech Synthesis (text-to-speech)



Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

# Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2M	1.94B	74.8	92.2
NASNet-A (5 @ 1538)	299×299	10.9M	2.35B	78.6	94.2
Inception V3 [60]	299×299	23.8M	5.72B	78.8	94.4
Xception [9]	299×299	22.8M	8.38B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8M	13.2B	80.1	95.1
NASNet-A (7 @ 1920)	299×299	22.6M	4.93B	80.8	95.3
ResNeXt-101 (64 x 4d) [68]	320×320	83.6M	31.5B	80.9	95.6
PolyNet [69]	331×331	92M	34.7B	81.3	95.8
DPN-131 [8]	320×320	79.5M	32.0B	81.5	95.8
SENet [25]	320×320	145.8M	42.3B	82.7	96.2
NASNet-A (6 @ 4032)	331×331	88.9M	23.8B	82.7	96.2

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image as reported in the table. Model size for [25] calculated from open-source implementation.

## Instance Segmentation

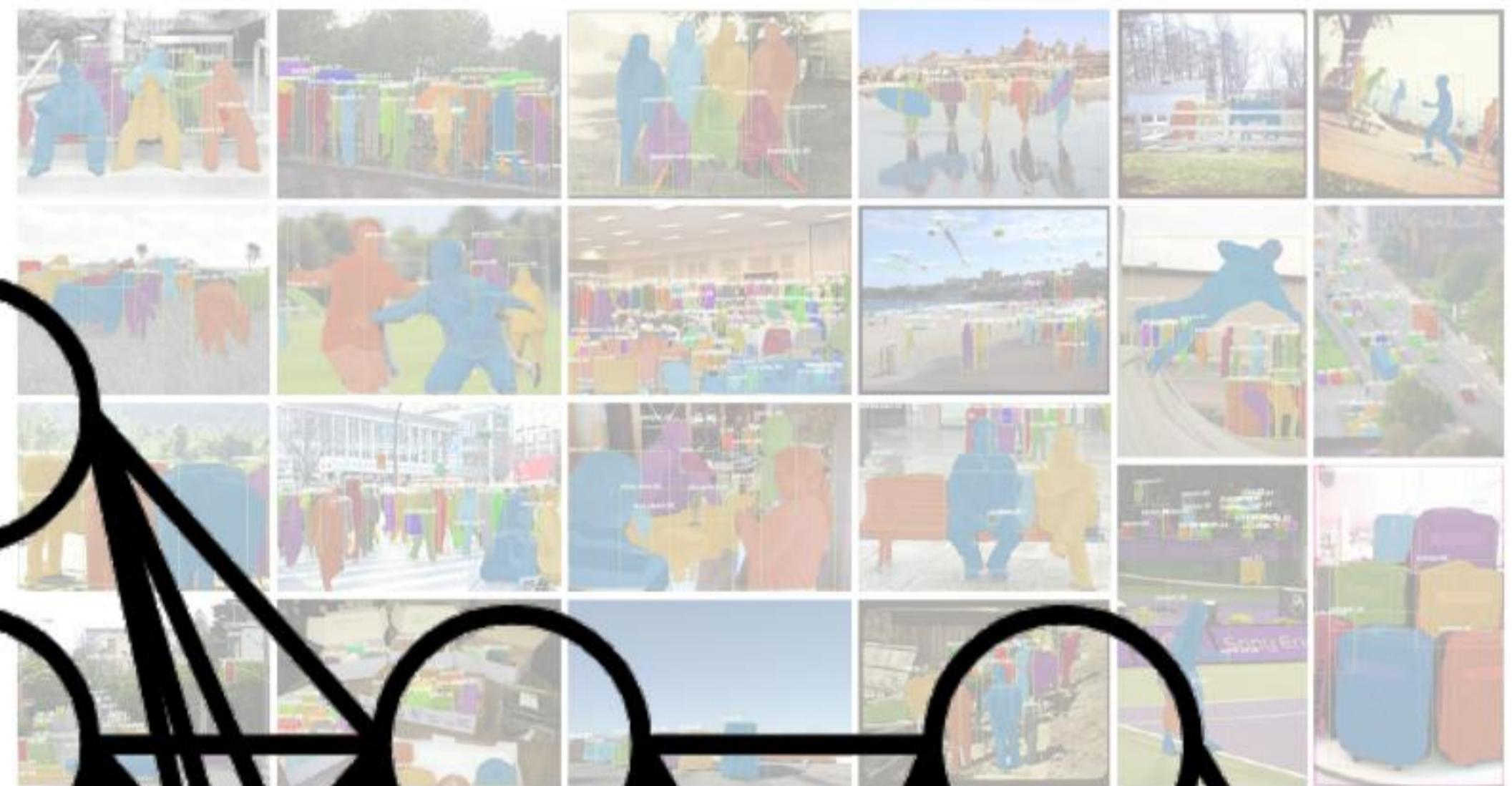


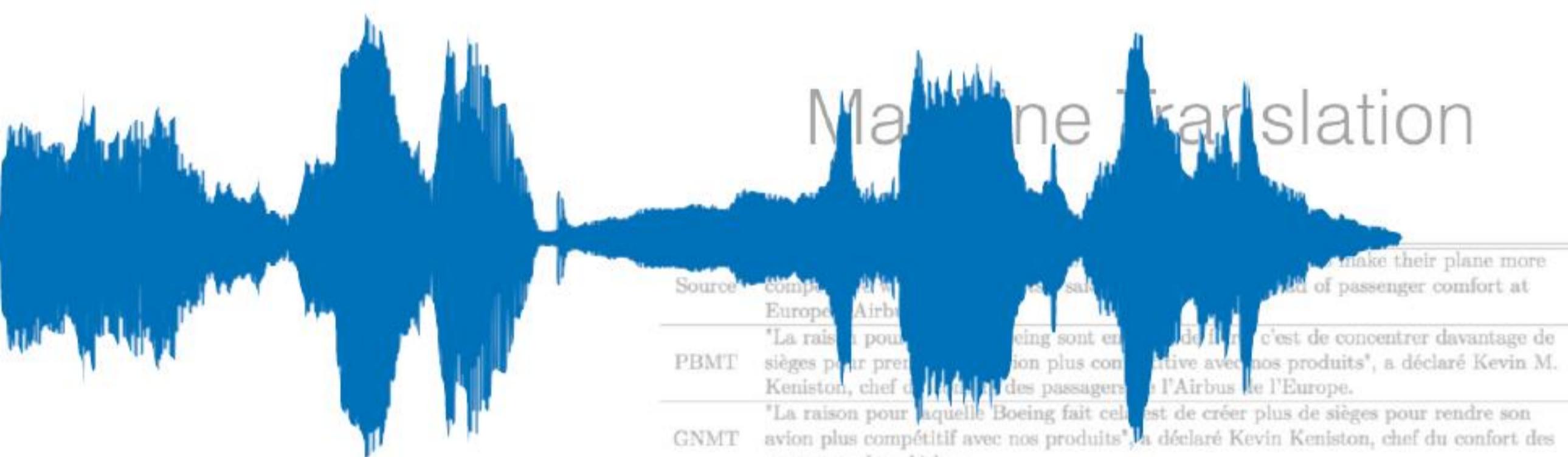
Figure 11. Qualitative results of Mask R-CNN on COCO test set, using ResNet-101 [28] and running at 5 FPS with 35.7 mask AP (Table 1).

## Visual Question Answering

Method	VQA v2 test dev			VQA v2 test			
	All	Youth	Adults	All	Youth	Adults	
prior (most common answer in training set) [1]	29.08	84.20	95.98	3.14	—	—	
LSTM Language only (base model) [1]	—	—	—	41.26	81.03	91.85	27.37
Deep LSTM Q (ours, 127) as reported in [1]	—	—	—	54.22	73.06	85.18	41.83
MCR [1] as reported in [1]	—	—	—	62.27	78.82	82.28	52.36
UPMC-LDNN [1]	—	—	—	63.71	81.17	87.17	47.17
LM-NUS	—	—	—	67.79	82.50	84.49	59.07
HDDU-SYD-UNCC	—	—	—	66.77	81.89	82.89	58.30
Proposed model	—	—	—	68.09	84.50	85.91	59.01

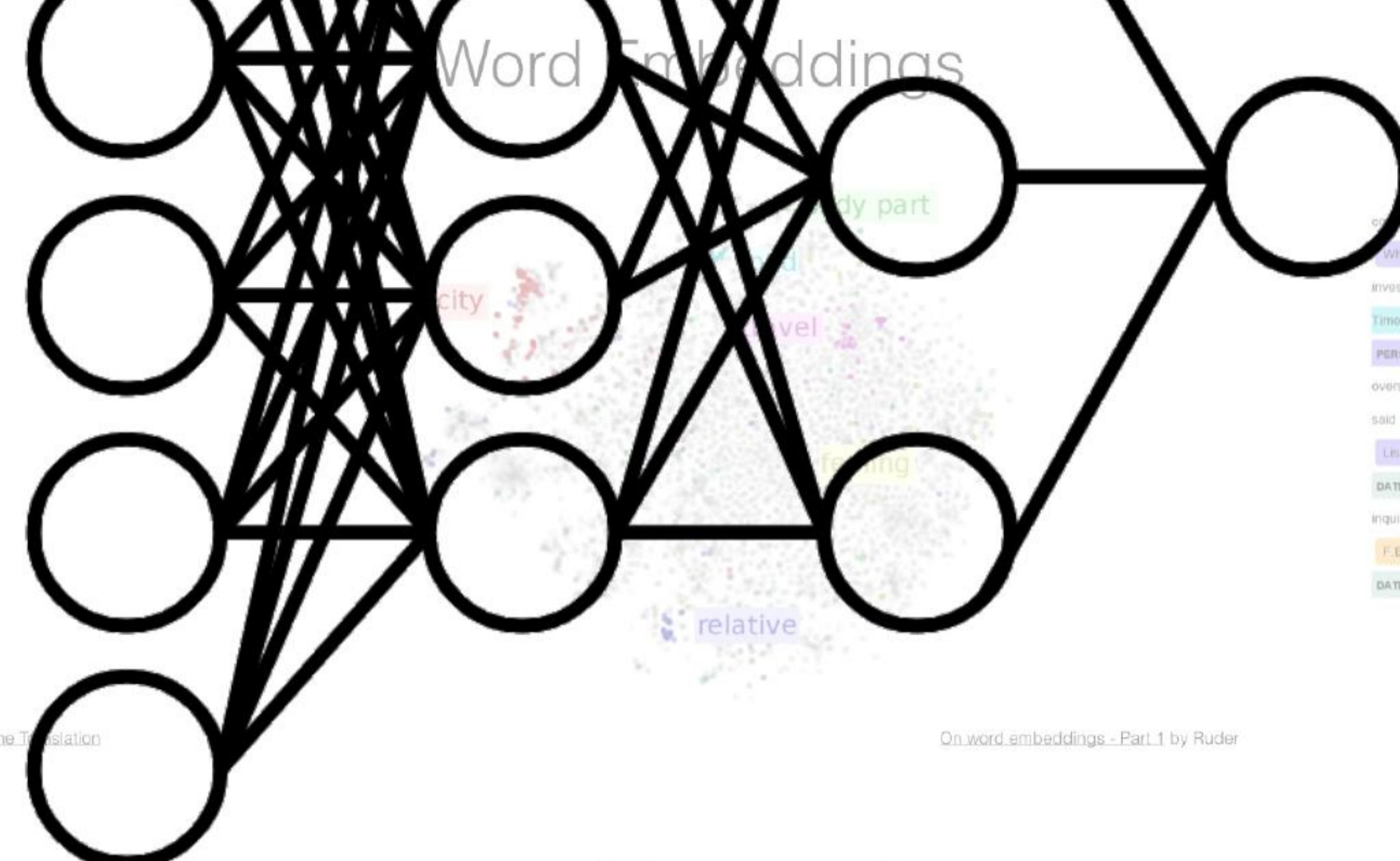
Table 3. Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

## Waveform



Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Word Embeddings



On word embeddings...Part\_1 by Ruder

## Named Entity

This is a supervised learning method

Named Entity Recognition and Classification with Scikit-Learn by Susan Li  
Esteves et al. Named Entity Recognition in Twitter using Images and Text

## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BIDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT <sub>BASE</sub> (Single)	80.8	88.5	-	-
BERT <sub>LARGE</sub> (Single)	84.1	90.9	-	-
BERT <sub>LARGE</sub> (Ensemble)	85.8	91.8	-	-
BERT <sub>LARGE</sub> (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
BERT <sub>LARGE</sub> (Ens.+TriviaQA)	86.2	92.2	87.4	93.2

Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

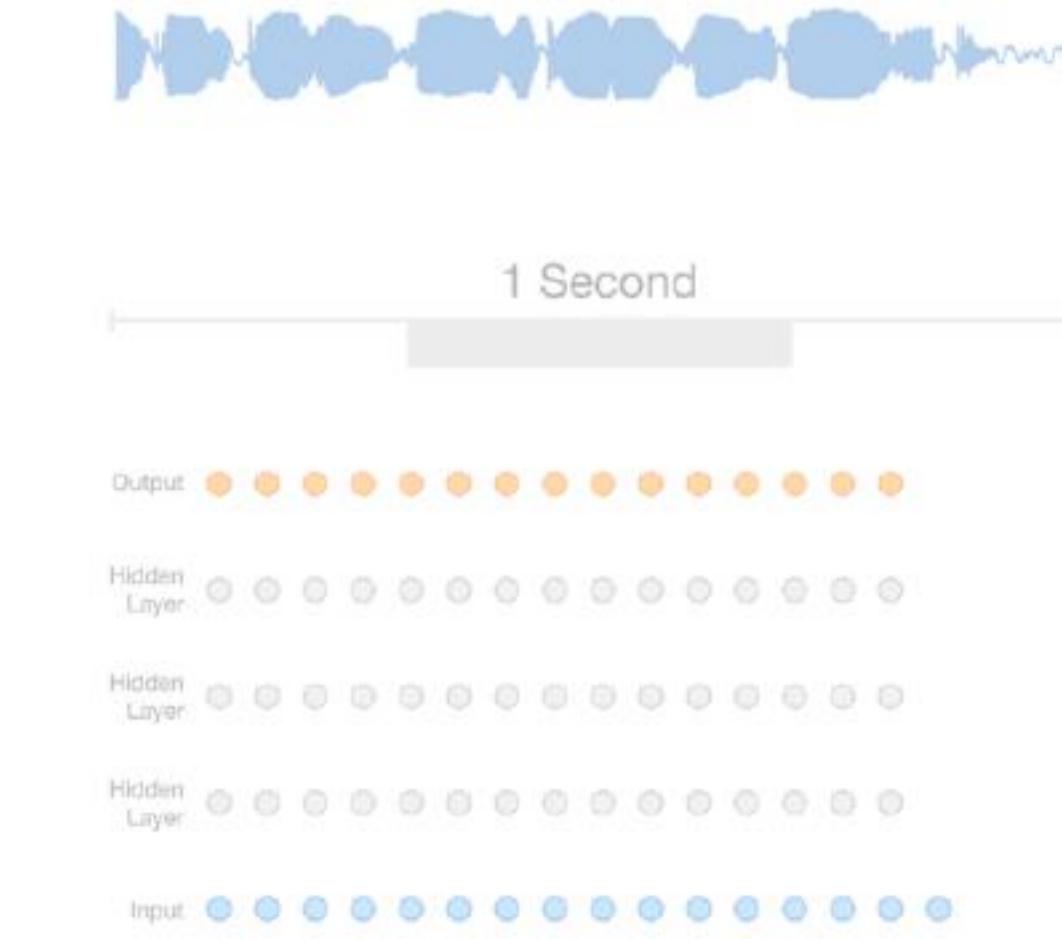
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## Speech Recognition

Senone set	Model/combination step	WER	WER	WER	WER
		devset	test	devset	test
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
27k-puhpum	BLSTM	11.3	8.0	9.2	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8
Confusion network combination		7.4	5.2		
+ LSTM rescoring		7.3	5.2		
+ ngram rescoring		7.2	5.2		
+ backchannel penalty		7.2	5.1		

Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

## Speech Synthesis (text-to-speech)



Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

# Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224x224	11.2M	1.94B	74.8	92.2
NASNet-A (5 @ 1538)	299x299	10.9M	2.35B	78.6	94.2
Inception V3 [60]	299x299	23.8M	5.72B	78.8	94.4
Xception [9]	299x299	22.8M	8.38B	79.0	94.5
Inception ResNet V2 [58]	299x299	55.8M	13.2B	80.1	95.1
NASNet-A (7 @ 1920)	299x299	22.6M	4.93B	80.8	95.3
ResNeXt-101 (64 x 4d) [68]	320x320	83.6M	31.5B	80.9	95.6
PolyNet [69]	331x331	92M	34.7B	81.3	95.8
DPN-131 [8]	320x320	79.5M	32.0B	81.5	95.8
SENet [25]	320x320	145.8M	42.3B	82.7	96.2
NASNet-A (6 @ 4032)	331x331	88.9M	23.8B	82.7	96.2

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image size reported in the table. Model size for [25] calculated from open-source implementation.

## Instance Segmentation

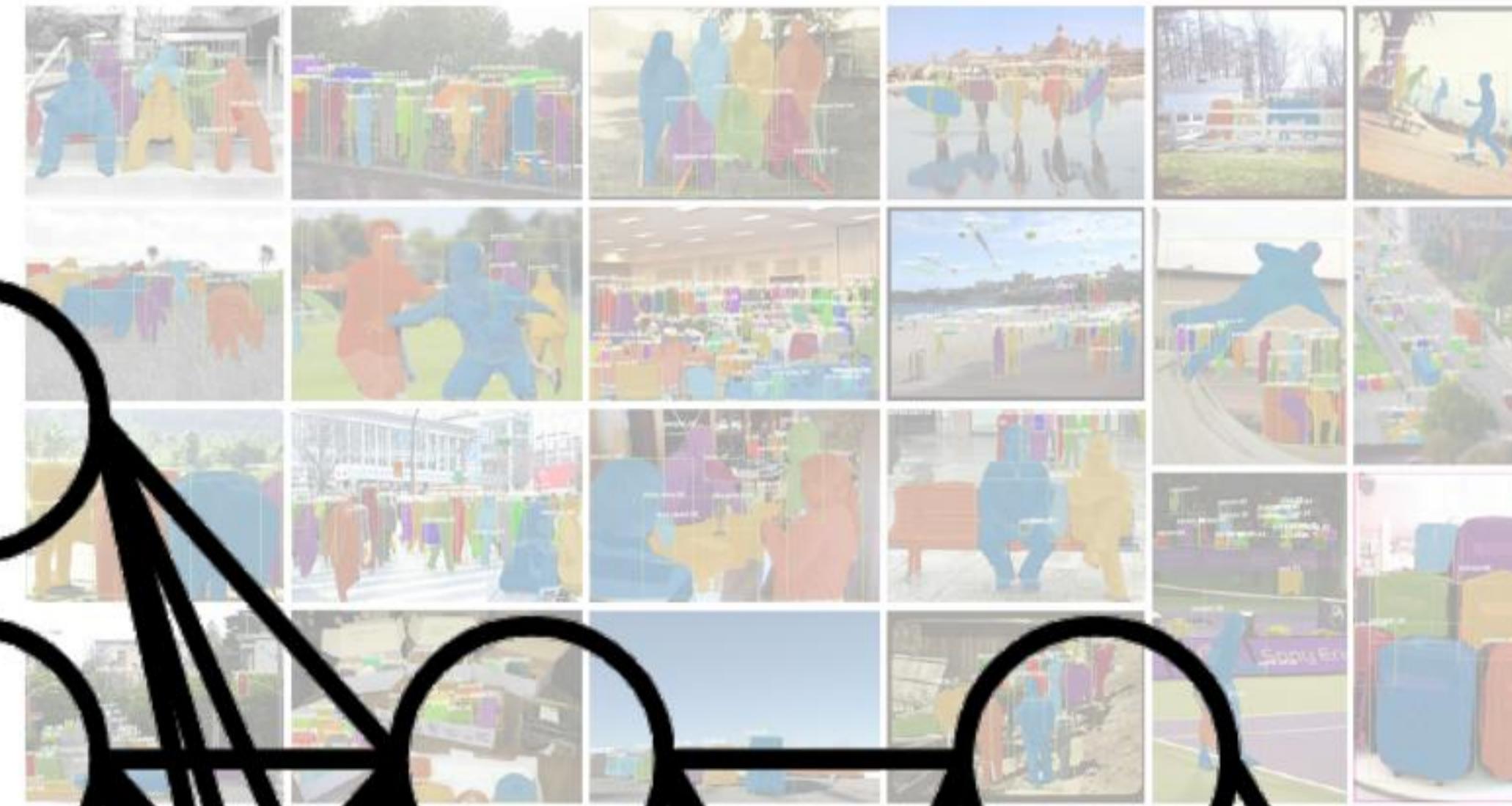


Figure 1. Qualitative results of Mask R-CNN on COCO test set, trained using ResNet-101 [28] and running at 5 FPS with 35.7 mask AP (Table 1).

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Method	VQA v2 test dev			VQA v2 test		
	All	Youth	Adults	All	Youth	Adults
prior (most common answer in training set) [1]	—	—	—	29.08	94.20	95.98
LSTM Language only (base model) [1]	—	—	—	41.26	91.01	91.85
Deep LSTM Q (ours, 127) as reported in [1]	—	—	—	54.22	73.06	85.18
Deep LSTM Q (ours, 127) as reported in [1]	—	—	—	62.27	78.82	82.28
MCR [1] (as reported in [1])	—	—	—	63.71	81.17	87.17
UPMC-LDNN [1]	—	—	—	67.79	82.50	84.49
LM-NUS	—	—	—	66.77	81.89	82.89
HDDU-SYD-UNCC	—	—	—	68.09	94.50	85.91
Proposed model	—	—	—	68.09	94.50	85.91
ResNet features 7x7, single network	62.07	79.20	89.46	82.82	82.27	89.71
Image features from bottom-up attention, adaptive K, single network	65.52	81.62	84.27	65.52	80.30	83.00
ResNet features 7x7, ensemble	66.31	81.29	83.17	81.10	84.23	85.20
Image features from bottom-up attention, adaptive K, ensemble	68.87	86.08	88.99	86.80	76.34	86.64

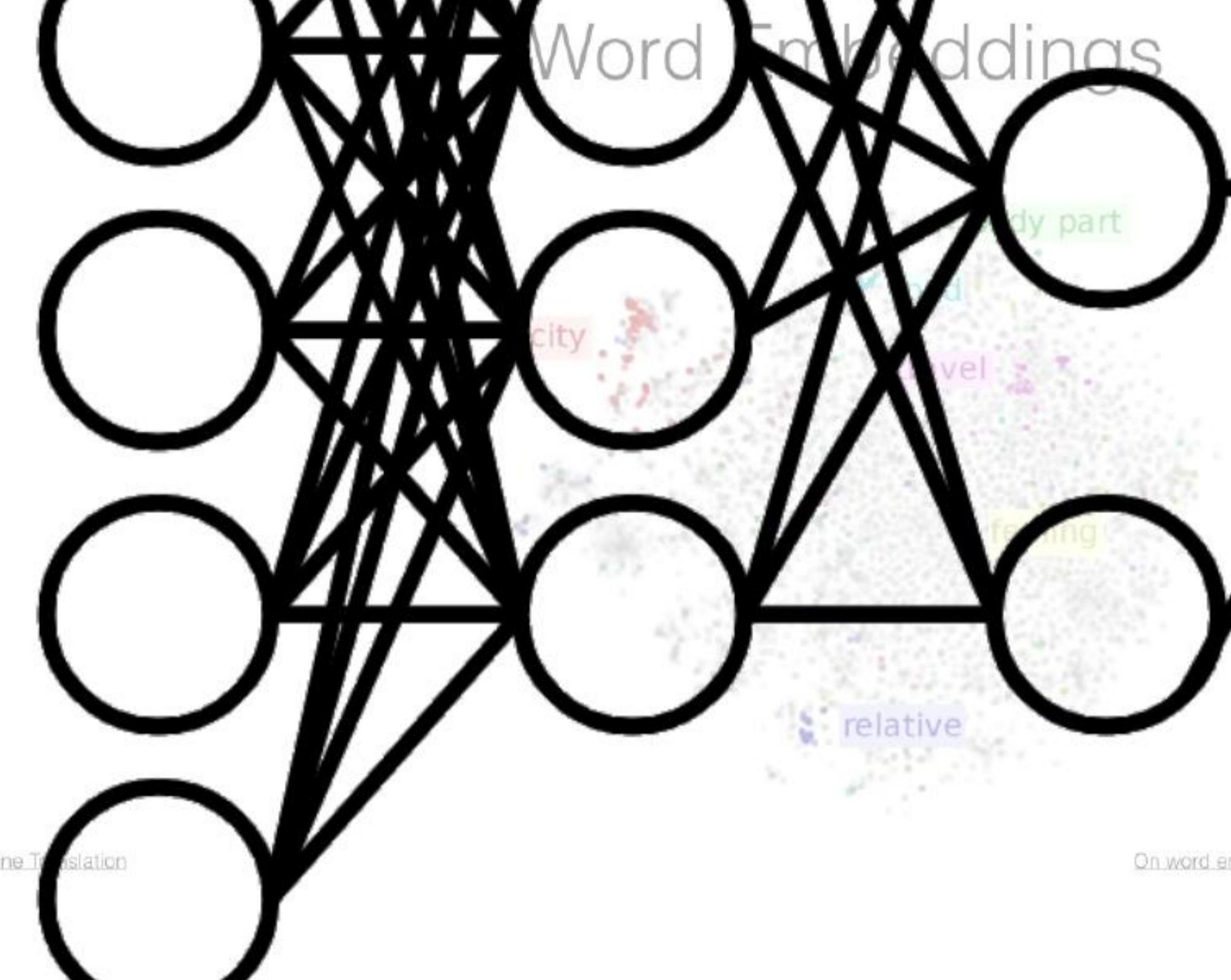
Table 3. Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

## Machine Translation

Input		
Source	"The rose competitive Europe's "La raison sièges pour Keniston, "La raison avion plus avion plus passagers chez Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs Human Source	"heir plane more ger confort at danser davantage de déclaré Kevin M. 3.0 leur rendre son chef du confort des 6.0 "Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs par rapport à nos produits", a déclaré Kevin Keniston, directeur de Confort Passager 6.0 chez l'avionneur Airbus. When asked about this, un officiel de l'American administration replied: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington." PBMT Etats-Unis n'est pas effectué une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington". Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington". Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".
PBMT	"La raison sièges pour Keniston, "La raison avion plus avion plus passagers chez Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs Human Source	"heir plane more ger confort at danser davantage de déclaré Kevin M. 3.0 leur rendre son chef du confort des 6.0 "Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs par rapport à nos produits", a déclaré Kevin Keniston, directeur de Confort Passager 6.0 chez l'avionneur Airbus. When asked about this, un officiel de l'American administration replied: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington." PBMT Etats-Unis n'est pas effectué une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington". Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington". Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".
GNMT	"The rose competitive Europe's "La raison sièges pour Keniston, "La raison avion plus avion plus passagers chez Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs Human Source	"heir plane more ger confort at danser davantage de déclaré Kevin M. 3.0 leur rendre son chef du confort des 6.0 "Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs par rapport à nos produits", a déclaré Kevin Keniston, directeur de Confort Passager 6.0 chez l'avionneur Airbus. When asked about this, un officiel de l'American administration replied: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington." PBMT Etats-Unis n'est pas effectué une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington". Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington". Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".
Human	"The rose competitive Europe's "La raison sièges pour Keniston, "La raison avion plus avion plus passagers chez Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs Human Source	"heir plane more ger confort at danser davantage de déclaré Kevin M. 3.0 leur rendre son chef du confort des 6.0 "Boeing fait ça pour pouvoir essayer plus de sièges et rendre ses avions plus compétitifs par rapport à nos produits", a déclaré Kevin Keniston, directeur de Confort Passager 6.0 chez l'avionneur Airbus. When asked about this, un officiel de l'American administration replied: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington." PBMT Etats-Unis n'est pas effectué une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington". Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington". Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

Wu et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation



On word embeddings...Part\_1 by Ruder

## Named Entity Recognition

Output

Named Entity Recognition and Classification with Scikit-Learn by Susan Li  
Esteves et al. Named Entity Recognition in Twitter using Images and Text

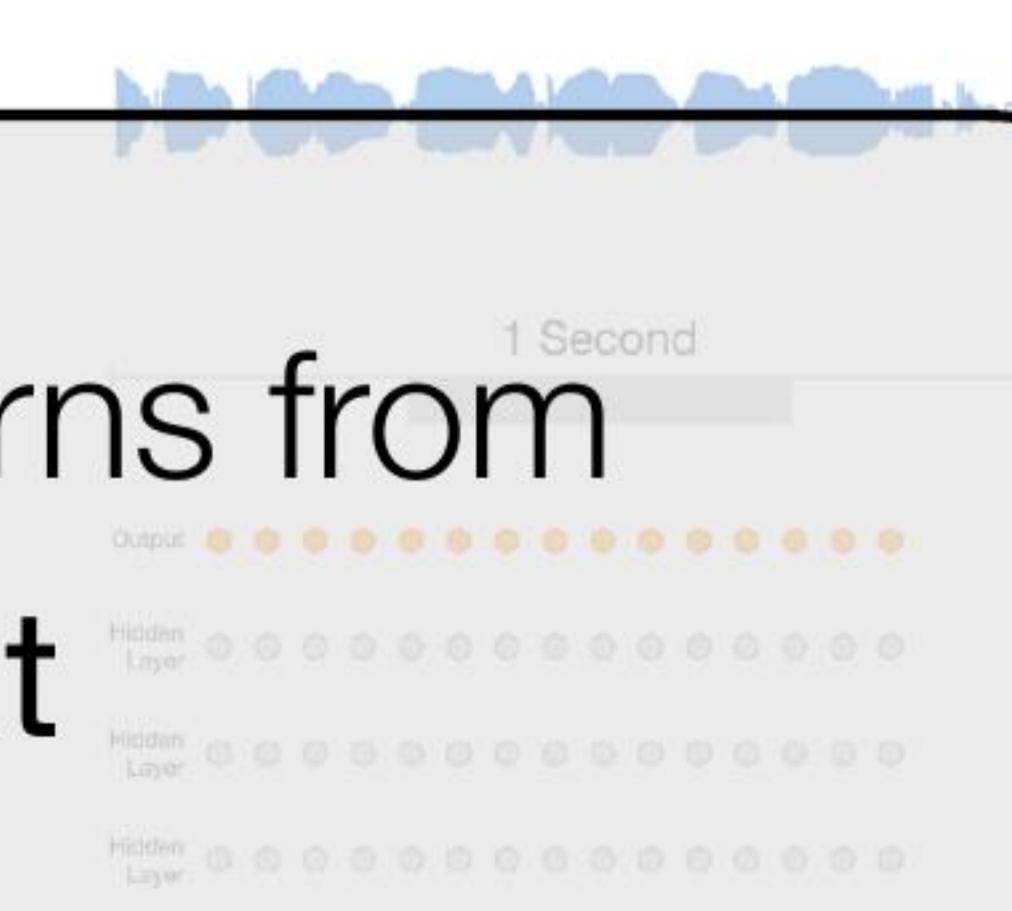
## Question Answering

Leaderboard (Oct 8th, 2018)				
	Dev	Test	EM	F1
<b>Input Question:</b>				
Where do water droplets collide with ice crystals to form precipitation?	-	-	86.0	91.7
<b>Input Paragraph:</b>				
... Precipitation forms as smaller droplets coalesce via collisions with other rain drops or ice crystals within a cloud. ...	-	-	84.5	90.5
<b>Output Answer:</b>				
within a cloud	-	-	83.5	90.1
	-	-	82.5	89.3

The model has no control over the dataset it learns from  
Ground truth output can be specified given input

Senone set	Model/combination step	WER devset	WER test	WER devset	WER test
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.2
9k-pulpm	BLSTM+LSTM+LACE+CNN	11.3	7.8	9.7	6.1
9k-pulpm	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-pulpm	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8

Xiong et al. The Microsoft 2017 Conversational Speech Recognition System



Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

Leaderboard (Oct 8th, 2018)				
	Dev	Test	EM	F1
<b>Input Question:</b>				
What is the capital of France?	80.8	88.5	-	-
<b>Input Paragraph:</b>				
Paris is the capital city of France. It is located in the northern part of the country, on the Seine River. Paris is known for its historical landmarks, such as the Eiffel Tower and the Louvre Museum. It is also a major center for fashion, art, and culture.	81.2	87.9	84.1	90.3
<b>Output Answer:</b>				
Paris	86.2	92.2	87.4	93.2

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

# Robot Learning via Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	299×299	10.9 M	2.35 B	78.6	94.2
Inception V3 [60]	299×299	23.8 M	5.72 B	78.8	94.4
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	299×299	22.6 M	4.93 B	80.8	95.3
ResNeXt-101 (64 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	320×320	145.8 M	42.3 B	82.7	96.2
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	96.2

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiply-accumulate operations are calculated for the image size reported in the table. Model size for [25] calculated from open-source implementation.

## Instance Segmentation



Figure 10: Results of Mask R-CNN on COCO test set, using ResNet-101 [27] and running at 5 FPS with 35.7 mask AP (Table 1).

## Visual Question Answering

Method	VQA v2 test dev			VQA v2 test		
	All	Youth	Adults	All	Youth	Adults
prior (most common answer in training set) [1]	—	—	—	29.08	84.20	92.98
LSTM Language only (base model) [11]	—	—	—	41.26	87.03	91.85
Deep LSTM (ours) [27] as reported in [1]	—	—	—	54.22	73.06	85.18
MCB [1] as reported in [1]	—	—	—	62.27	78.82	82.28
CPMC-LSTM [1]	—	—	—	63.71	81.17	87.17
LM-NUS	—	—	—	67.70	82.50	84.49
HICO-SYD-UNCC	—	—	—	66.77	81.89	86.26
Proposed model	—	—	—	68.09	84.50	85.91

Table 3. Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 11: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 12: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 13: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 14: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 15: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 16: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 17: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 18: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 19: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 20: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 21: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 22: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 23: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 24: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 25: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 26: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 27: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 28: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 29: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 30: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 31: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 32: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 33: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 34: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 35: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 36: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 37: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 38: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 39: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 40: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 41: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 42: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 43: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 44: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 45: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 46: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 47: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 48: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 49: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 50: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 51: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

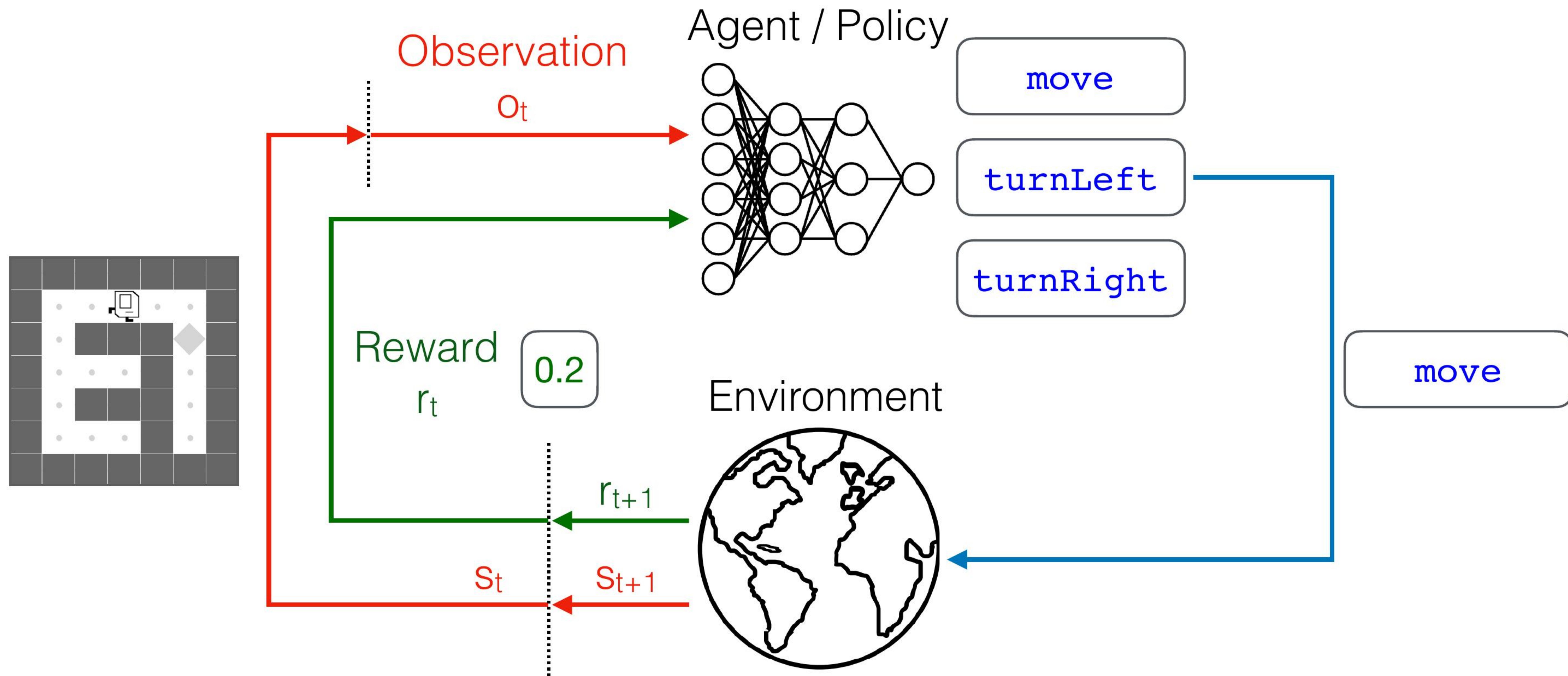
Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 52: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

Tanay et al. Tips and Tricks for Visual Question Answering: Learnings from the VQA v2 Leaderboard

Figure 53: Comparison of our best model with competing methods. Excerpt from the official VQA v2 Leaderboard [1].

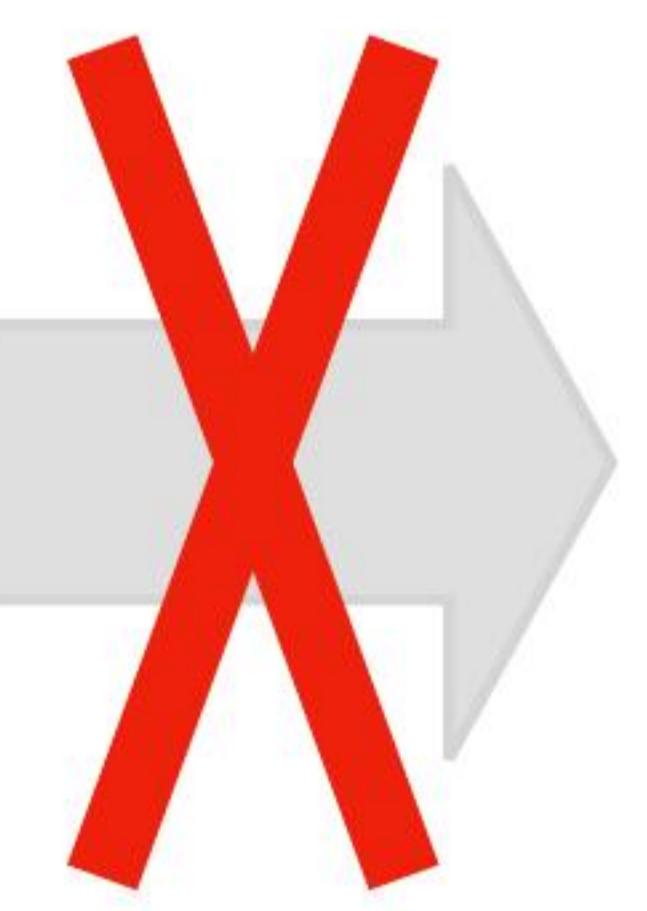
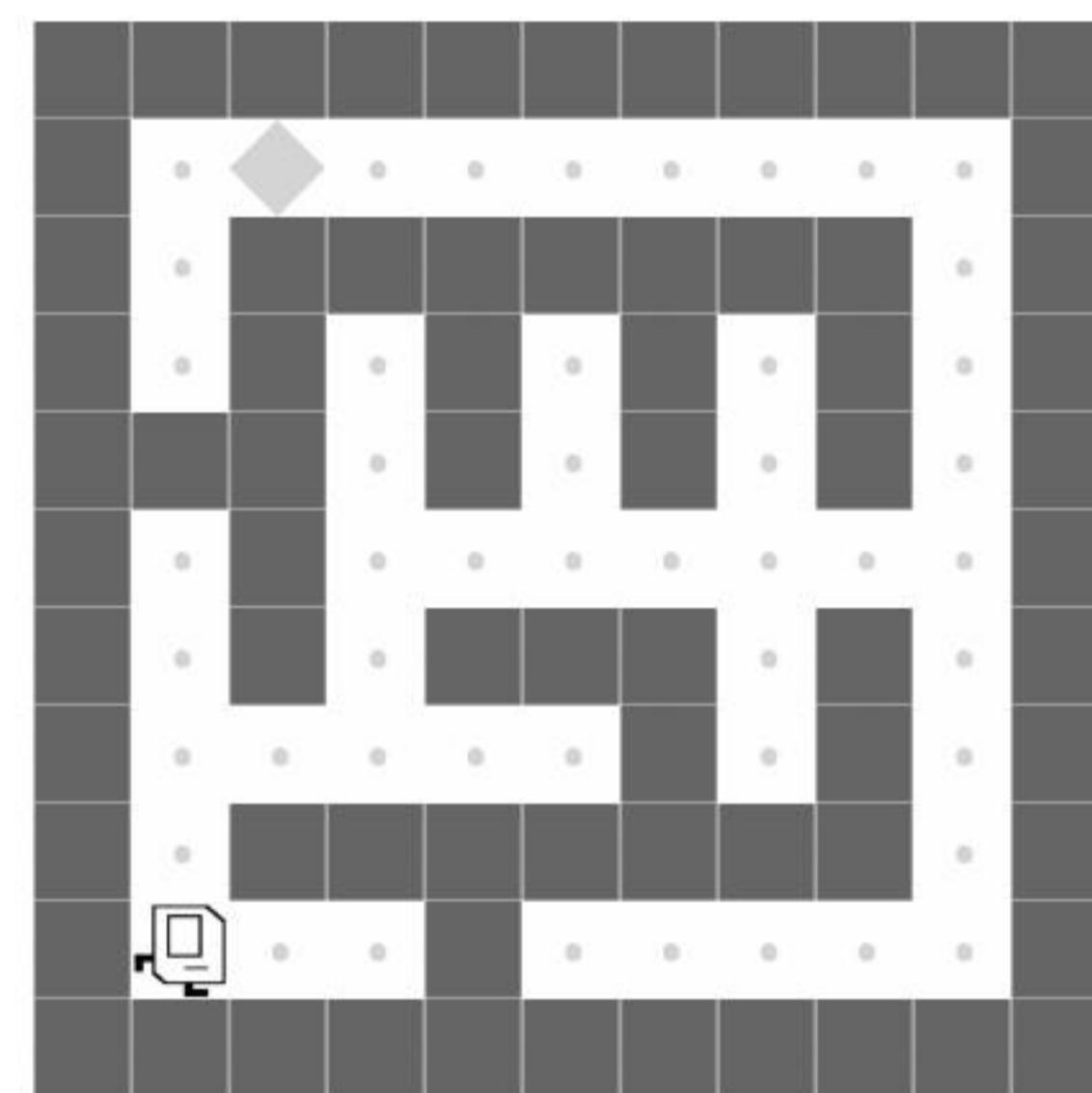
# Robot Learning via Deep Reinforcement Learning



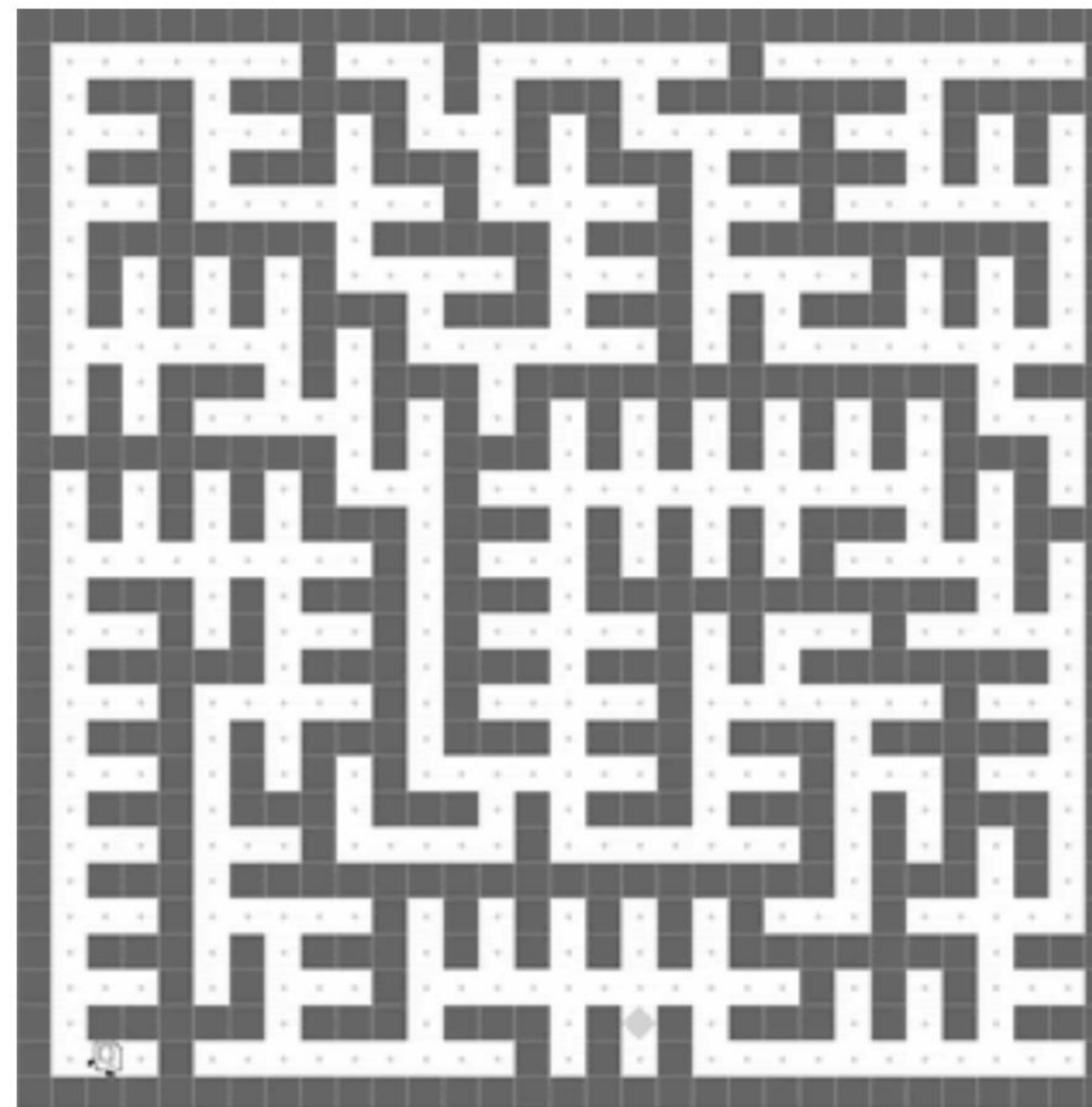
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Deep Reinforcement Learning - Issues

## Generalization



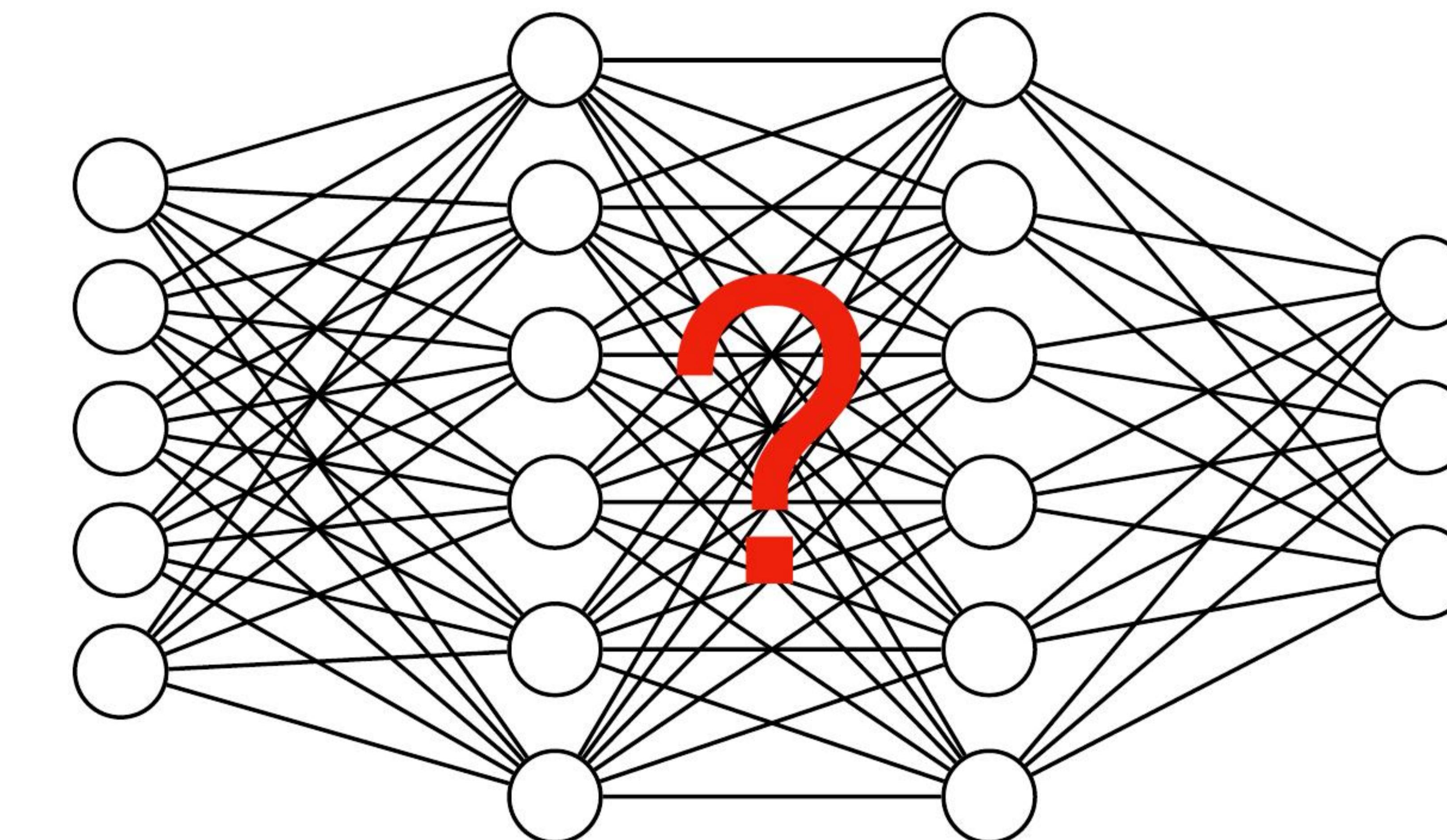
Simple task



Complex task


## Interpretability

*Trust, Safety, and Contestability*



Deep neural network policy

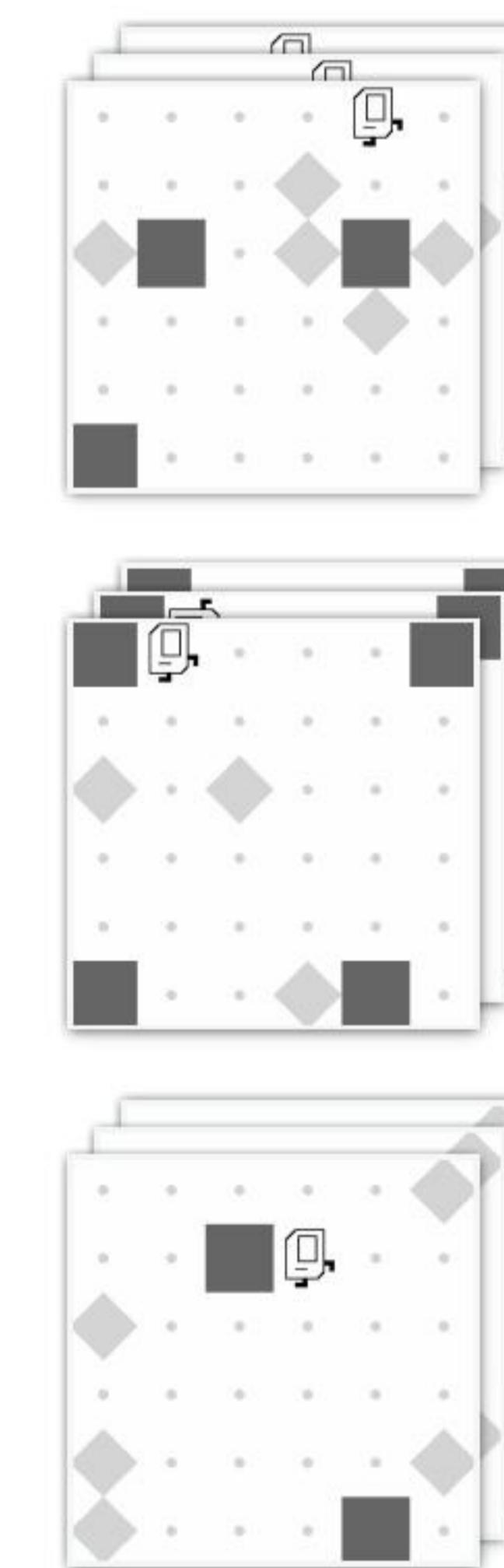
# Program as Reinforcement Learning Policies



# Neural Program Synthesis from Diverse Demonstration Videos

ICML 2018

Demonstrations      Program Policy      Execution

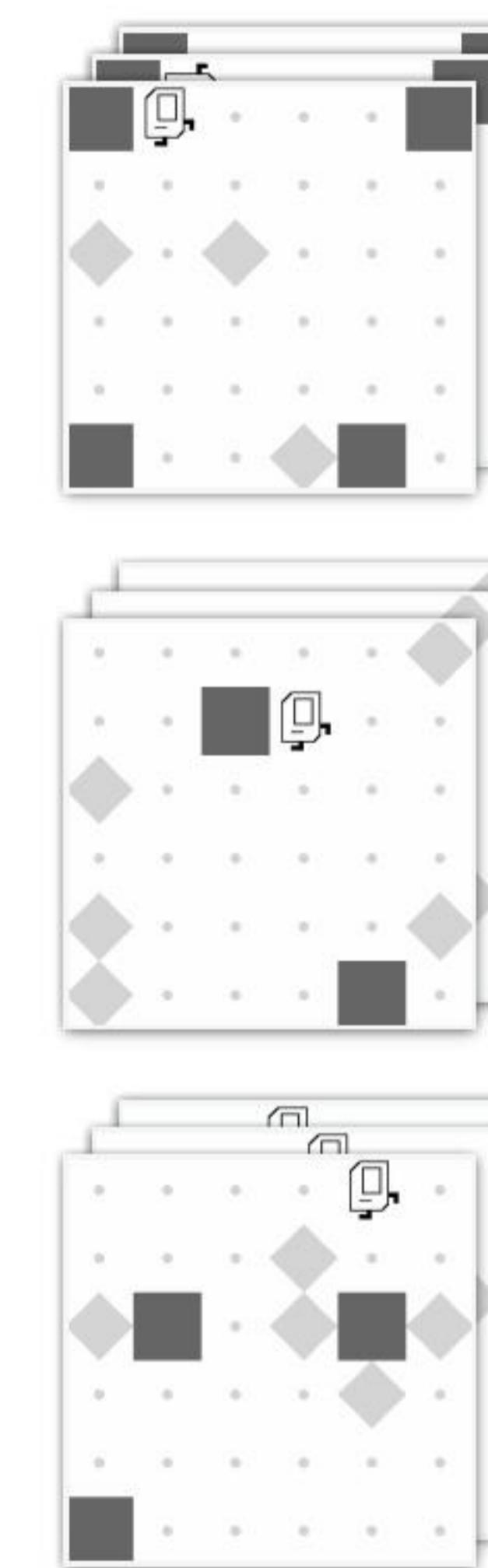
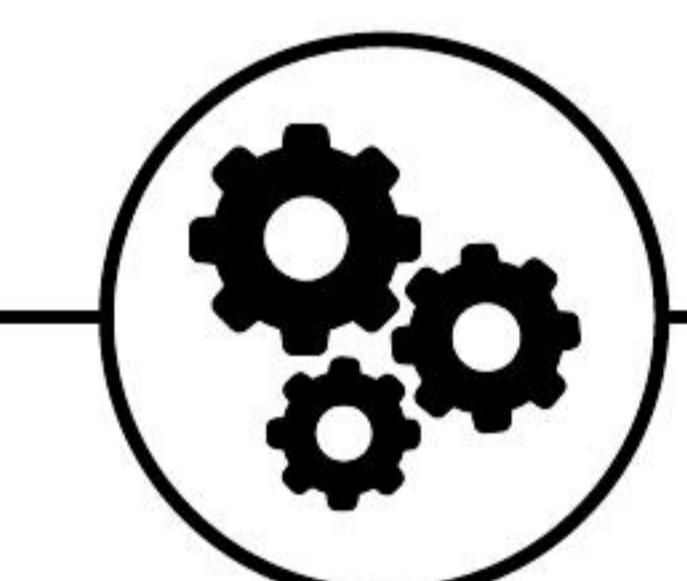


Synthesize

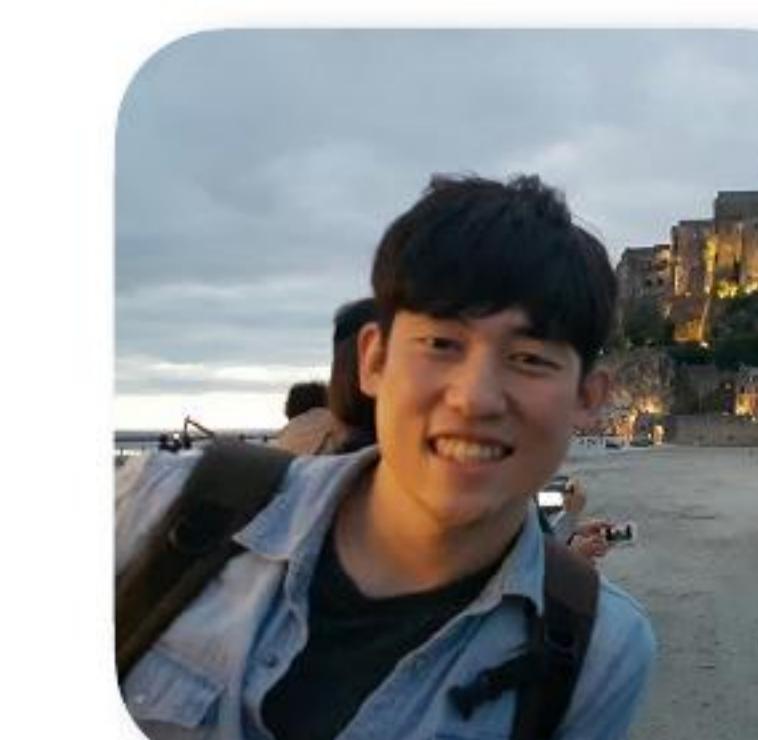


```
DEF run()
  IF frontIsClear
    move
  ELSE
    turnLeft
    move
    turnLeft
  REPEAT(2)
    turnRight
    putMarker
```

Execute



Shao-Hua Sun



Hyeonwoo Noh

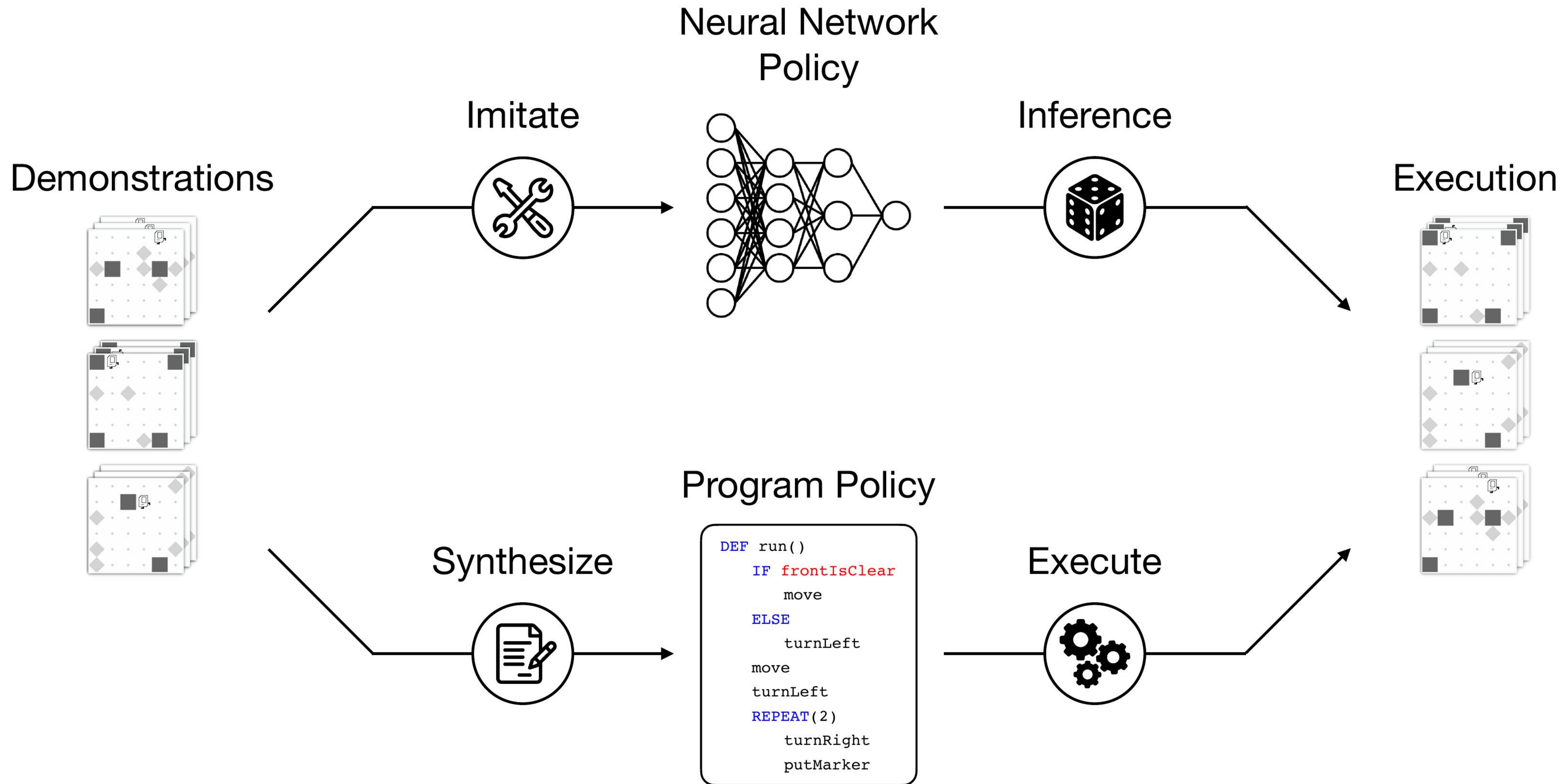


Sriram Somasundaram



Joseph J. Lim

# Imitation Learning via Synthesizing Programs

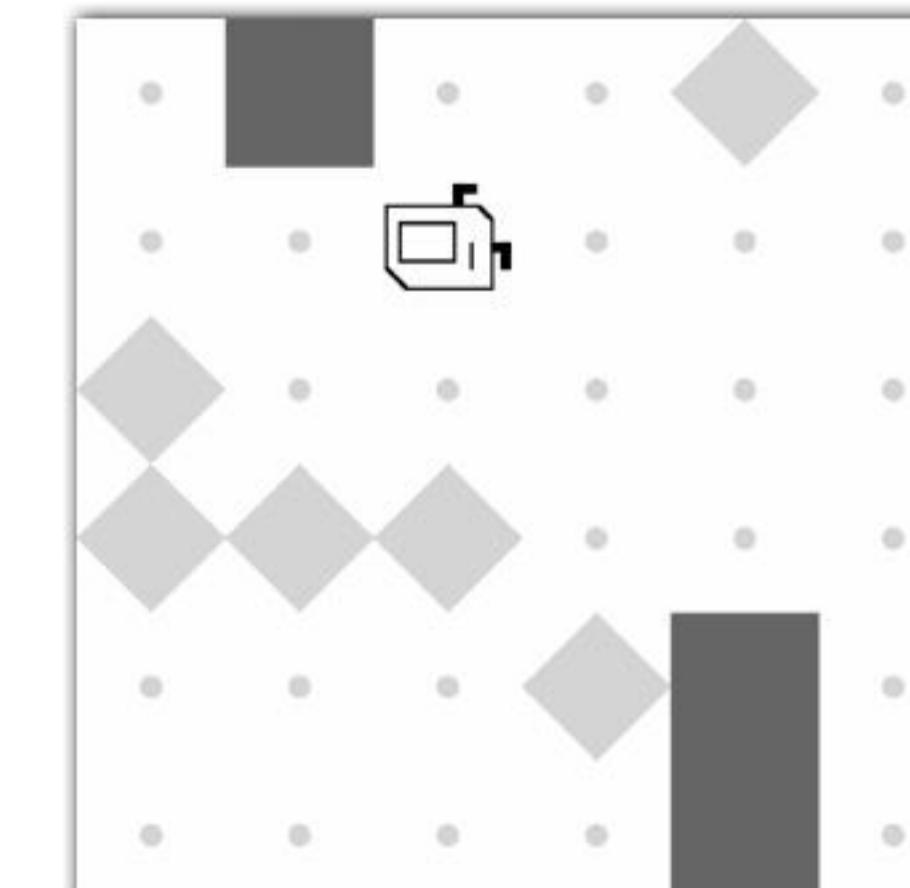
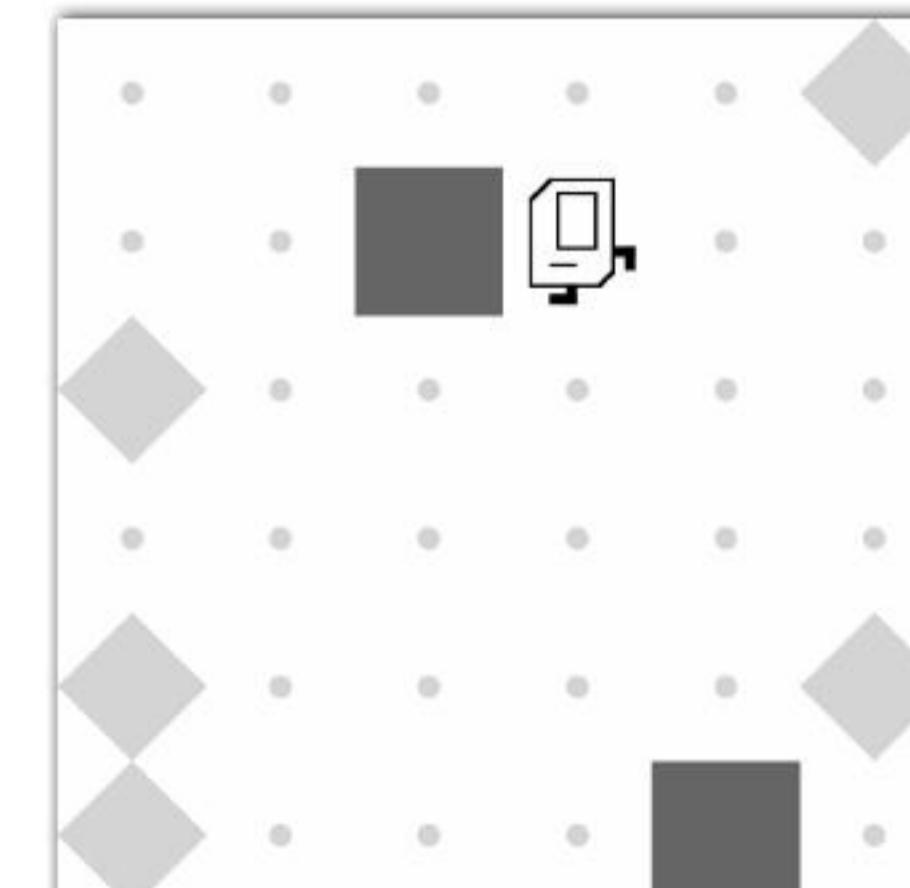
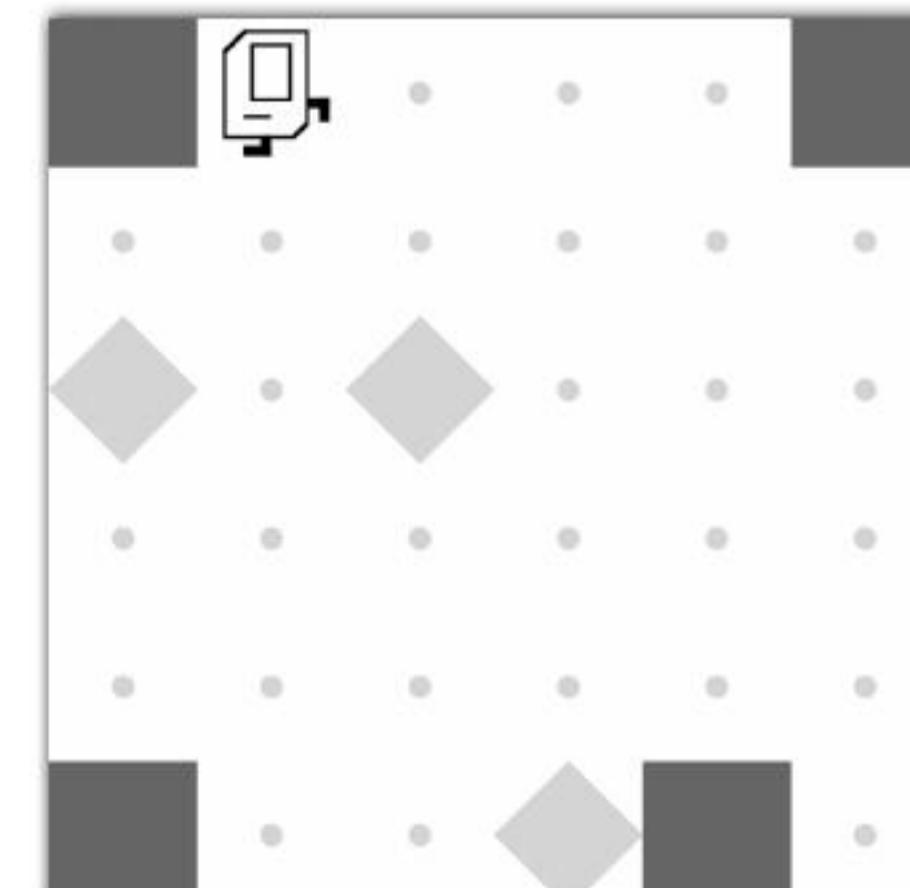
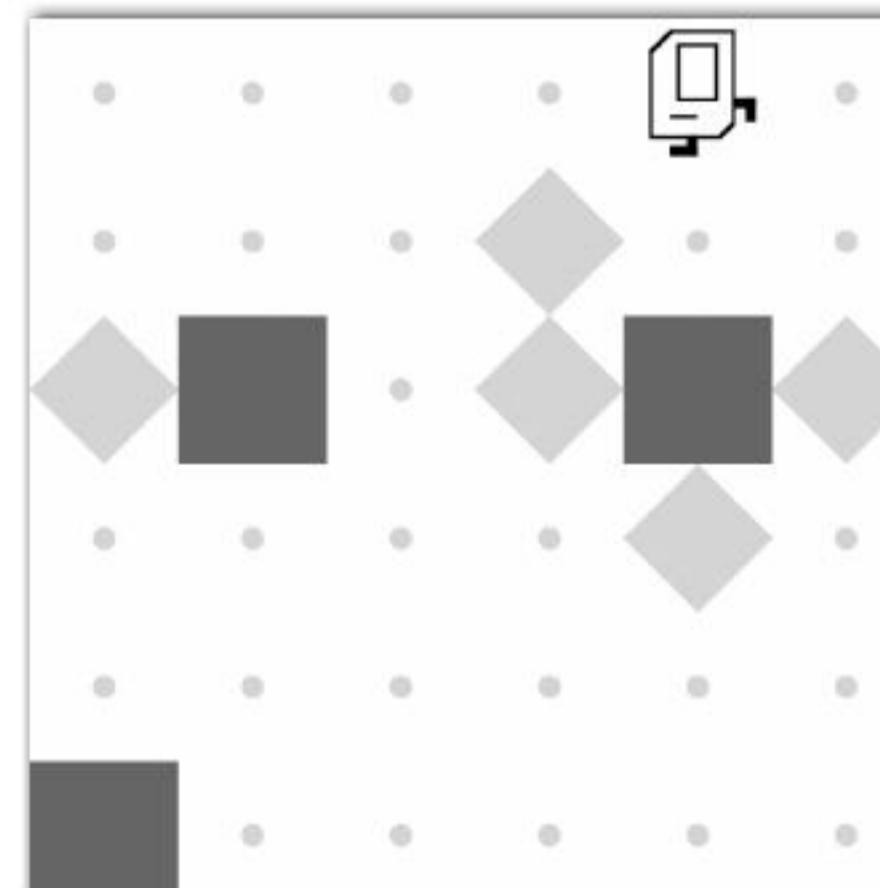


# Environments

Karel

Program

```
DEF run()
  IF frontIsClear
    move
  ELSE
    turnLeft
    move
    turnLeft
  REPEAT(2)
    turnRight
    putMarker
```



ViZDoom

Program

```
DEF run()
  WHILE frontIsClear(HellKnight)
    attack
    moveForward
  IF thereIs(Demon)
    moveRight
  ELSE
    moveLeft
    moveBackward
```

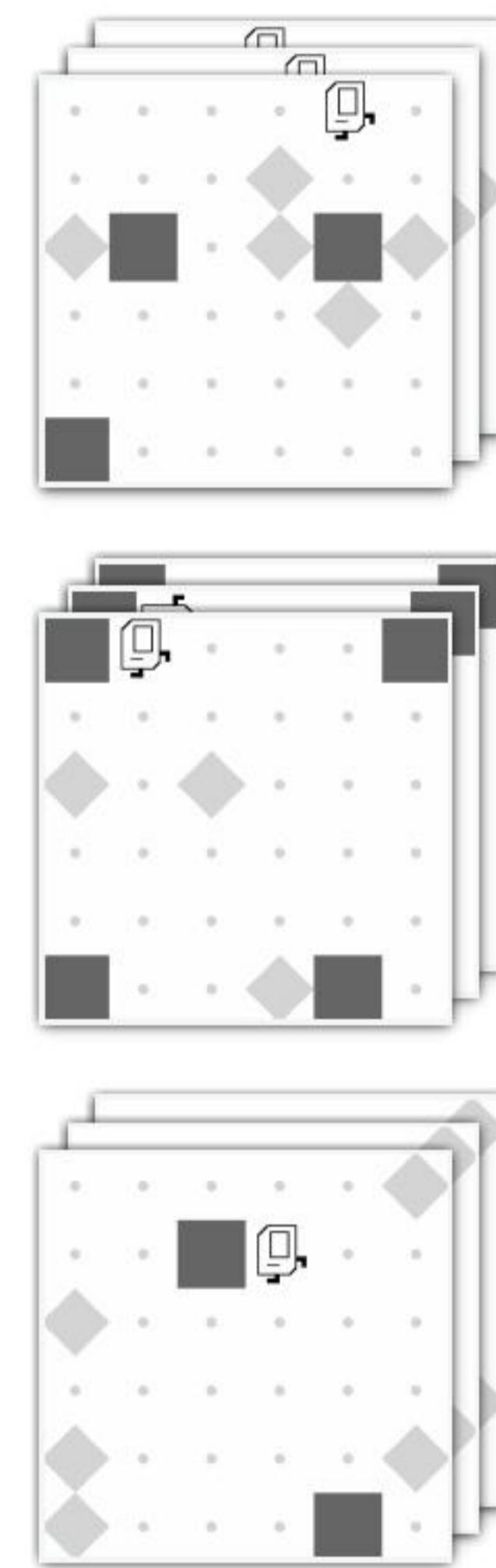


Richard E Pattis. "Karel the robot: a gentle introduction to the art of programming." John Wiley & Sons, Inc., 1981

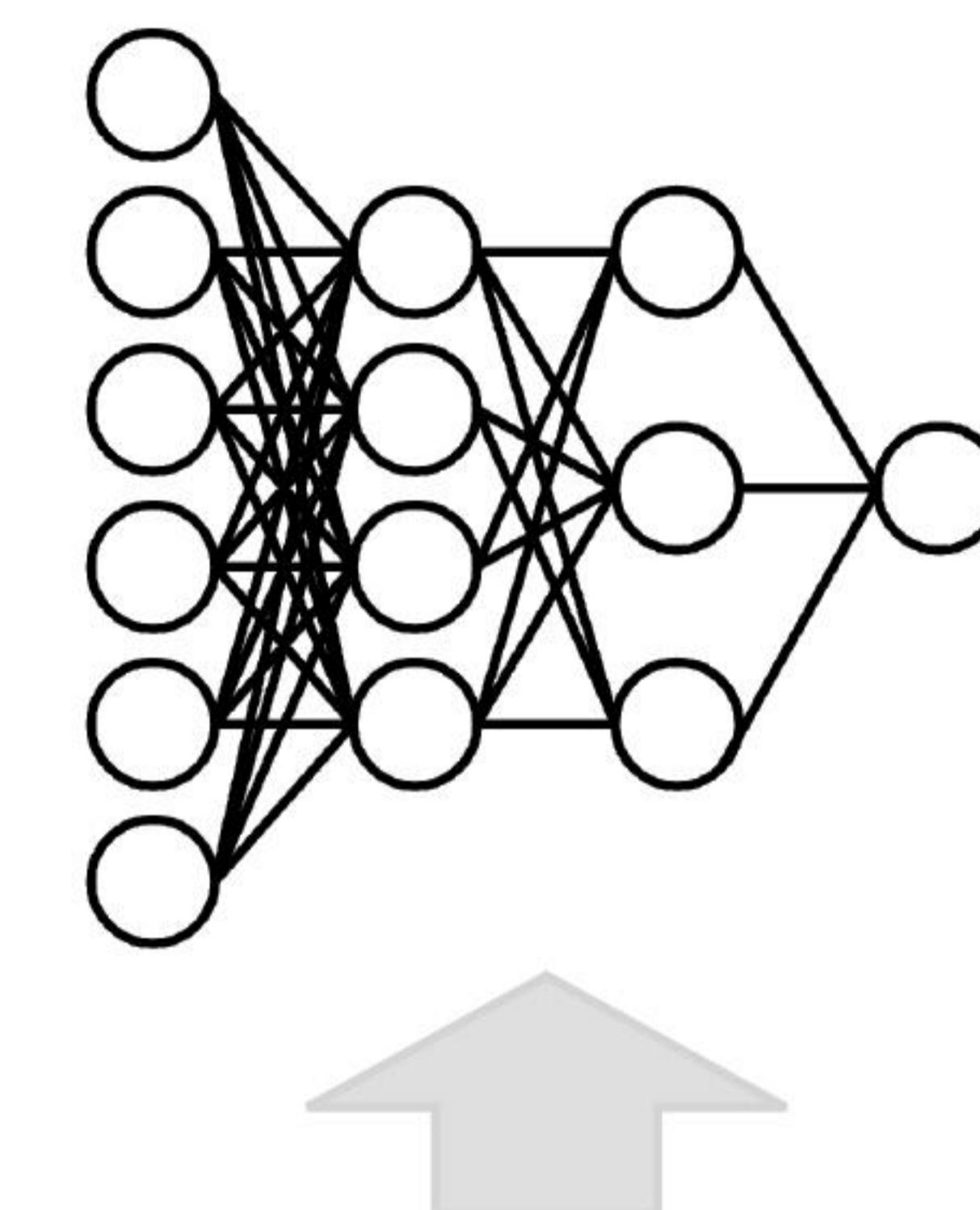
Kempka et al., "Vizdoom: A doom-based ai research platform for visual reinforcement learning." in CIG, 2016

# Imitation Learning with Neural Network Policy

Demonstrations



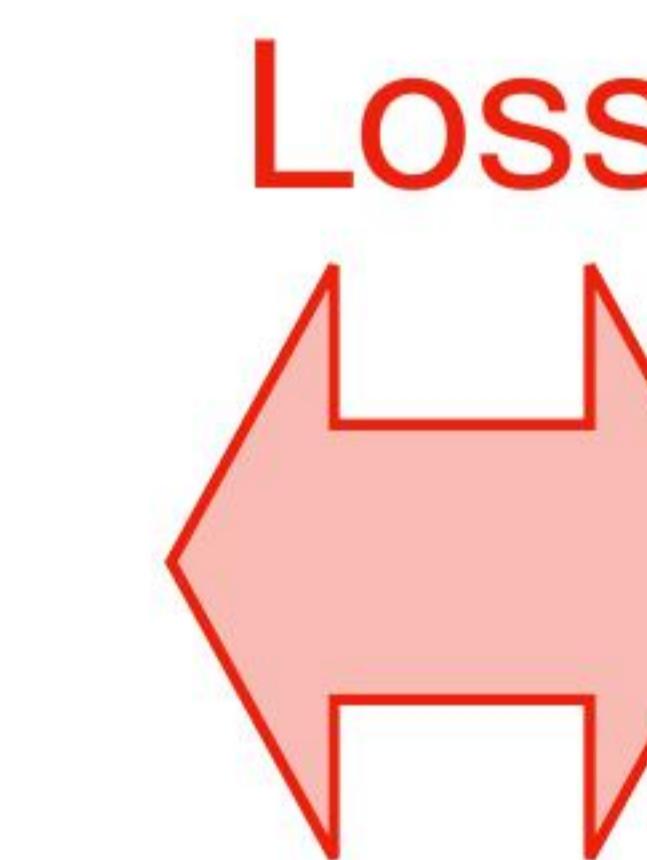
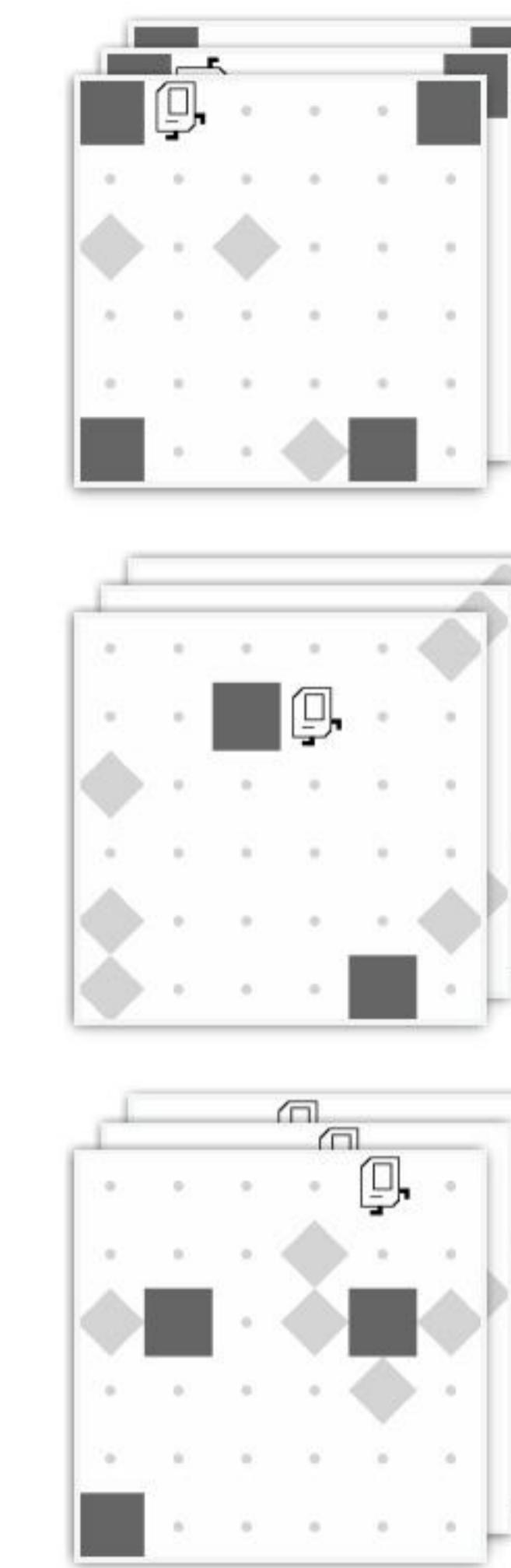
Neural Network  
Policy



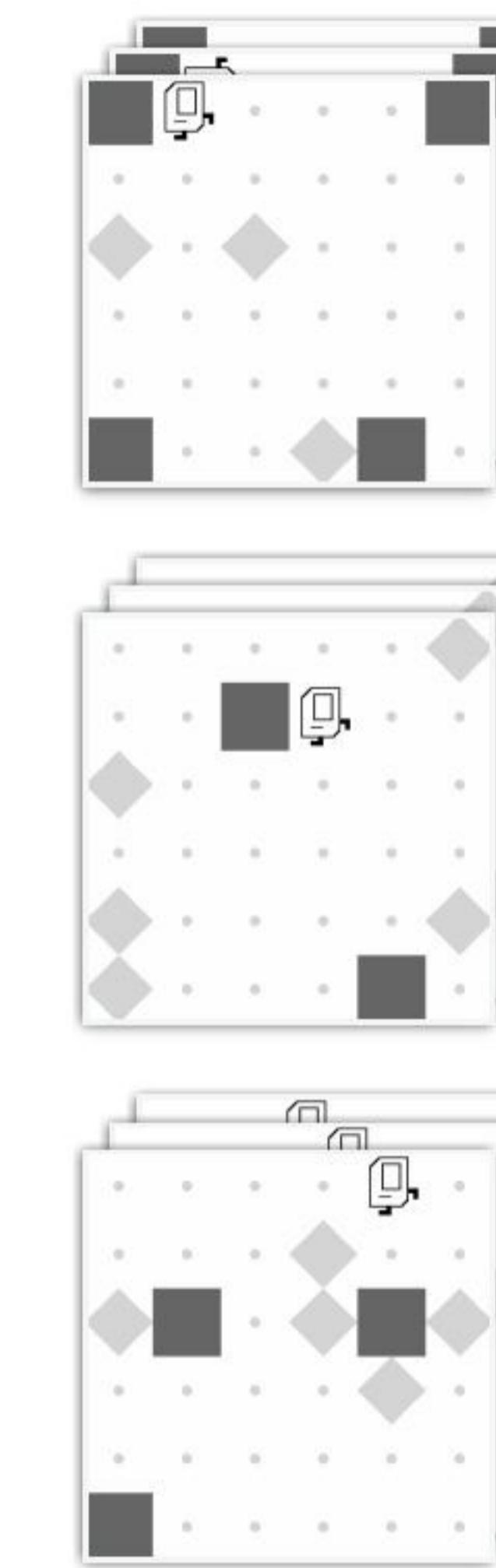
Infer



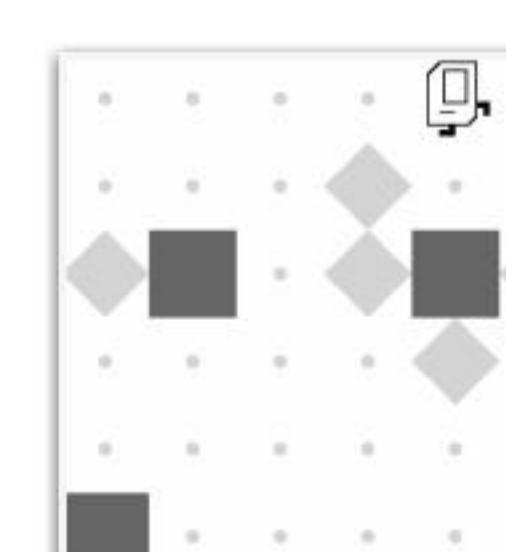
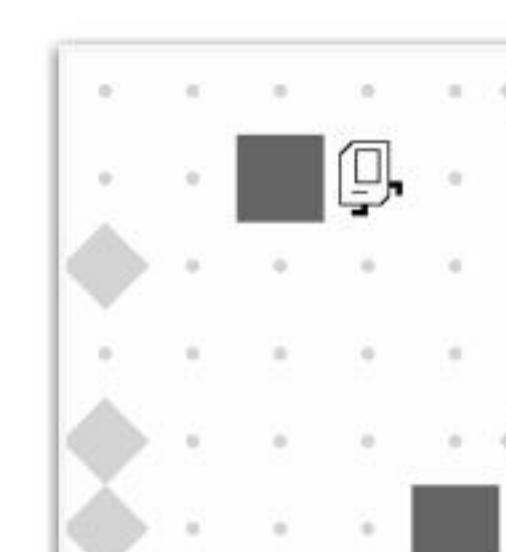
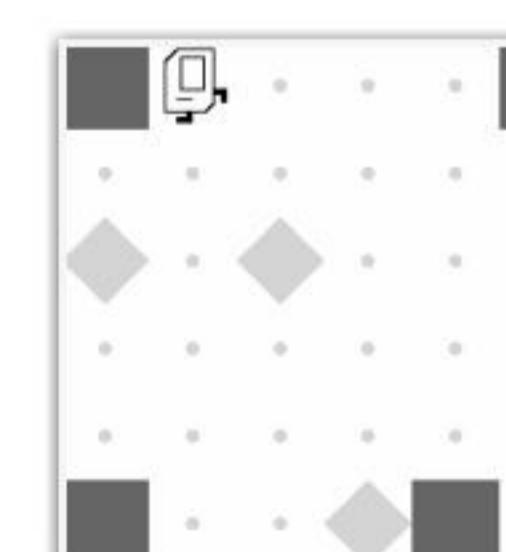
Predicted Execution



Ground Truth Execution

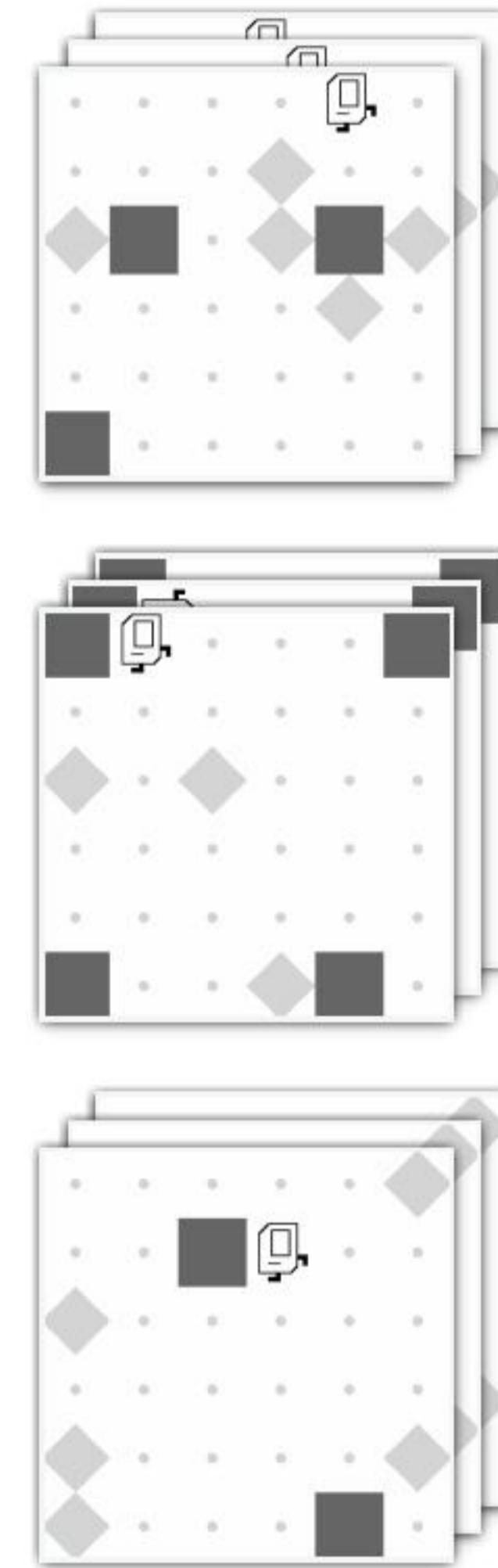


Initial States

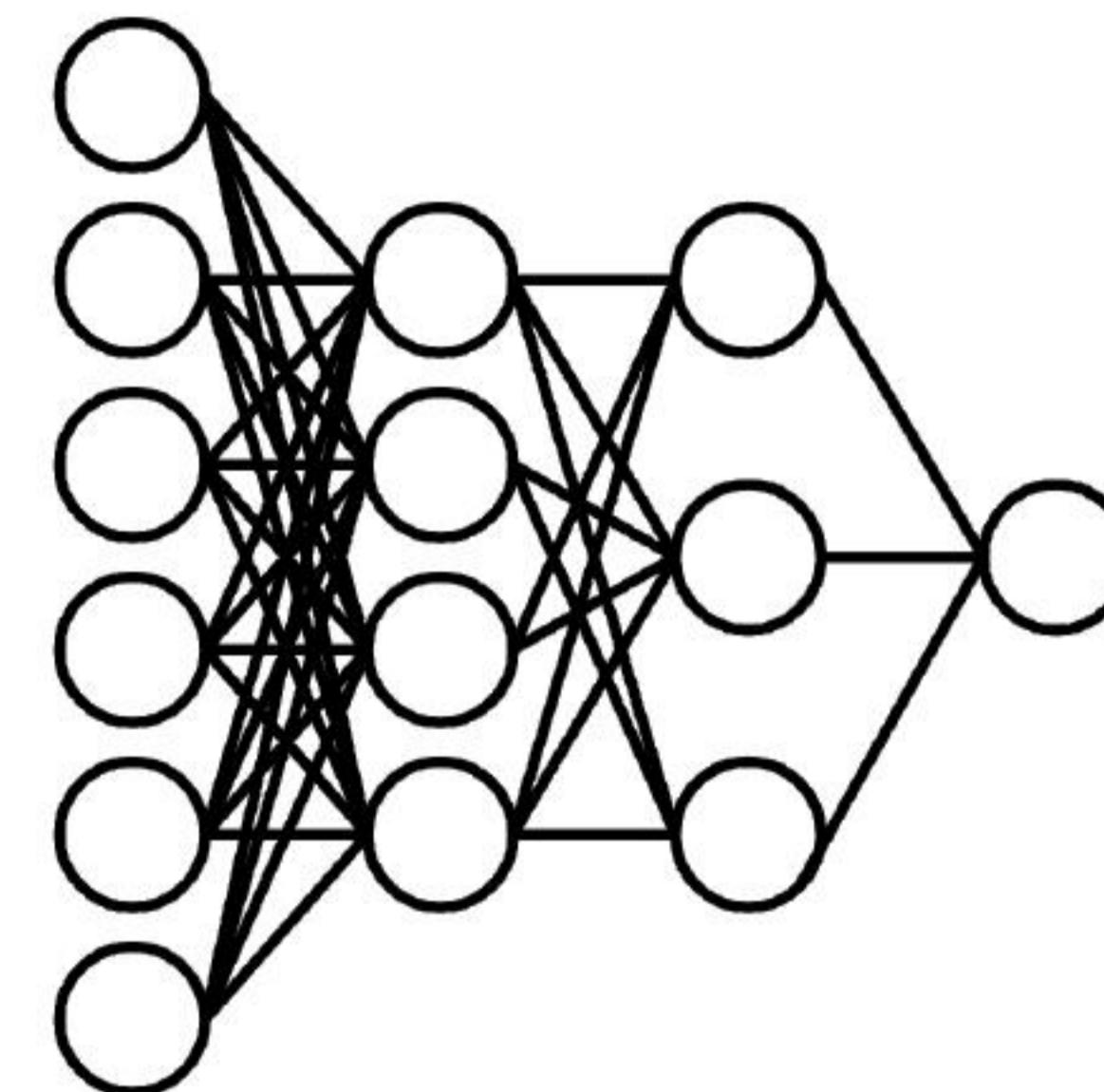


# Imitation Learning with Program Policy

Demonstrations



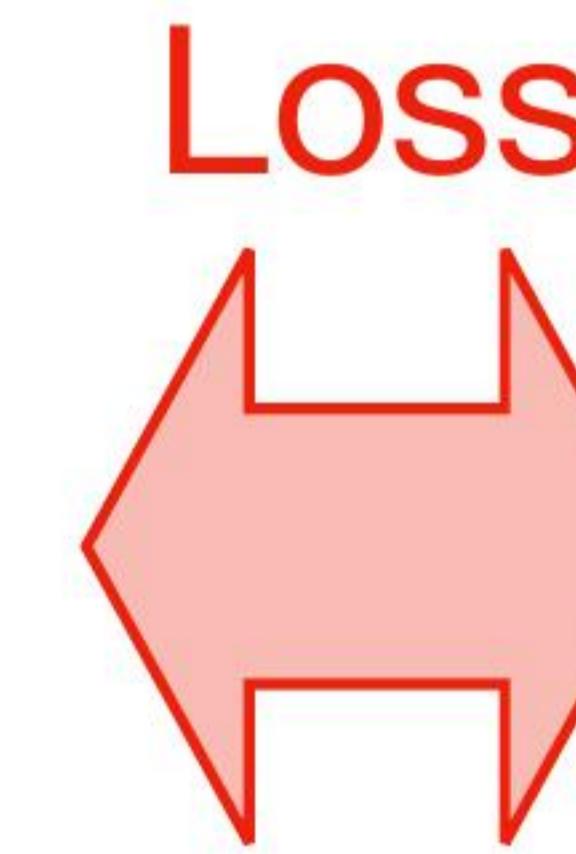
Neural Network  
Program Synthesizer



Synthesize

Predicted Program

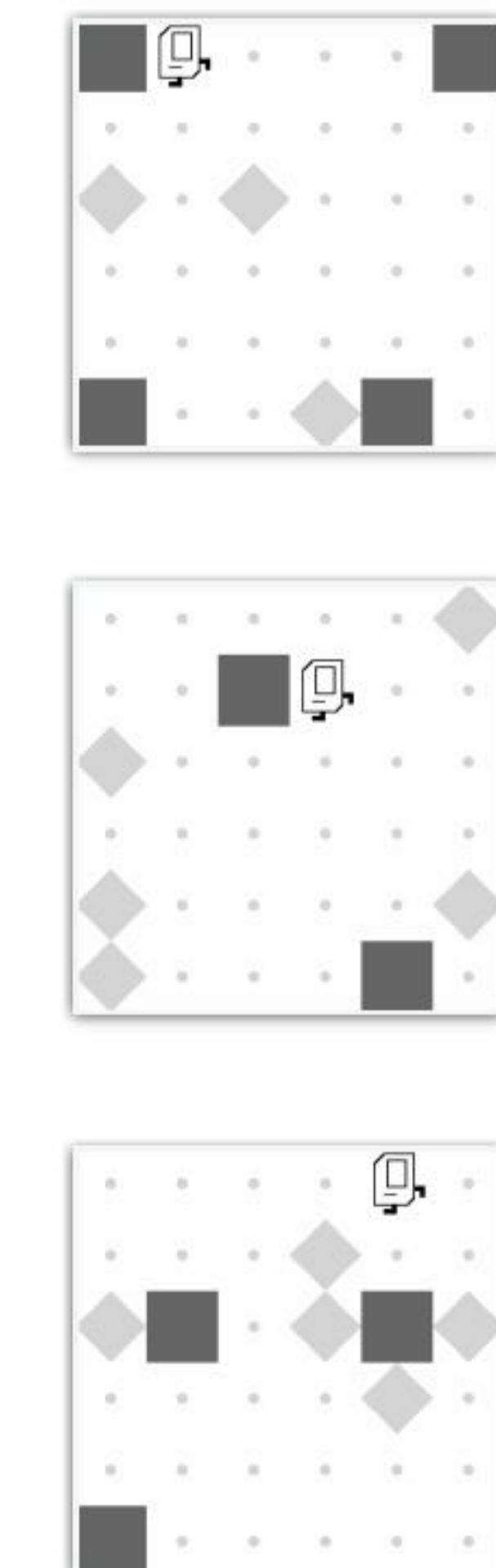
```
DEF run()
  IF frontIsClear
    move
  ELSE
    turnLeft
    move
  REPEAT(2)
    turnRight
```



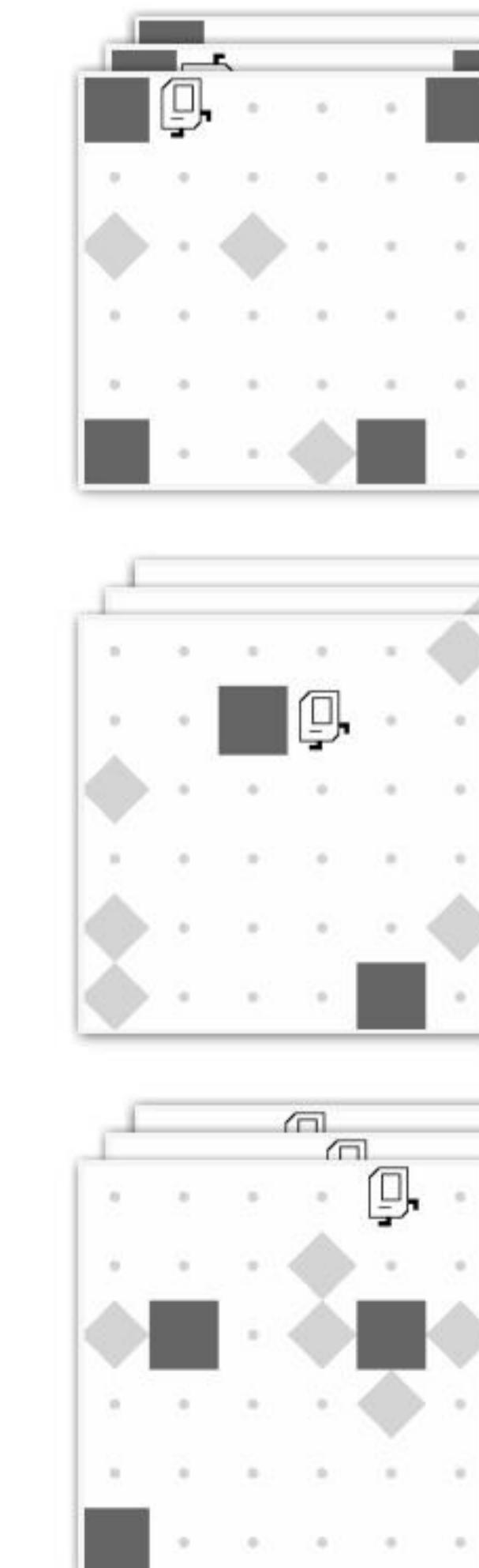
Ground Truth Program

```
DEF run()
  IF frontIsClear
    move
  ELSE
    turnRight
    move
  REPEAT(2)
    turnLeft
```

Initial States

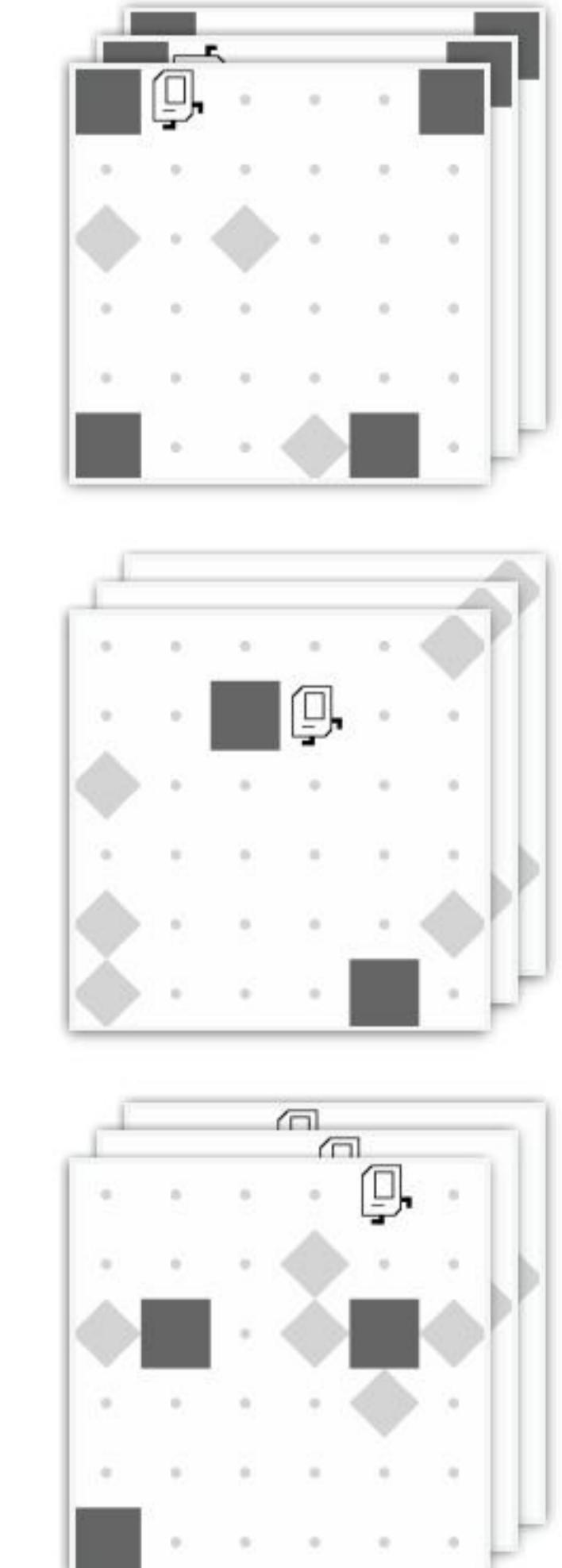


Predicted Execution



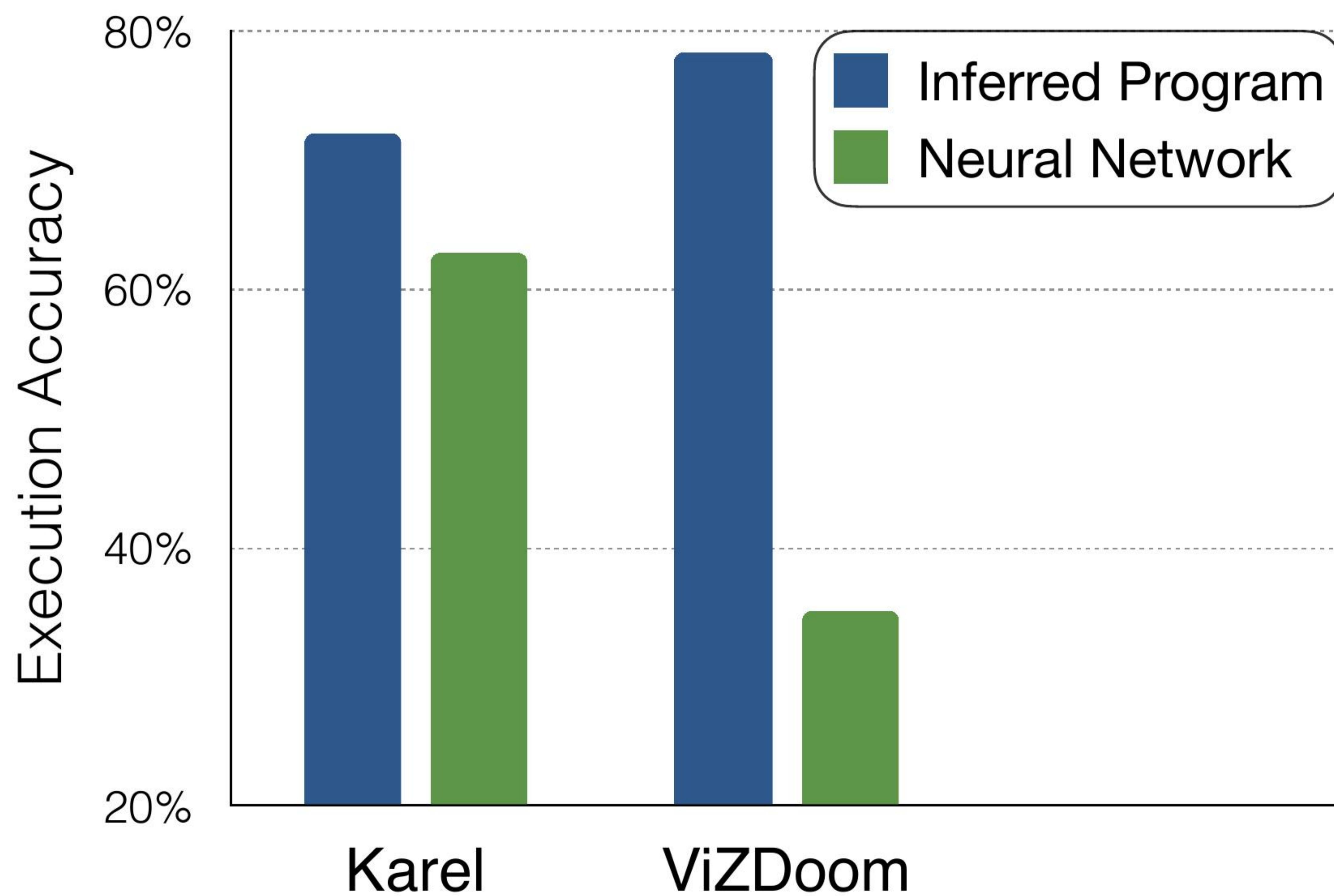
Execute

Ground Truth Execution



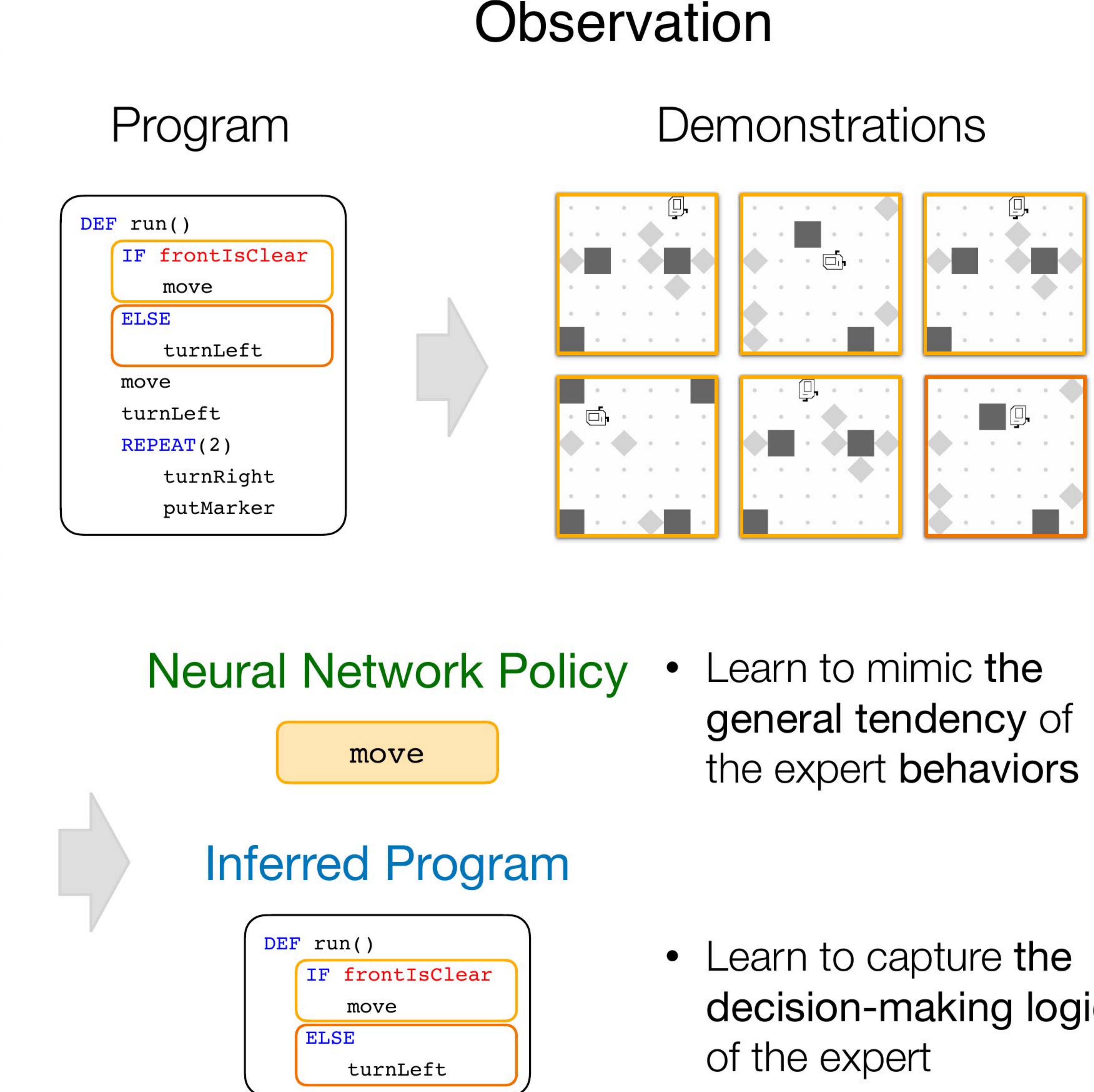
# Experimental Results

## Quantitative Results



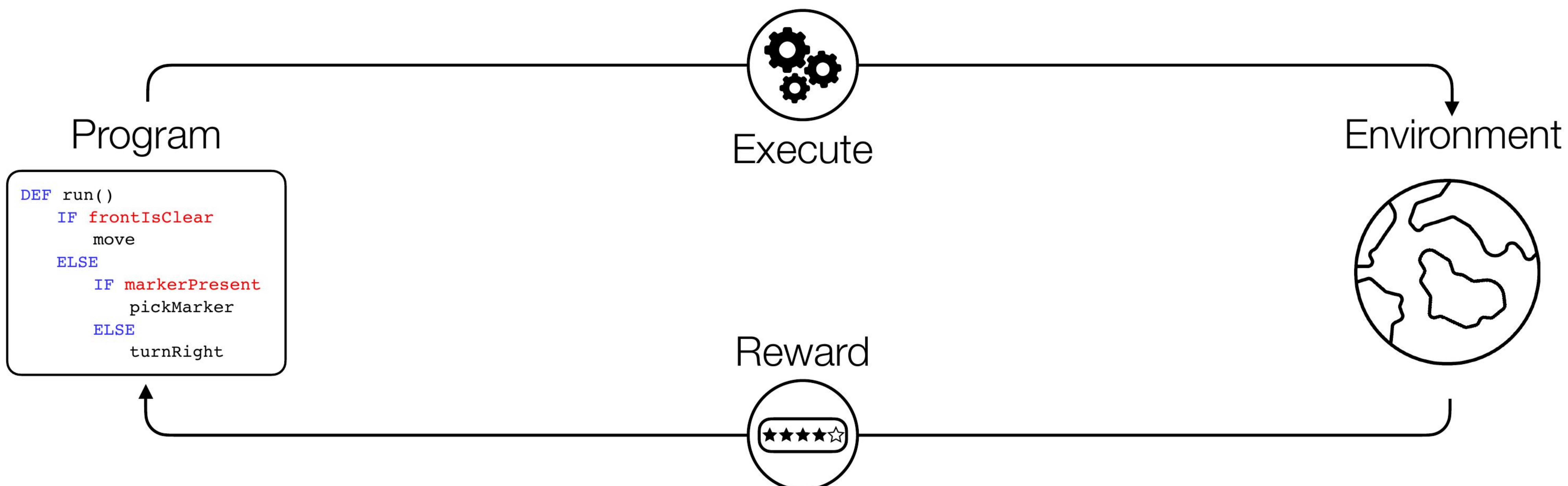
Evaluation: Execute the **inferred program** and the **learned neural network policy** on a set of unseen initial states and compare them to the **ground truth demonstrations**

## Observation



# Learning to Synthesize Programs as Interpretable and Generalizable Policies

NeurIPS 2021



Dweep Trivedi



Jesse Zhang




Shao-Hua Sun

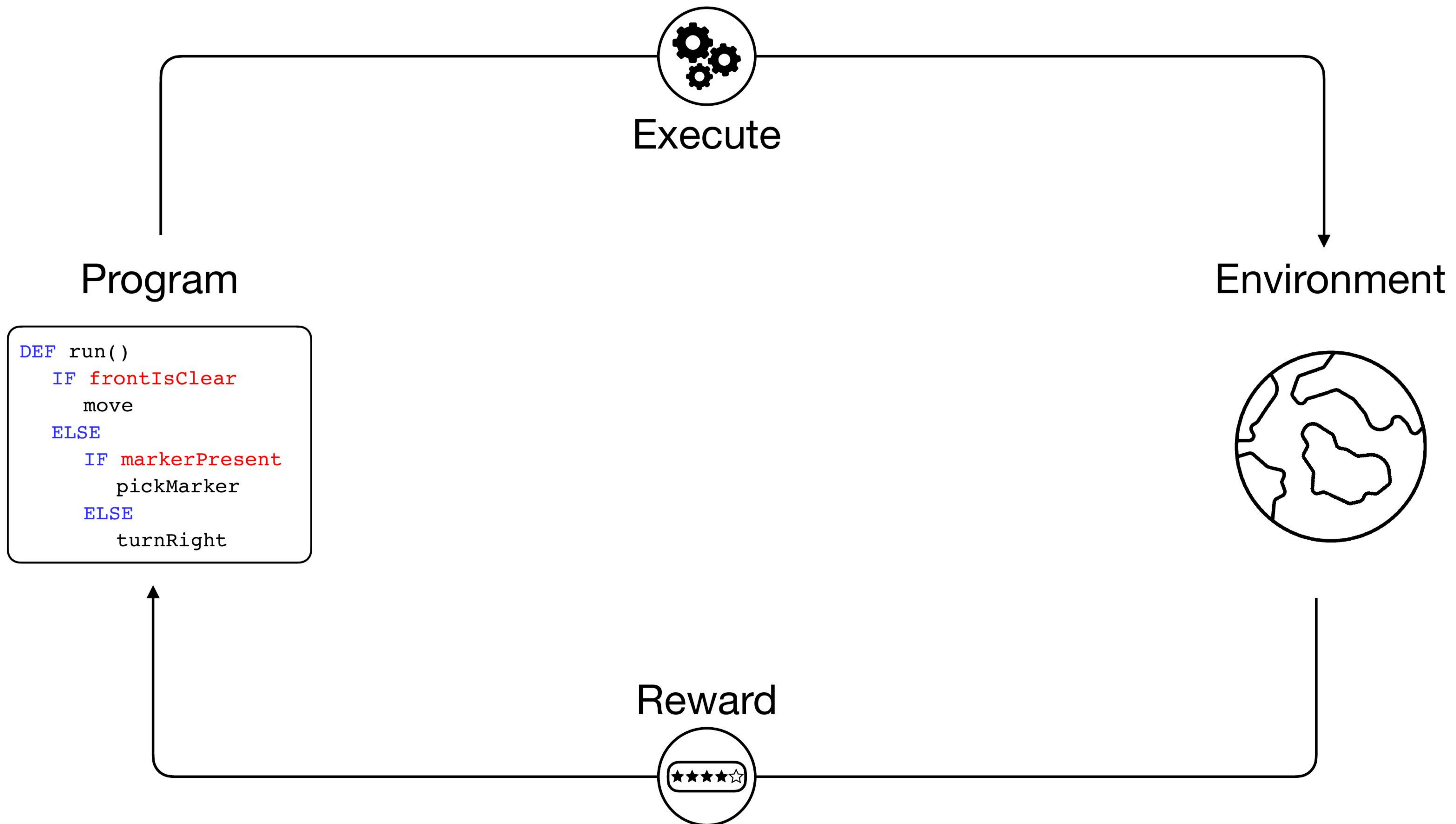


Joseph J. Lim

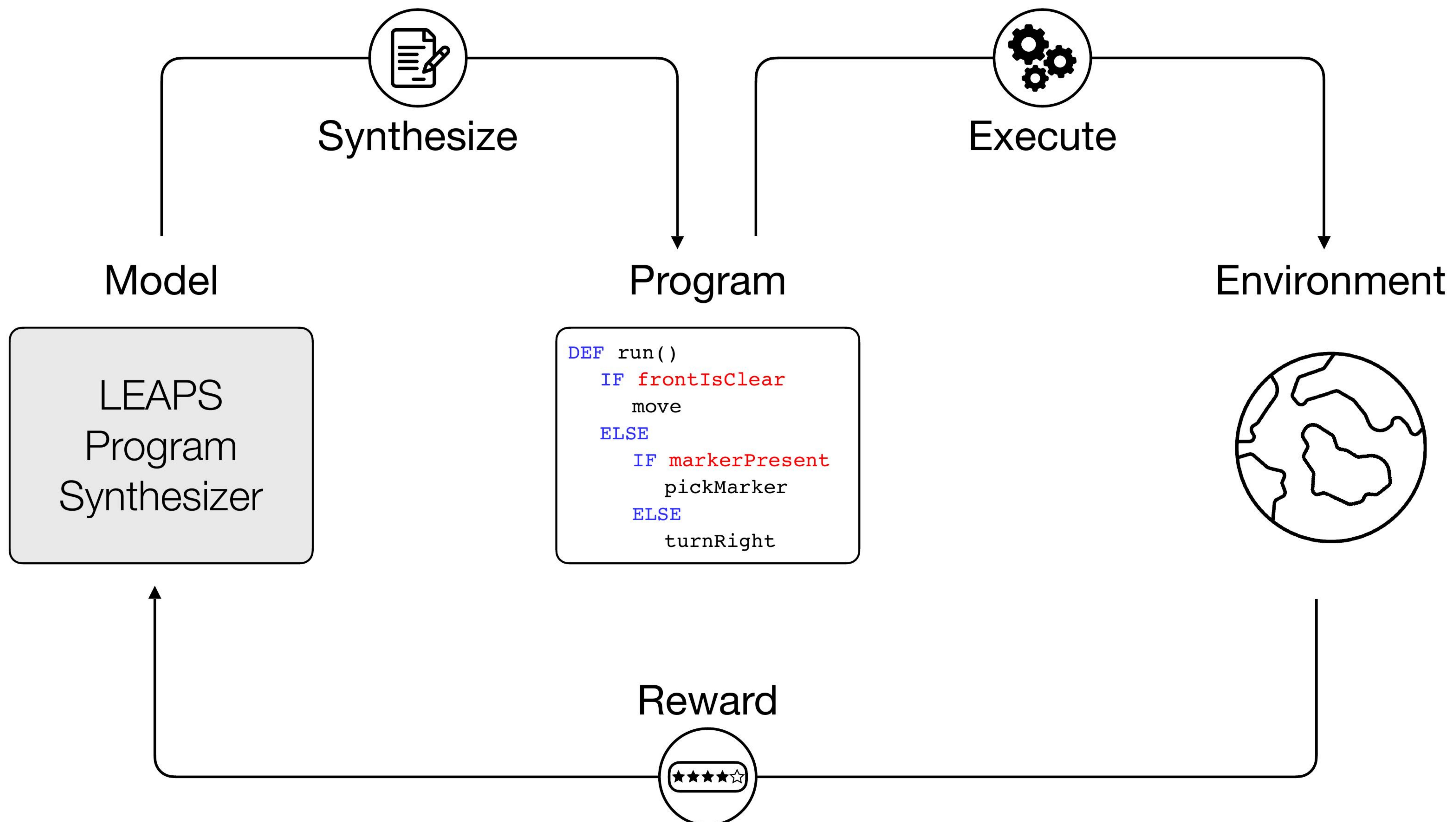
# Deep Reinforcement Learning



# Reinforcement Learning via Synthesizing Programs



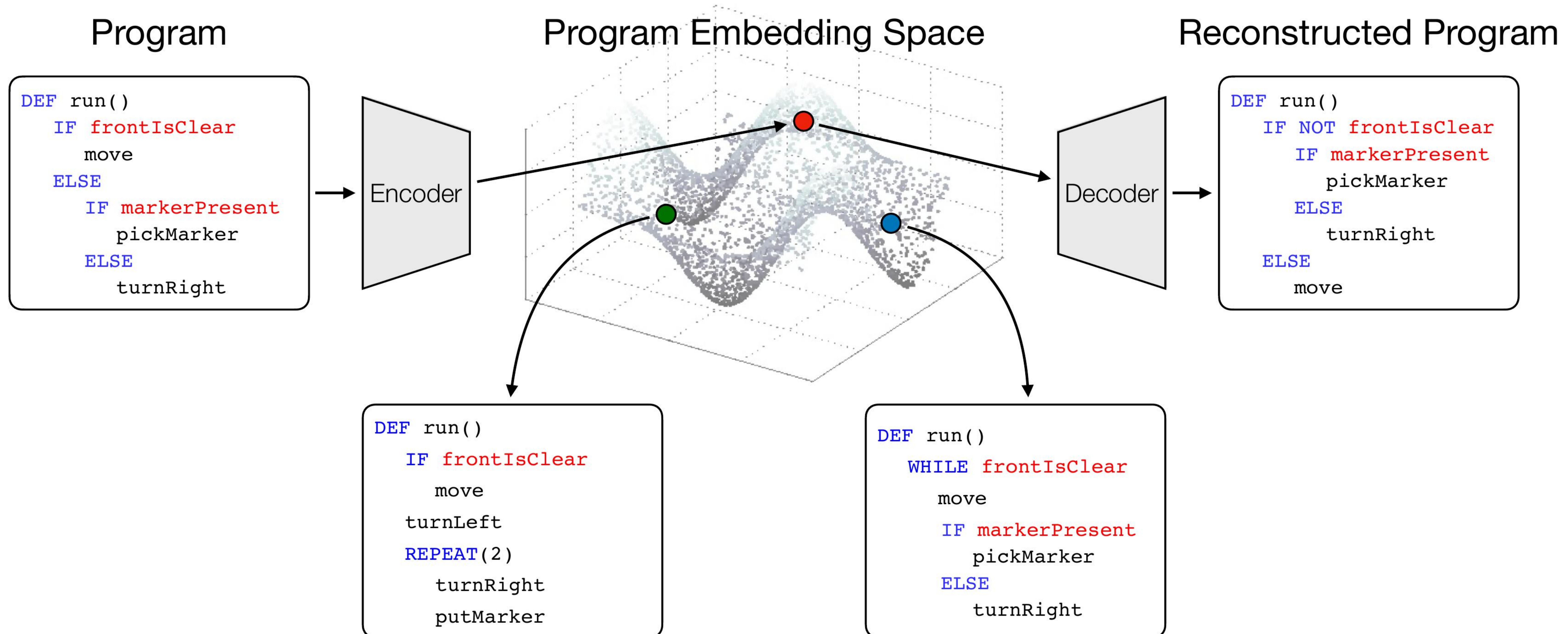
# Reinforcement Learning via Synthesizing Programs



# LEAPS: Learning Embeddings for Latent Program Synthesis

Stage 1 Learn a program embedding space from randomly generated programs

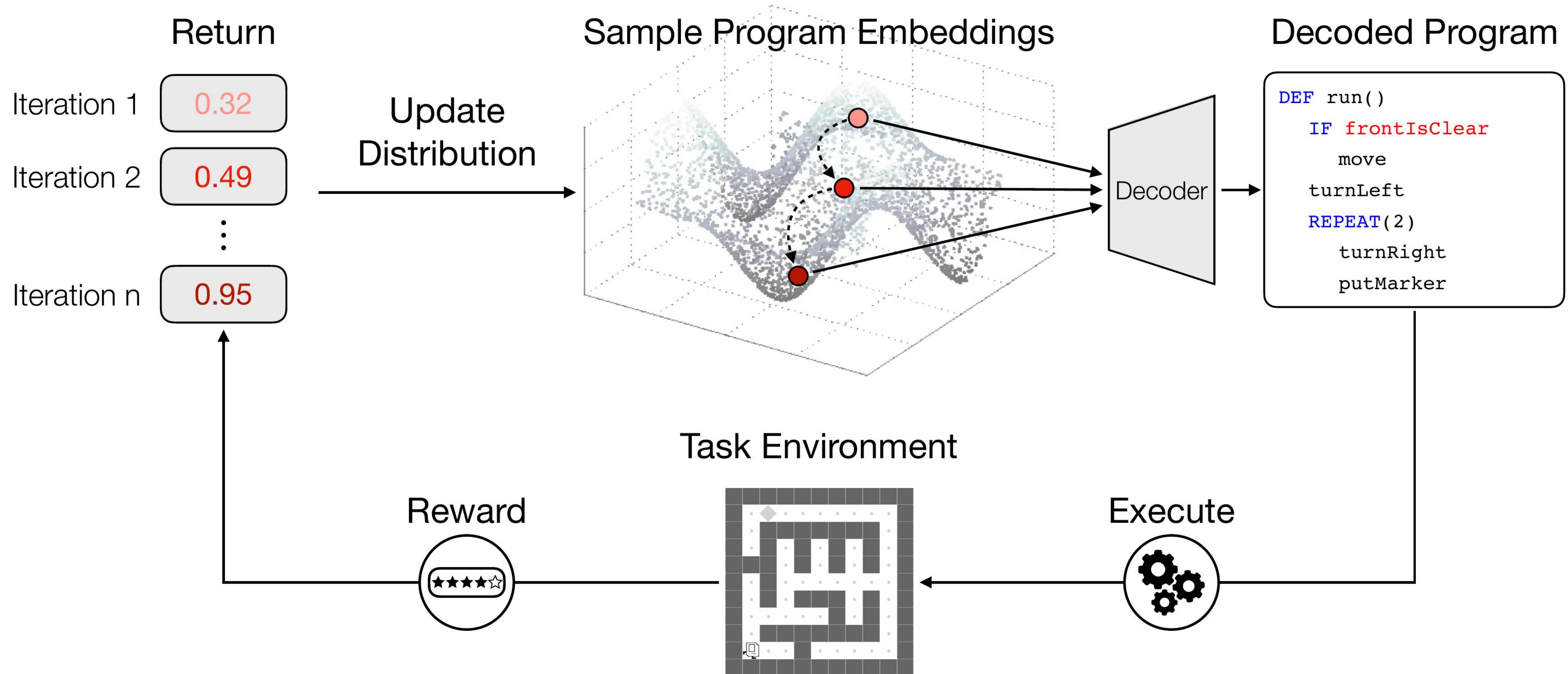
Goal Learn the **grammar** and the **environment dynamics**



# LEAPS: Learning Embeddings for Latent Program Synthesis

Stage 2 Search for a task-solving program using the cross-entropy method (CEM)

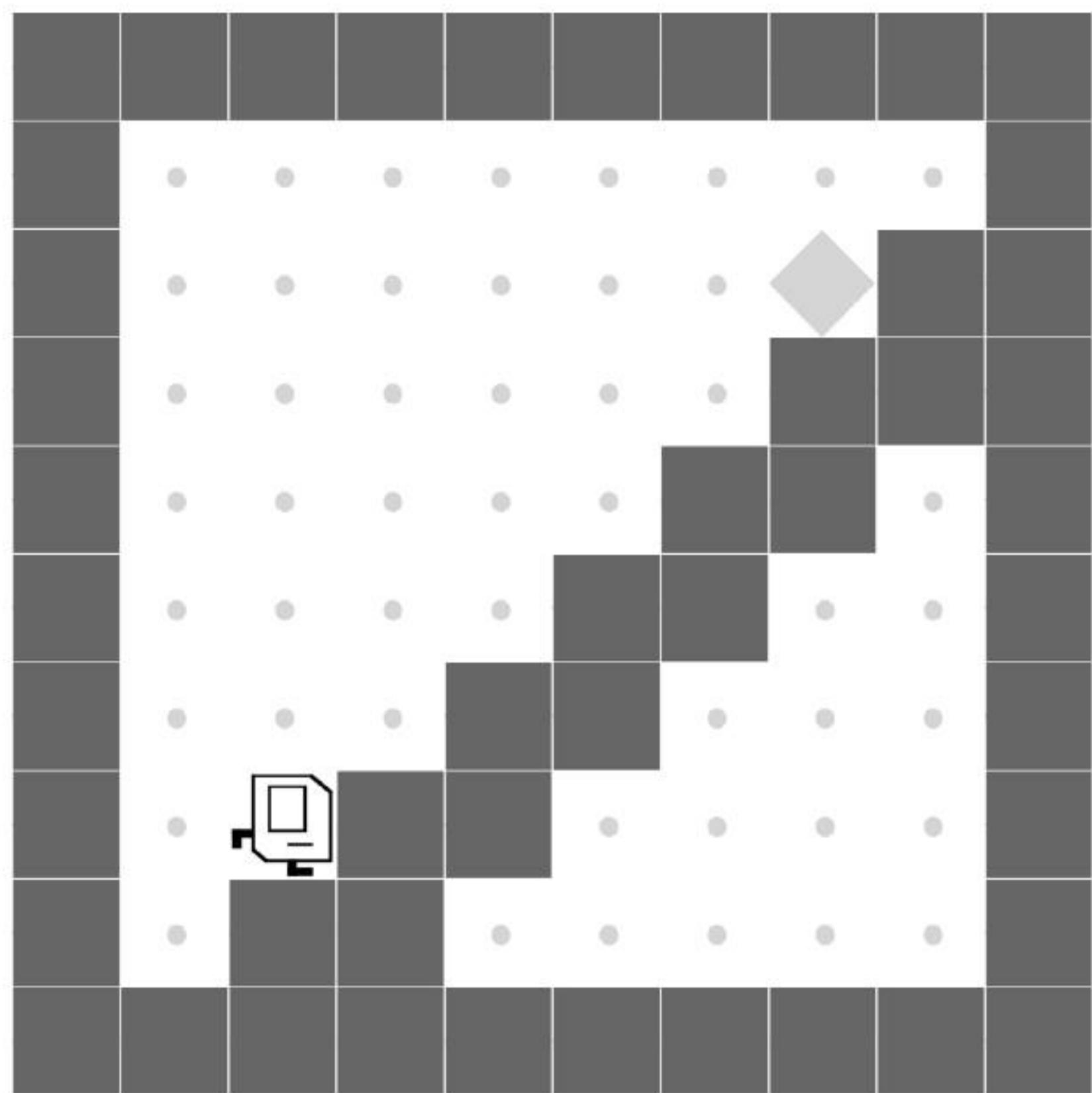
Goal Optimize the **task performance**



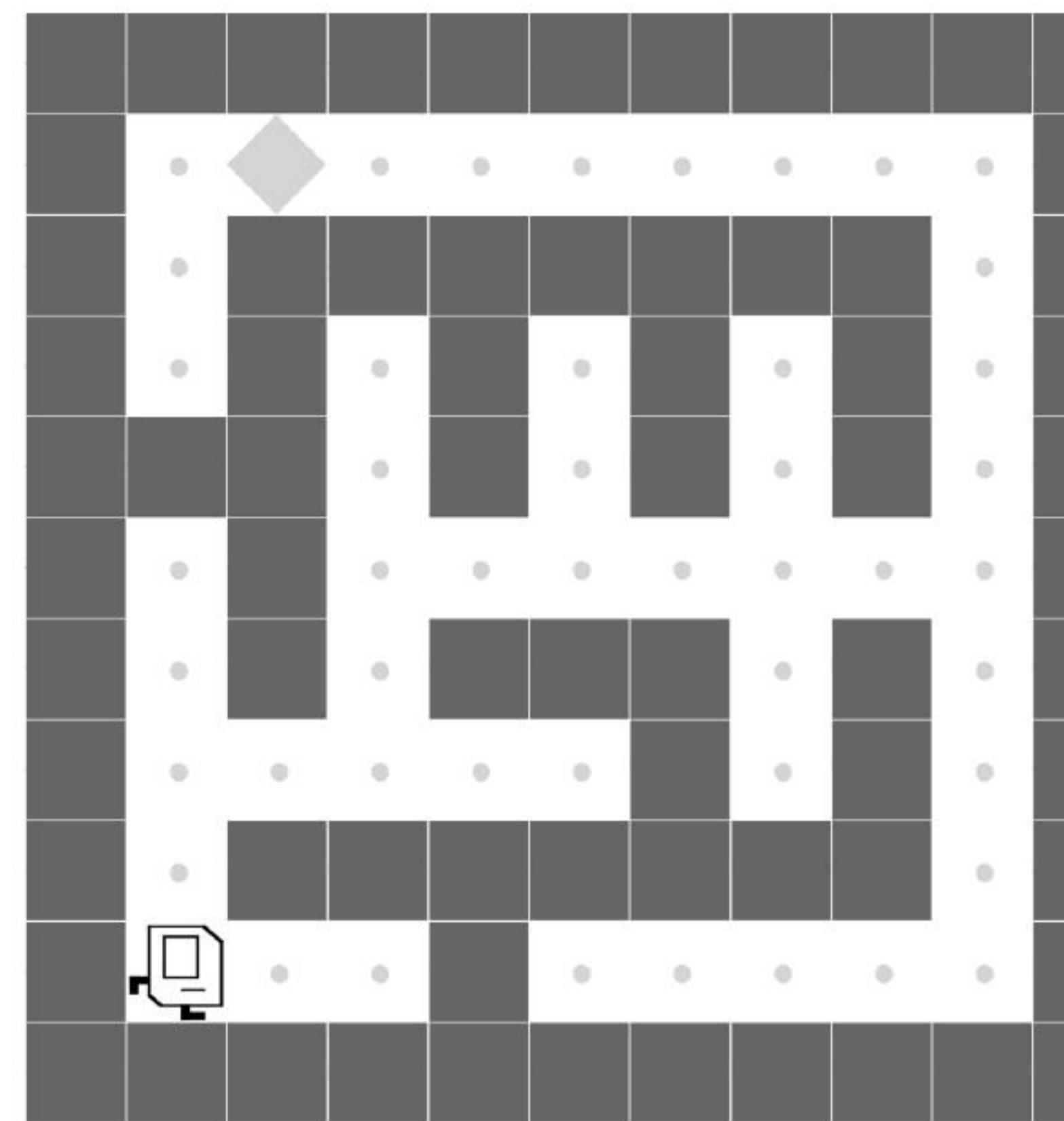
# Karel Tasks

---

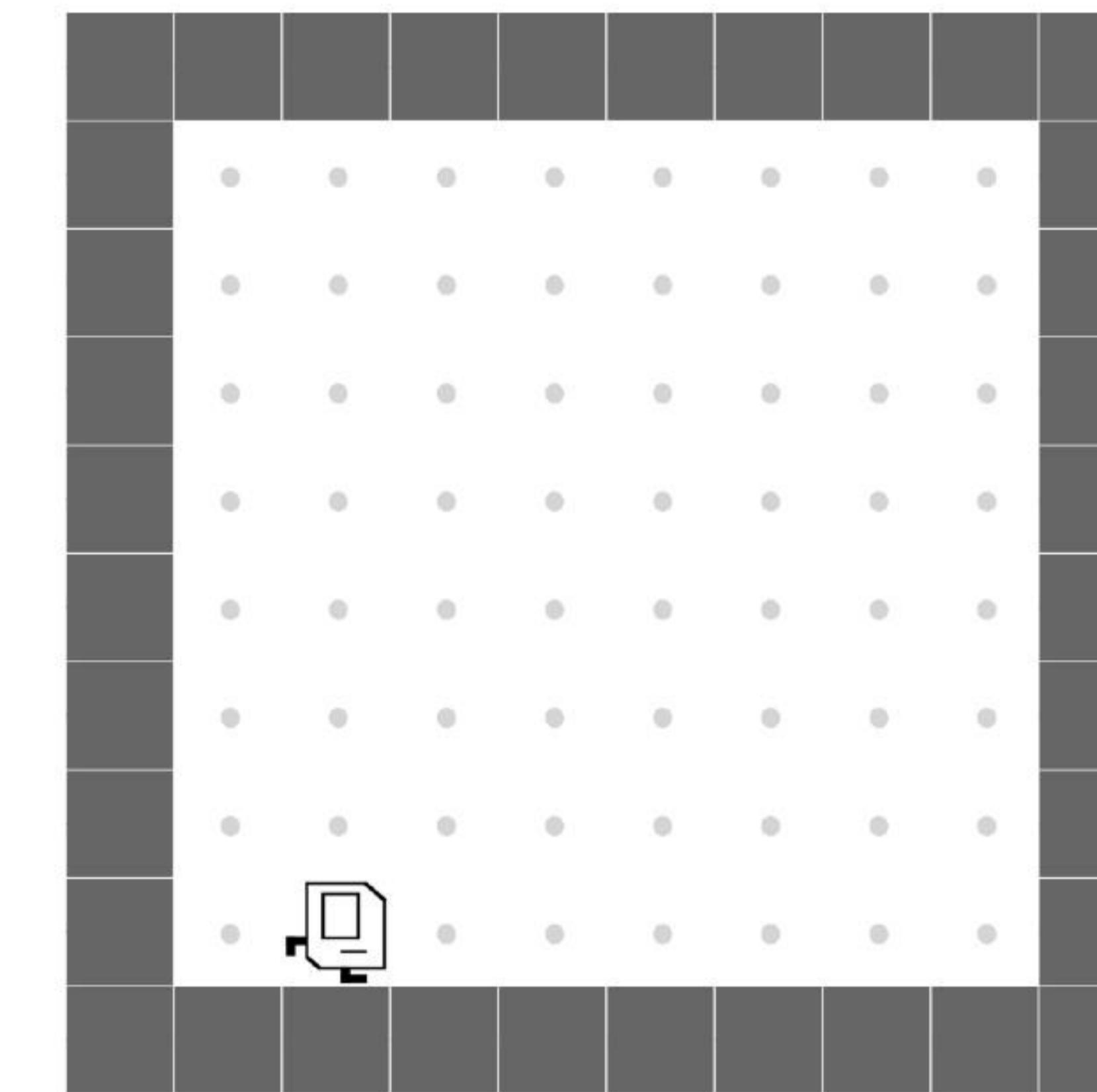
StairClimber



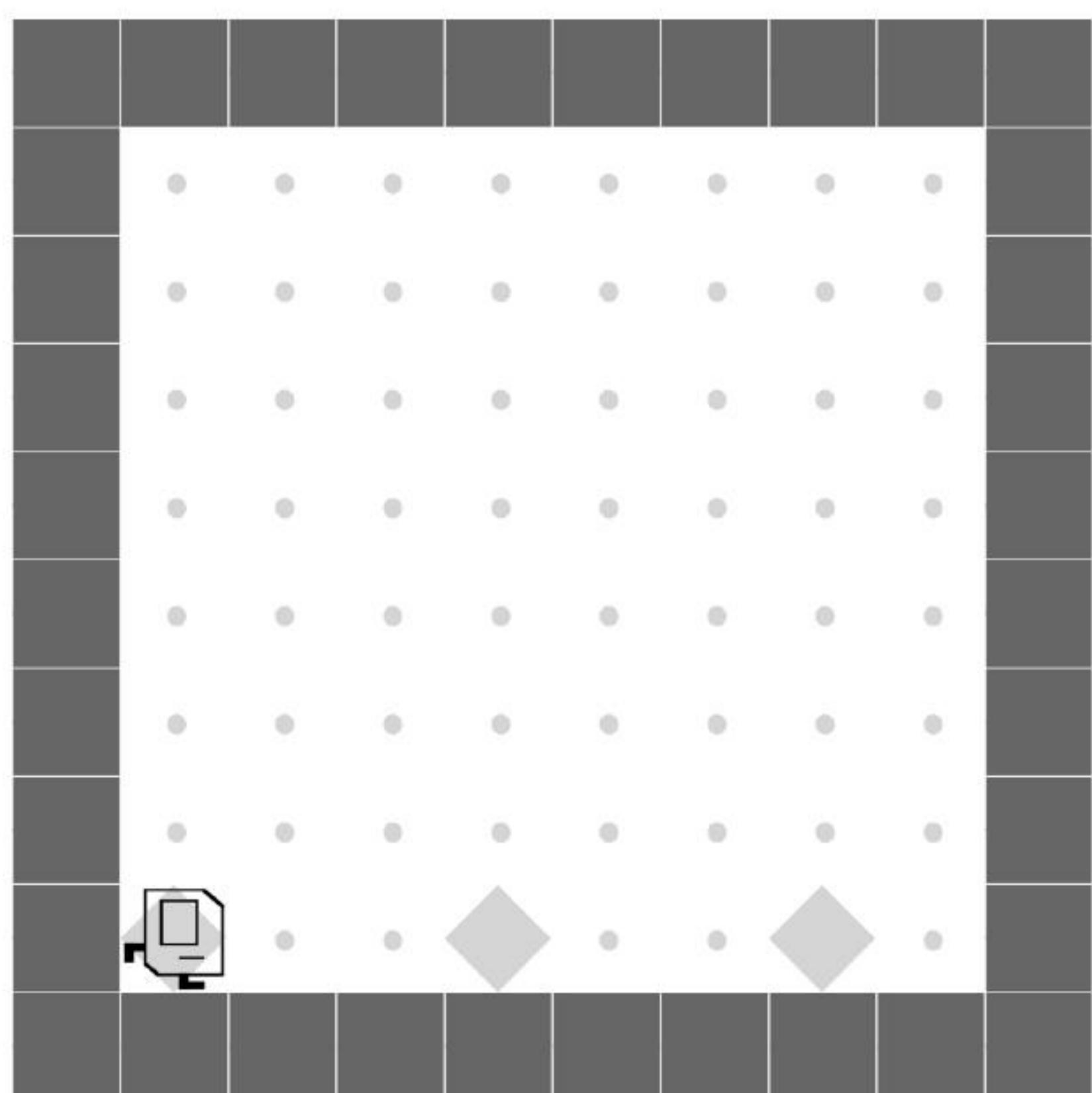
Maze



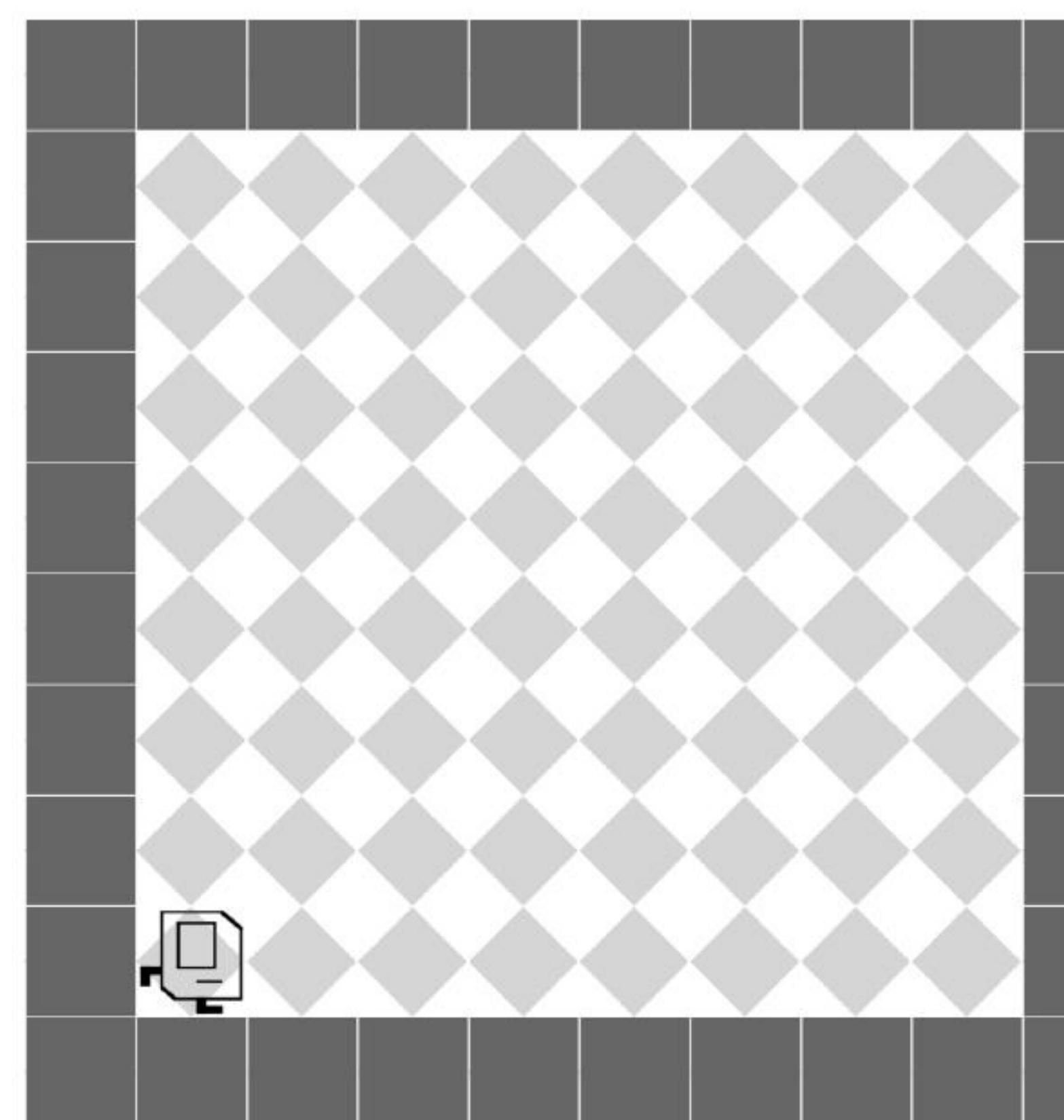
FourCorners



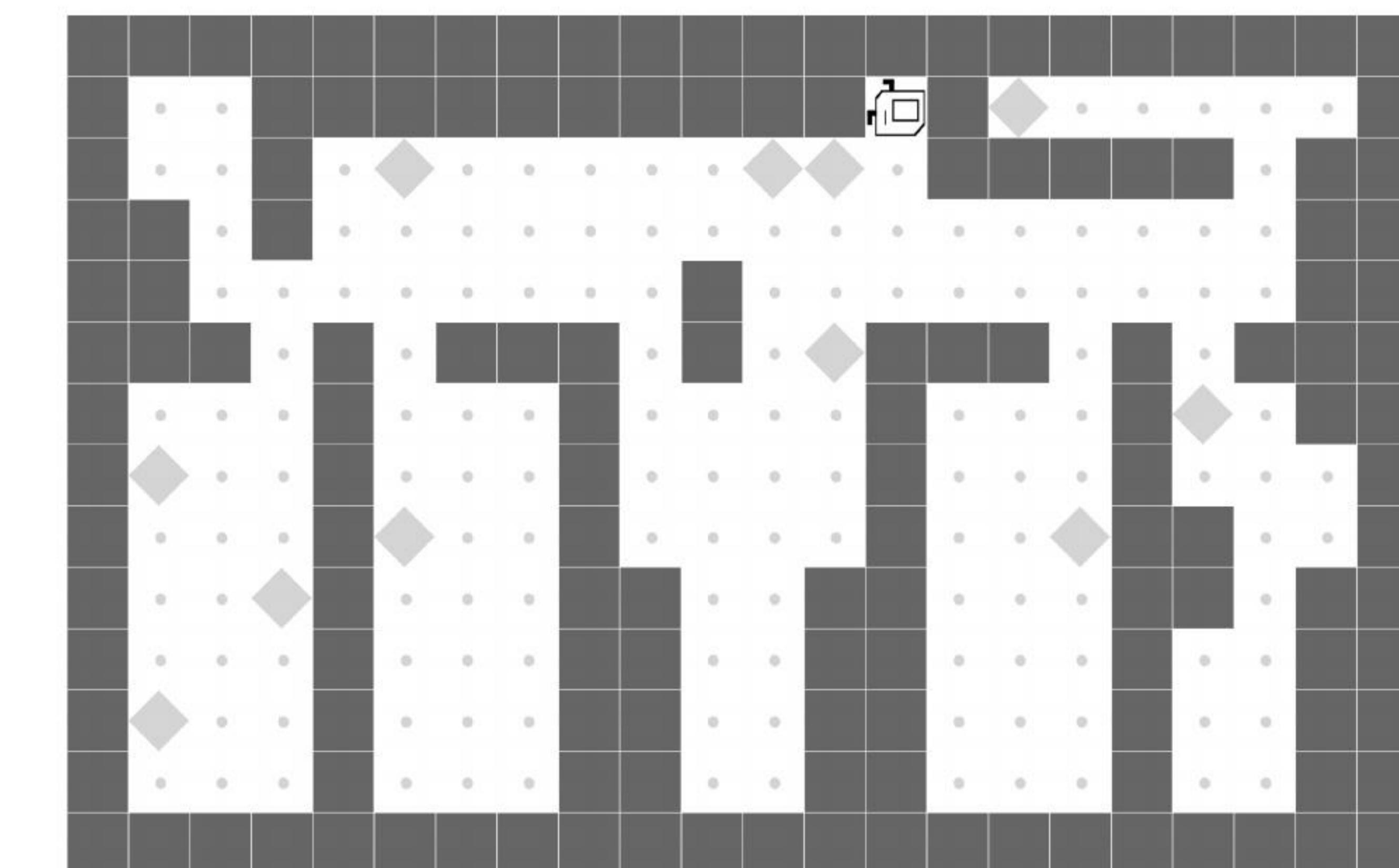
TopOff



Harvester

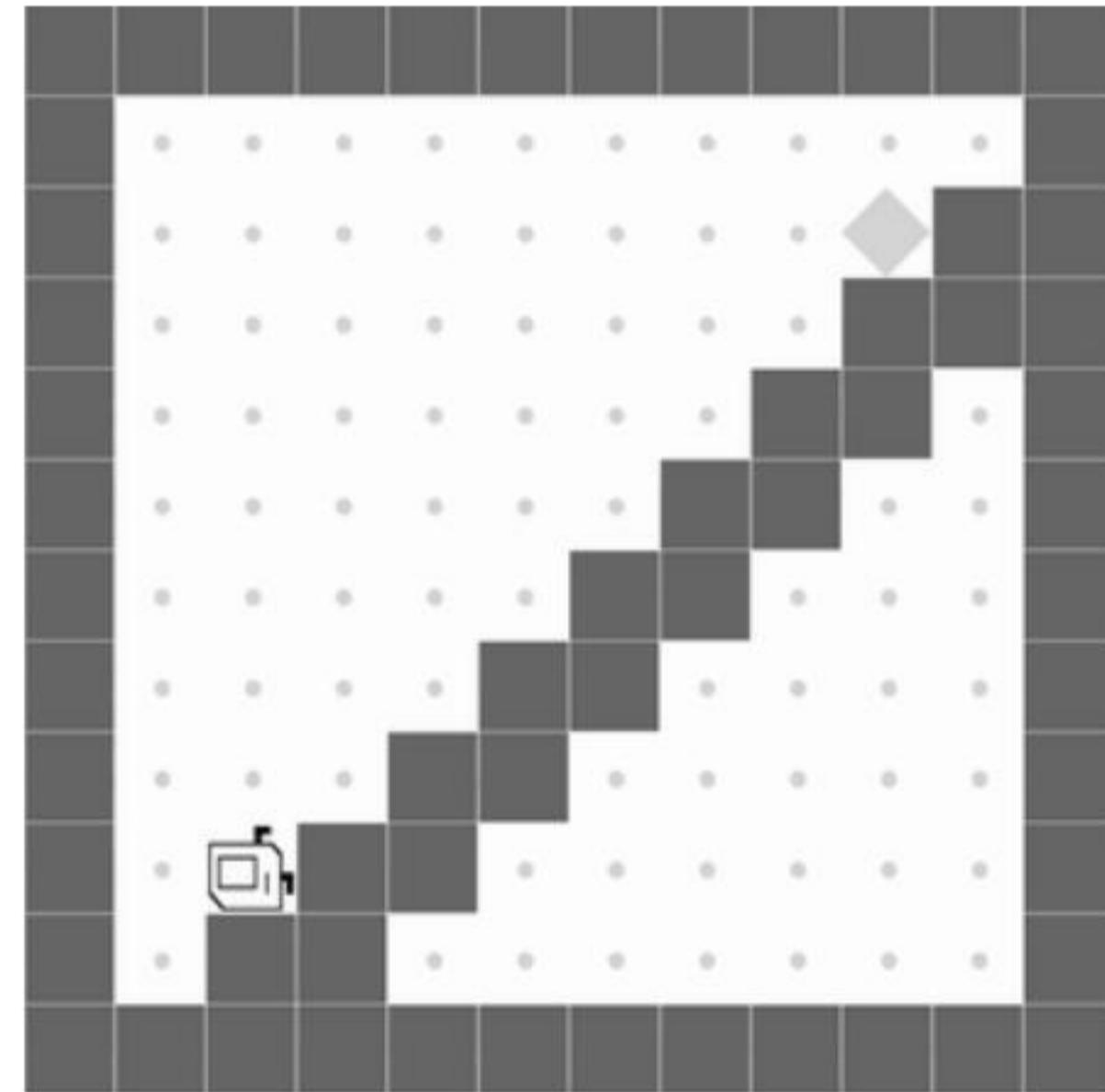


CleanHouse

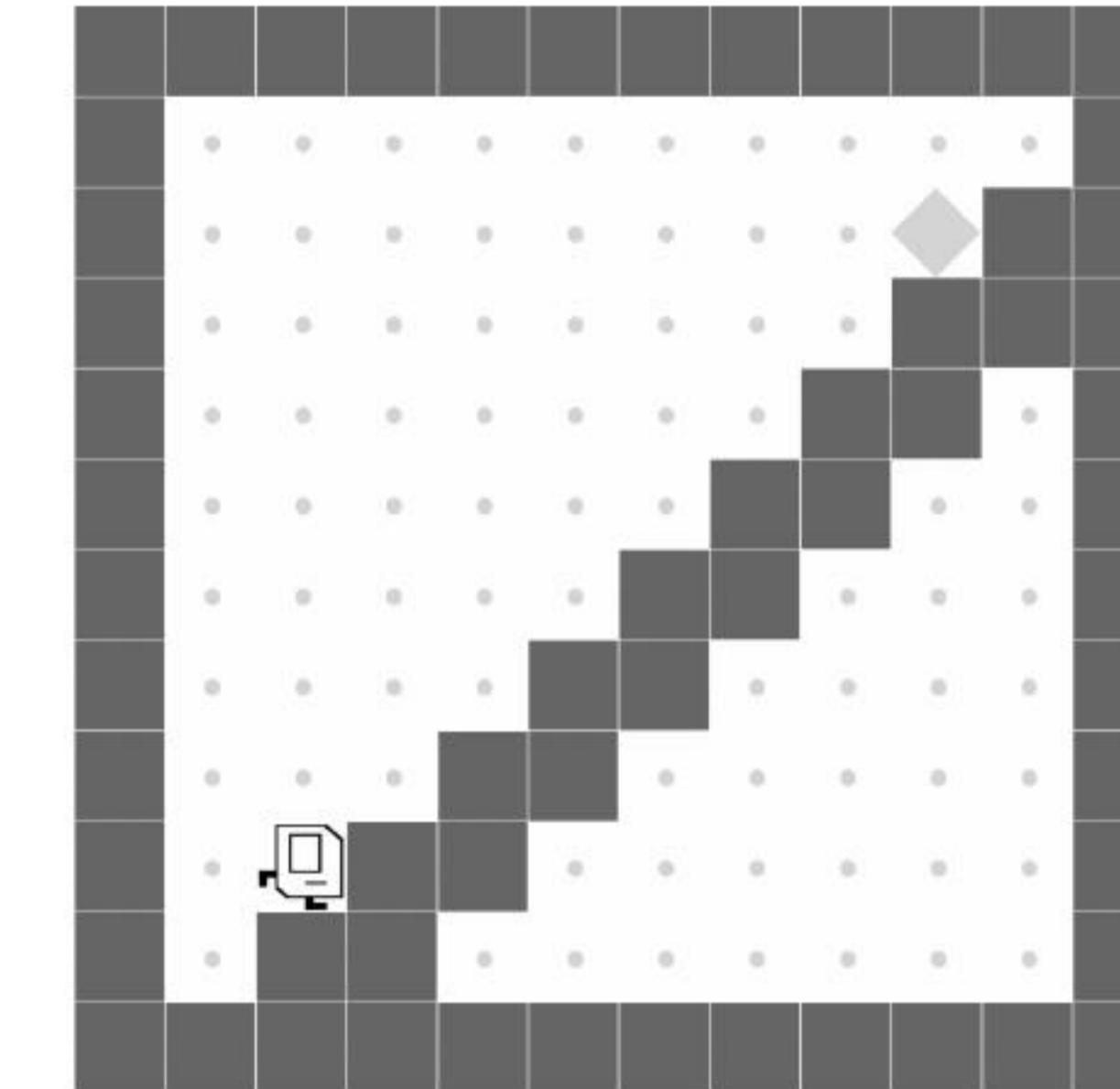


# Qualitative Results

StairClimber

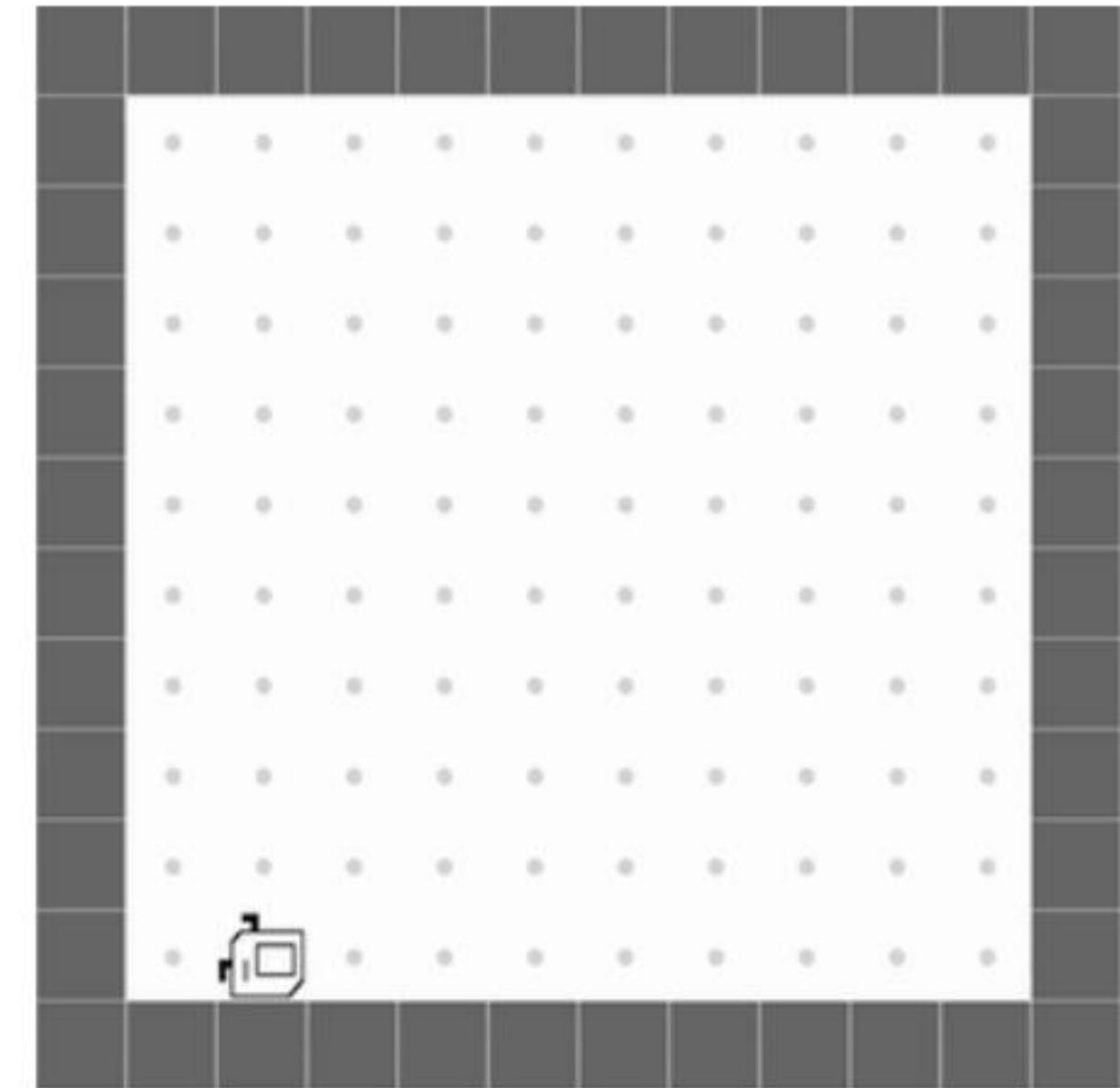


Deep RL

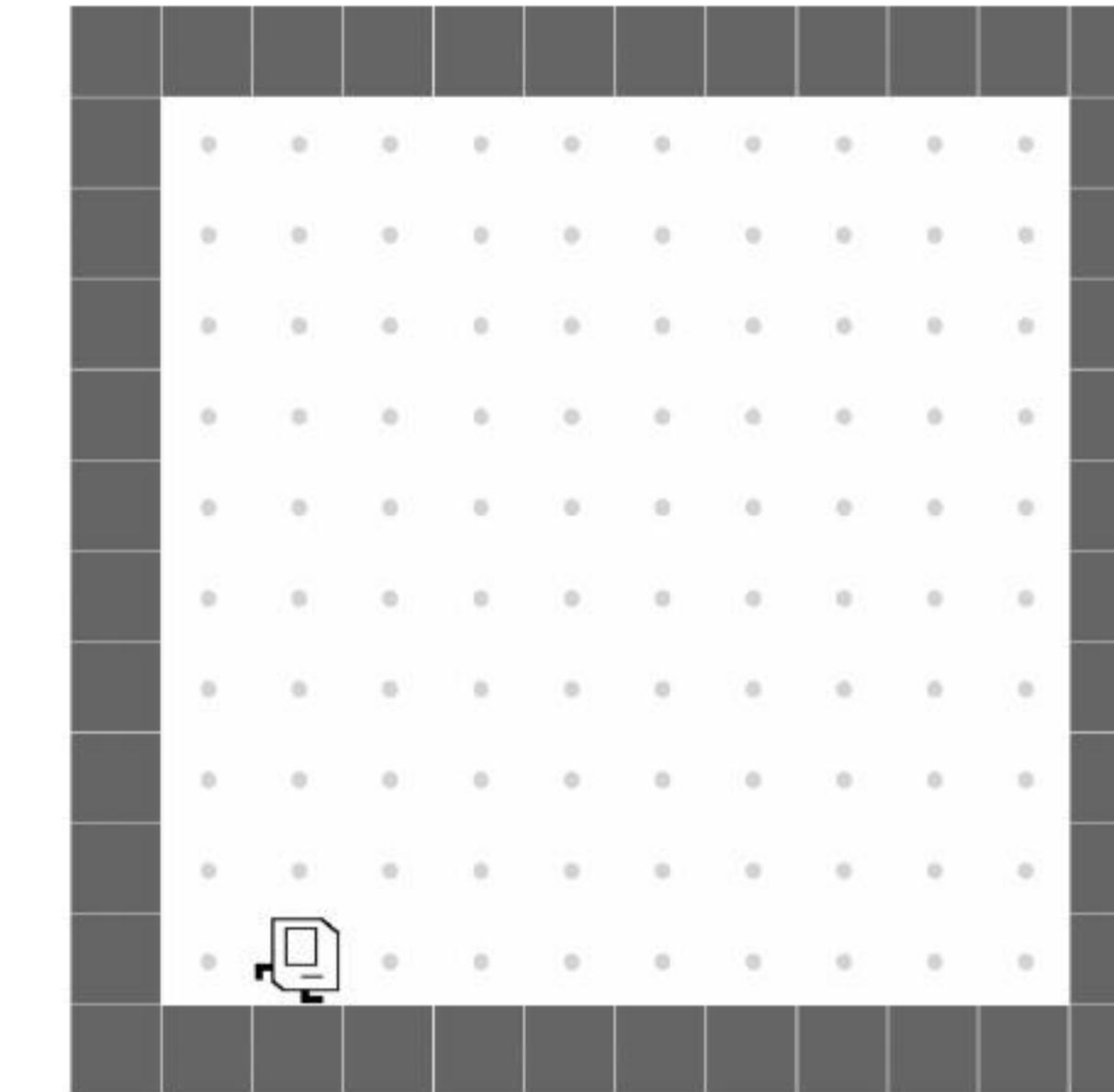


LEAPS

FourCorners

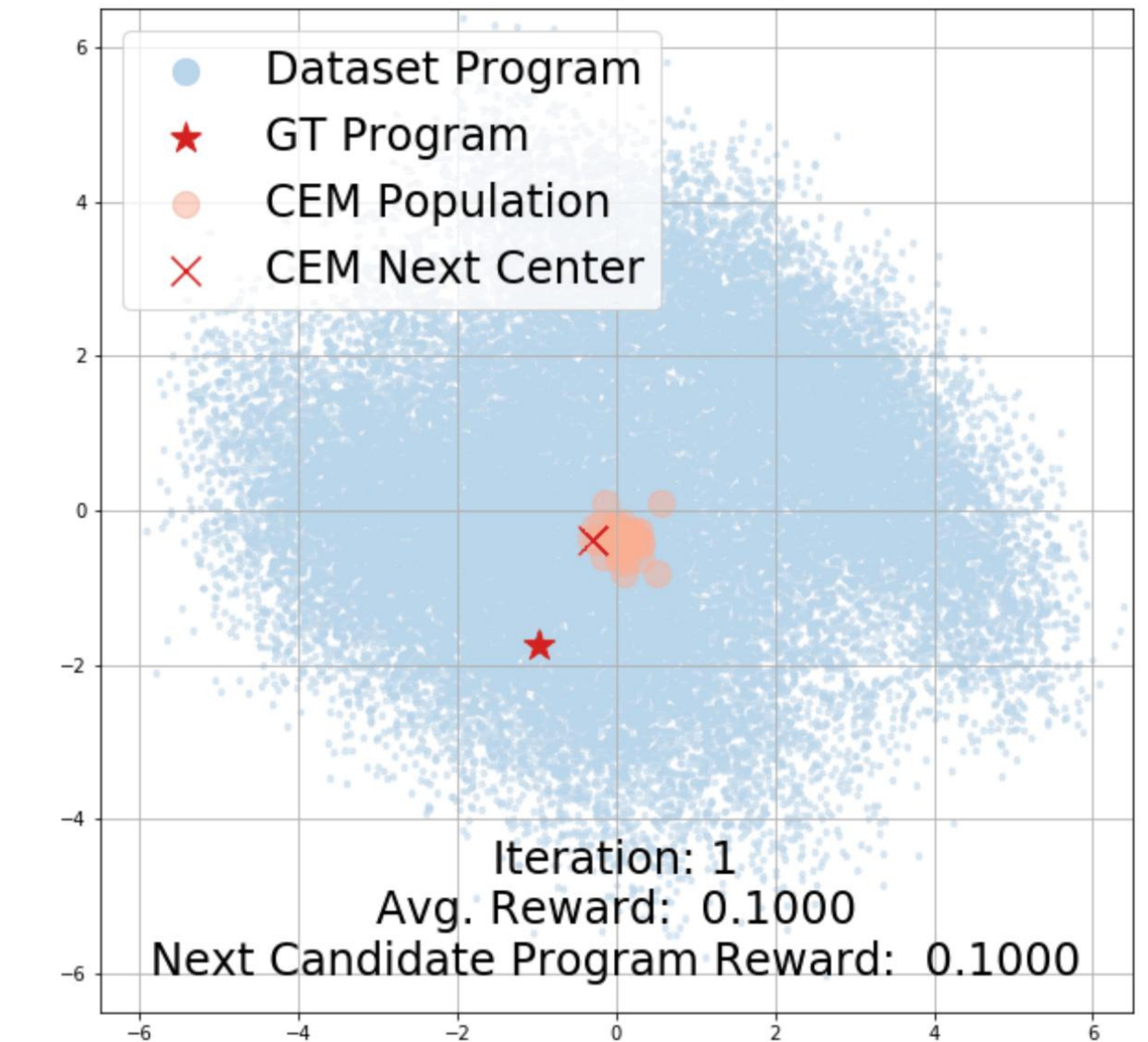


Deep RL

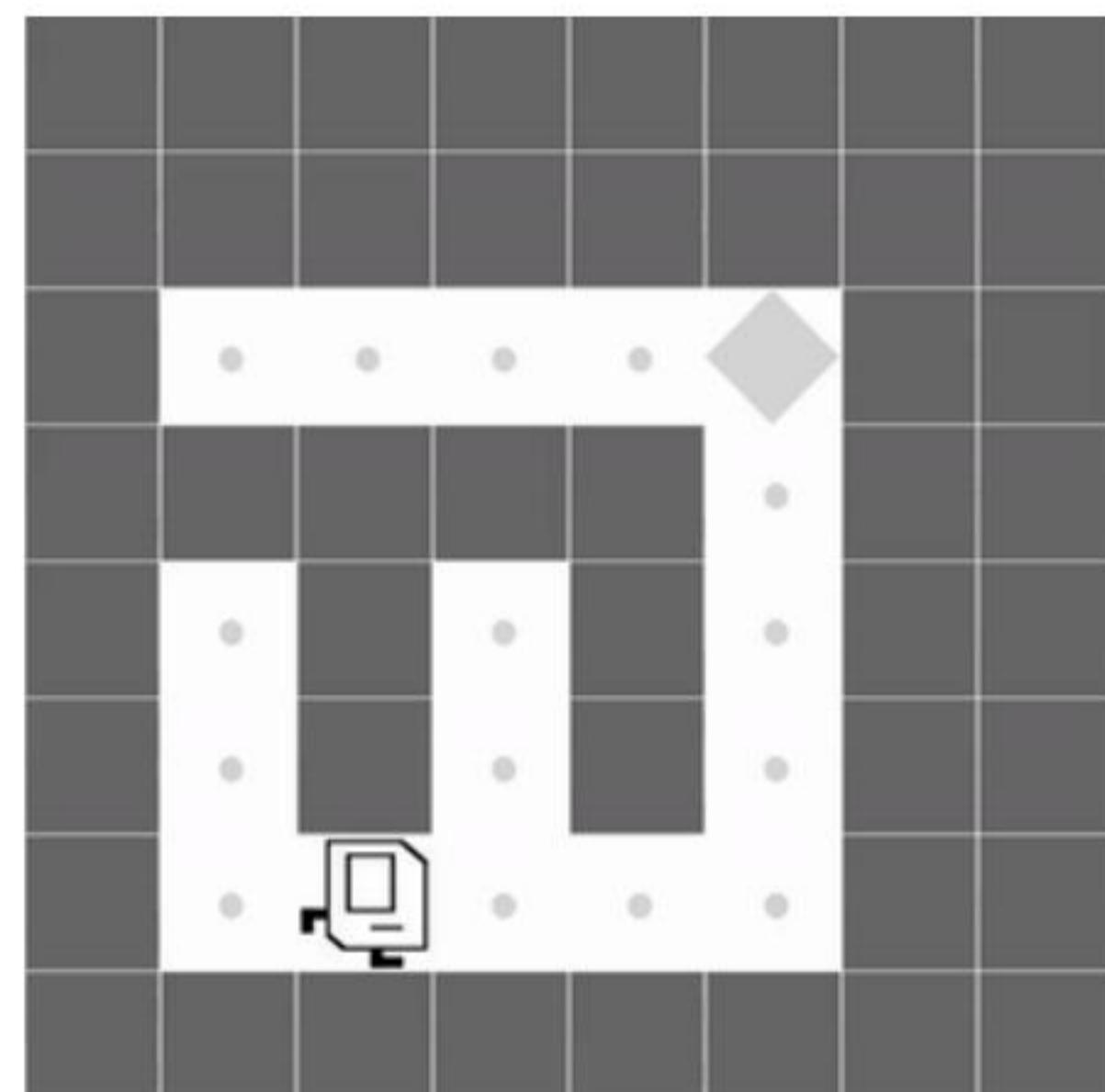


LEAPS

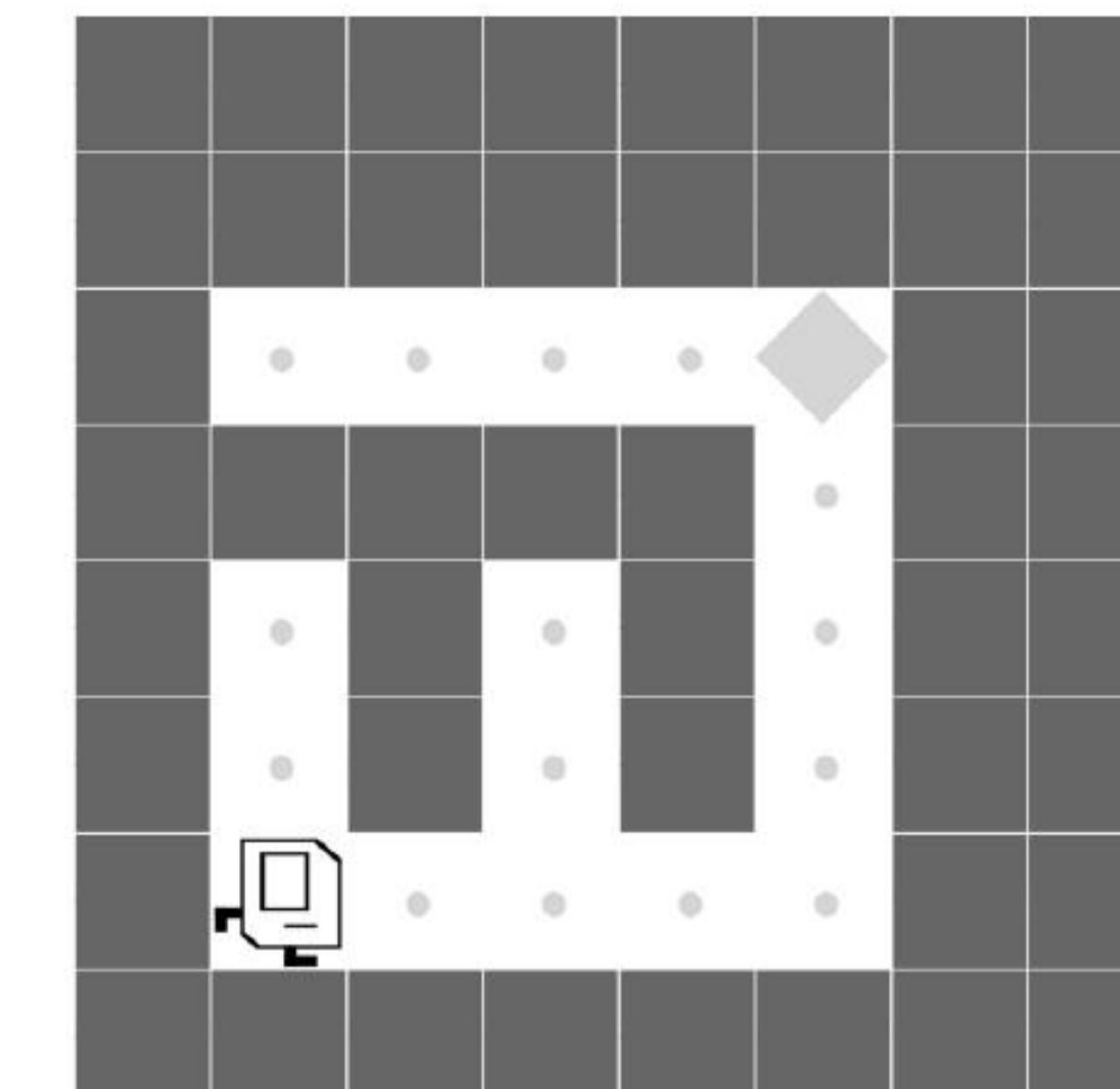
CEM trajectory Visualization



Maze

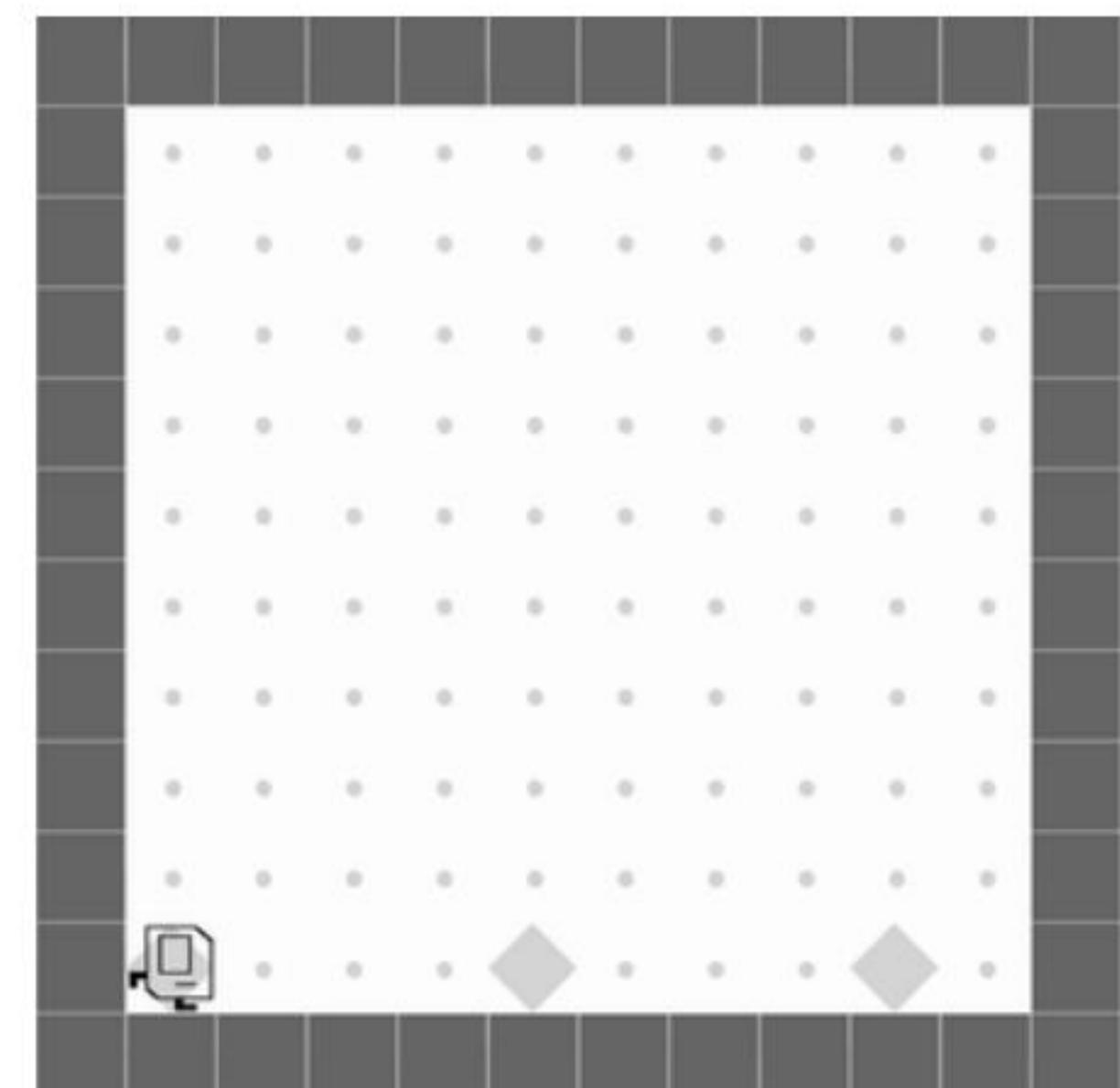


Deep RL

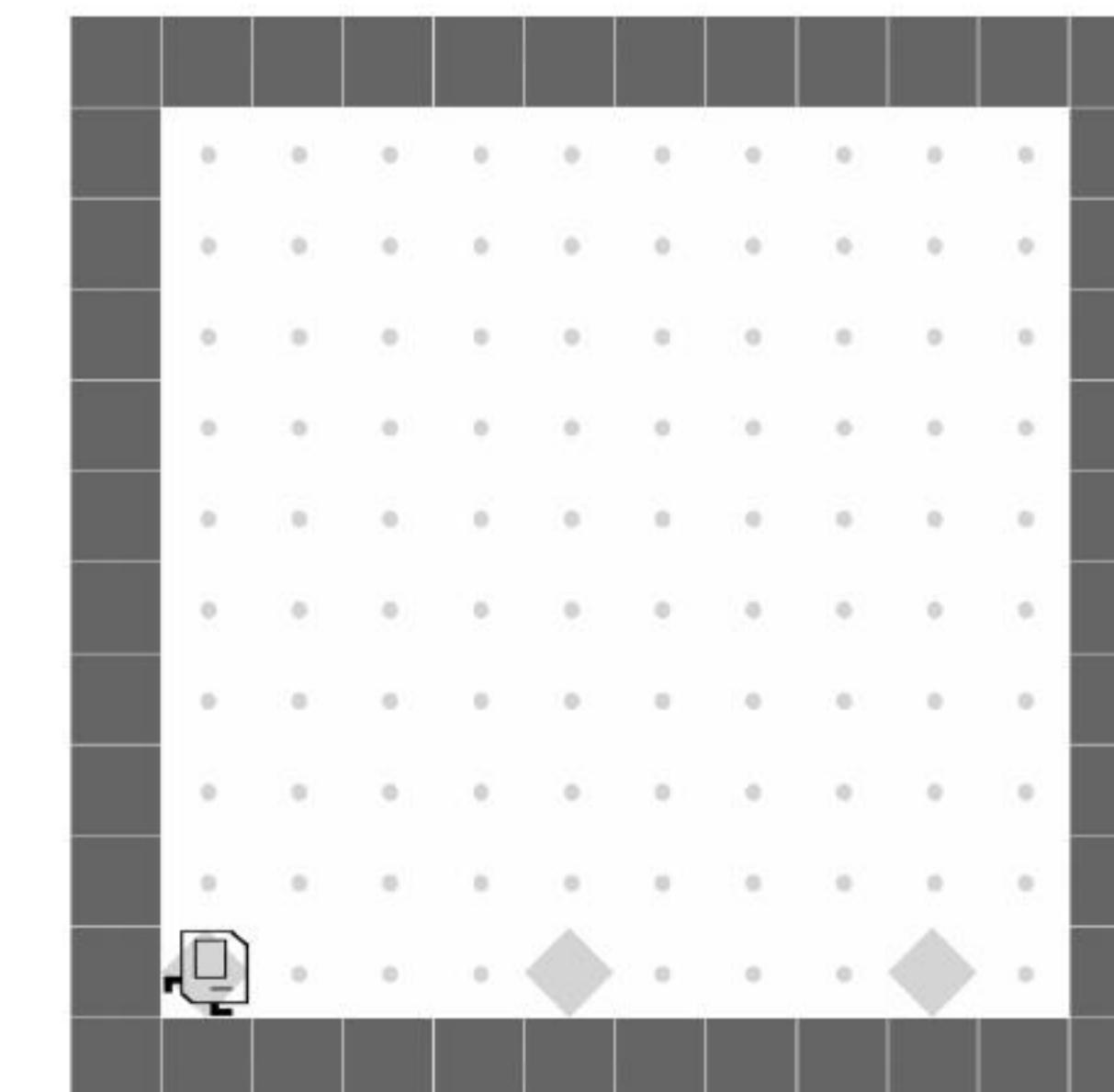


LEAPS

TopOff



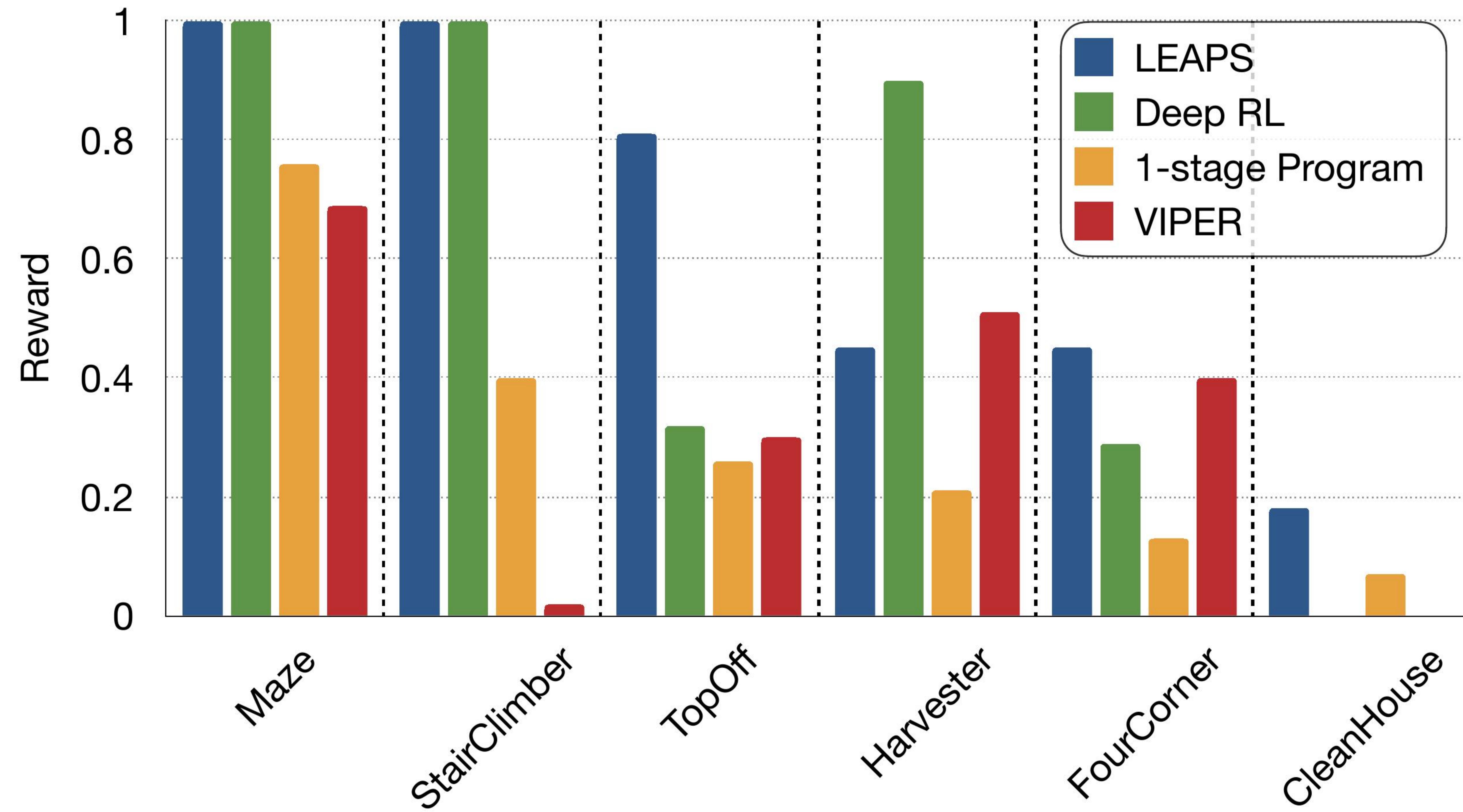
Deep RL



LEAPS

Goal: Search for a StairClimber program  
in the learned program embedding space

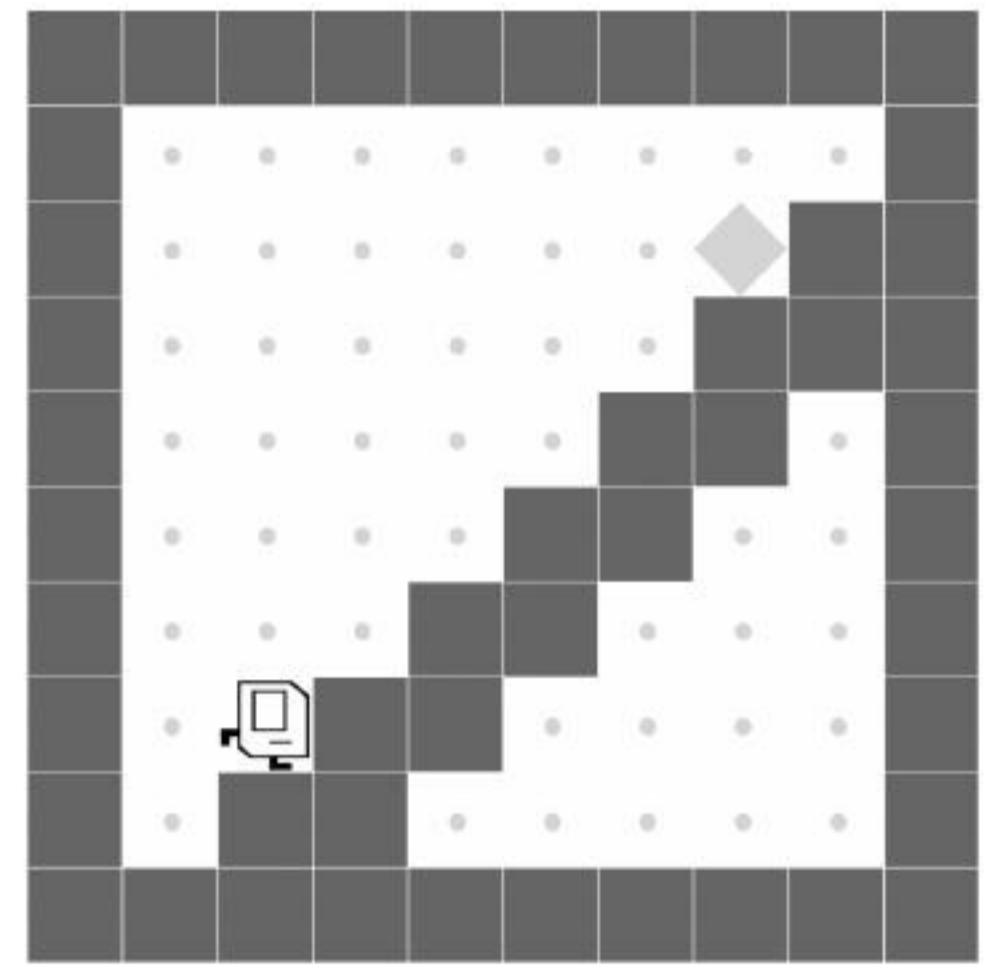
# Quantitative Results



# LEAPS Zero-shot Generalization

Learning on 8 x 8

StairClimber

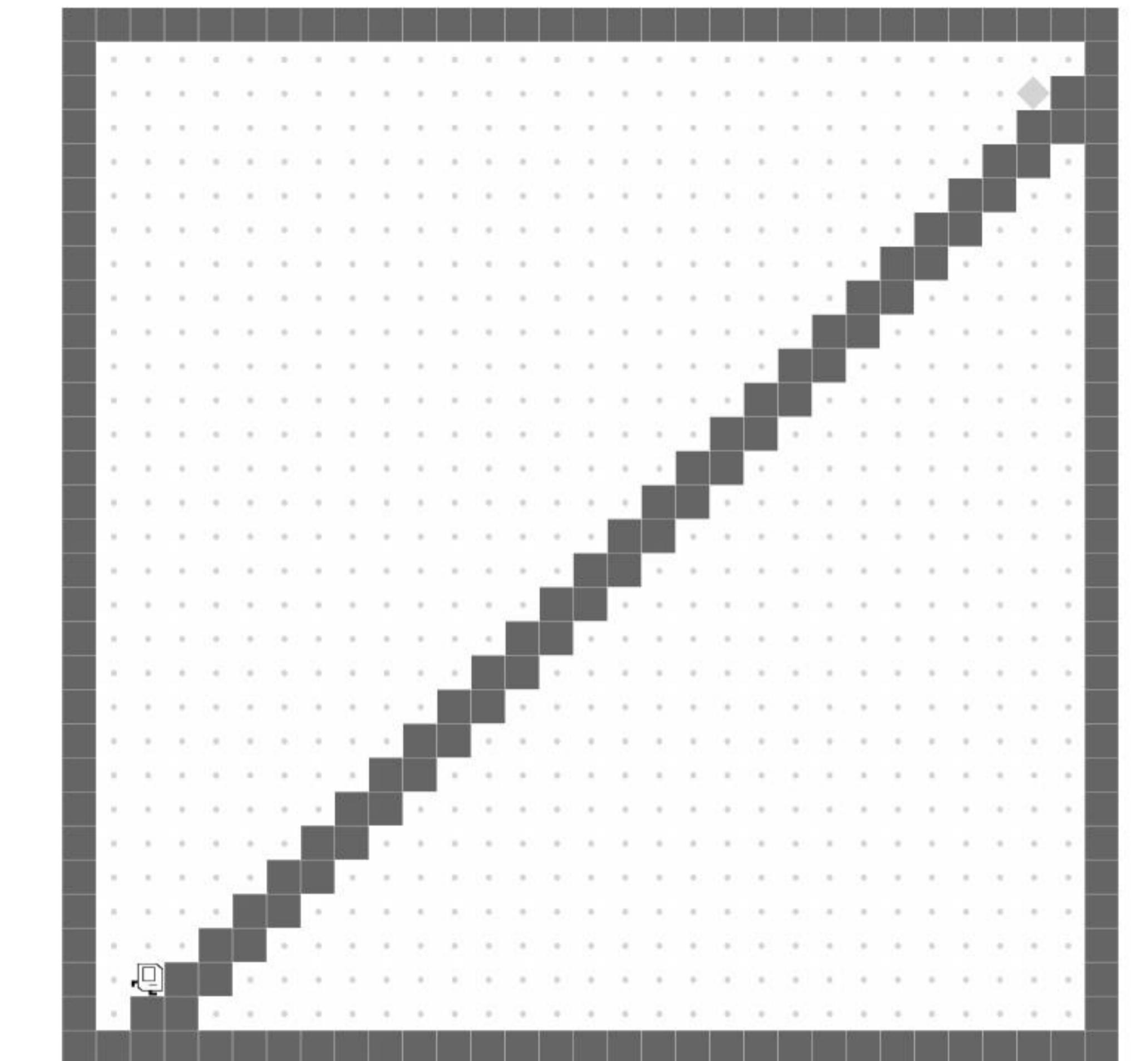


LEAPS  
Program  
Synthesizer

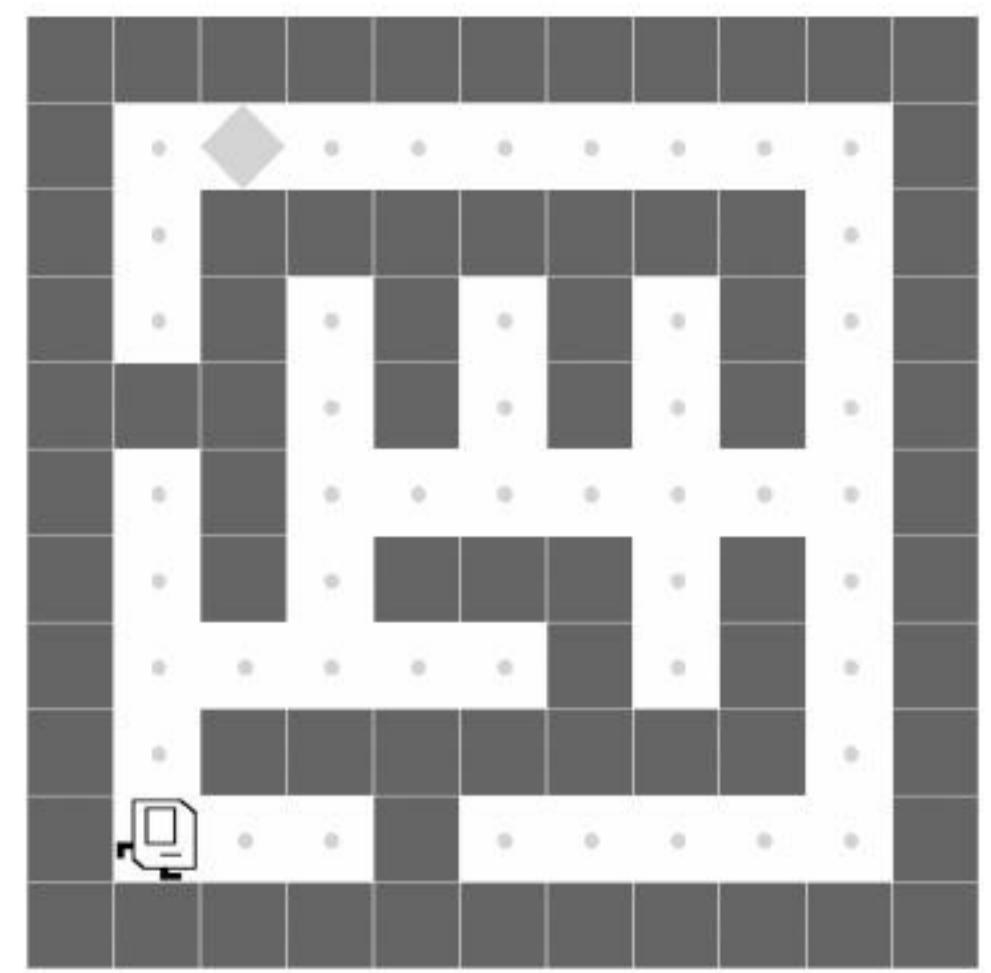
```
DEF run()
  while noMarkersPresent()
    turnRight
    move
  while rightIsClear()
    turnLeft
```

LEAPS Program Policy

Evaluation on 100 x 100



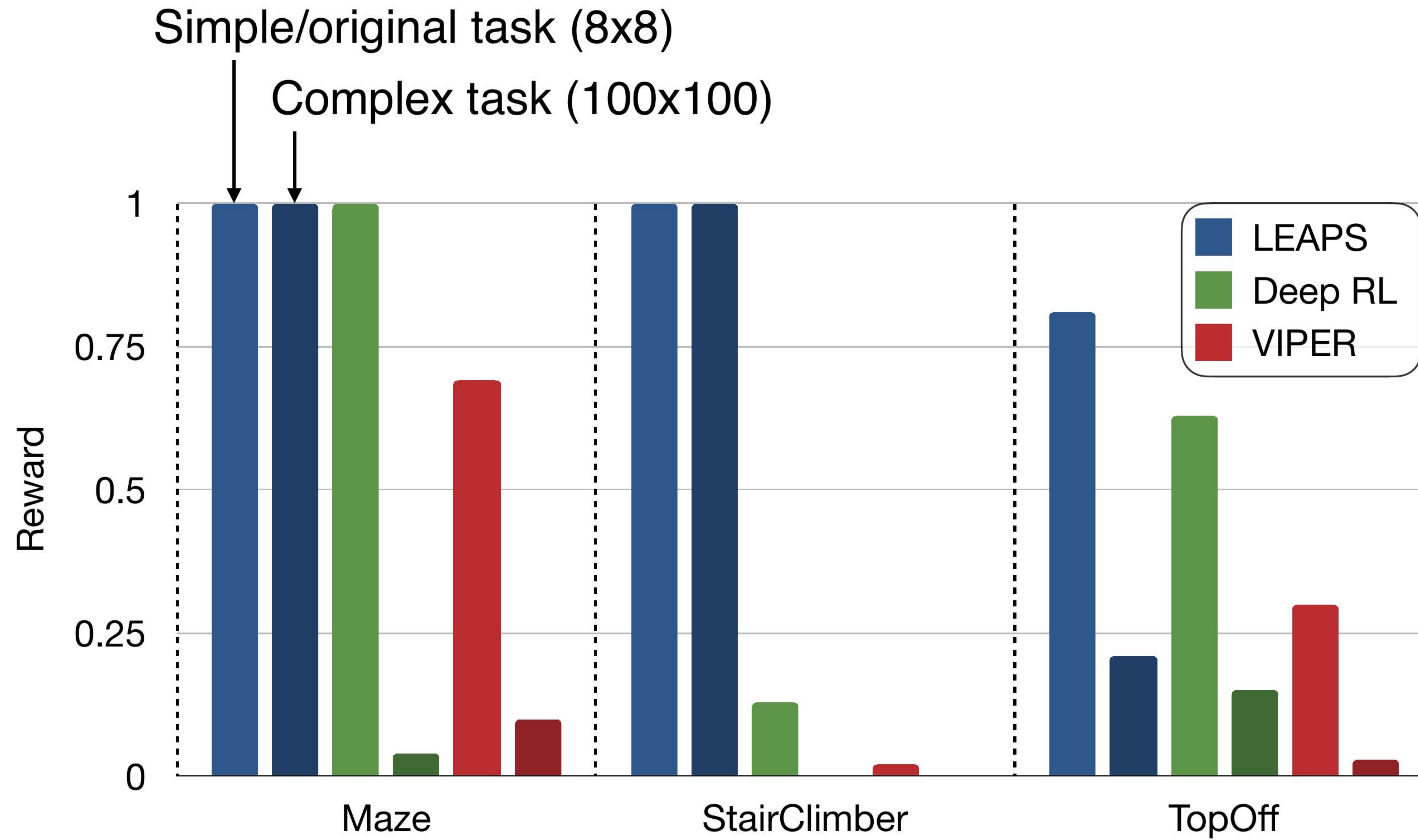
Maze



LEAPS  
Program  
Synthesizer

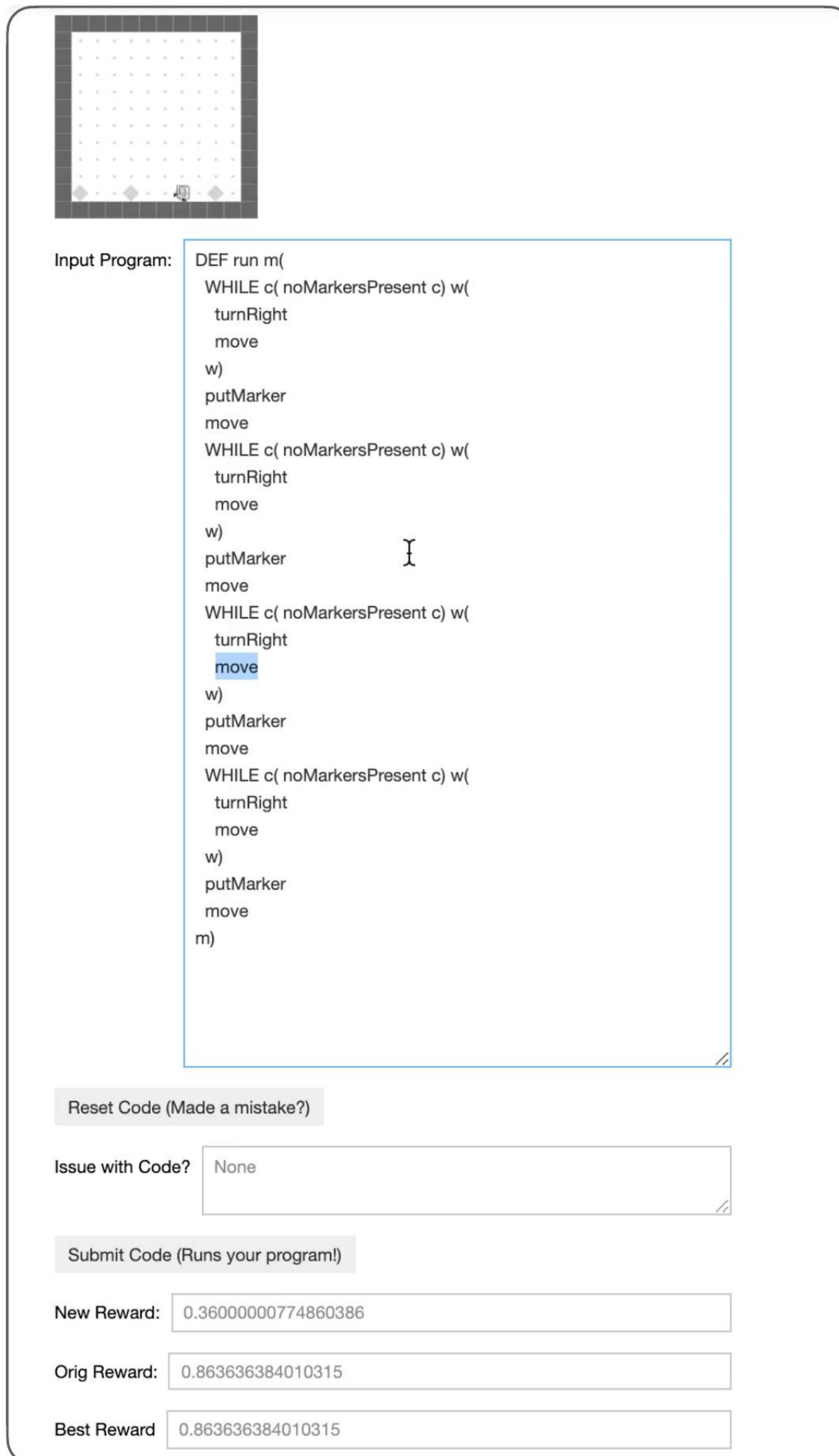
```
DEF run()
  if frontIsClear()
    turnLeft
  while noMarkersPresent()
    turnRight
    move
```

# Experimental Results - Zero-shot Generalization

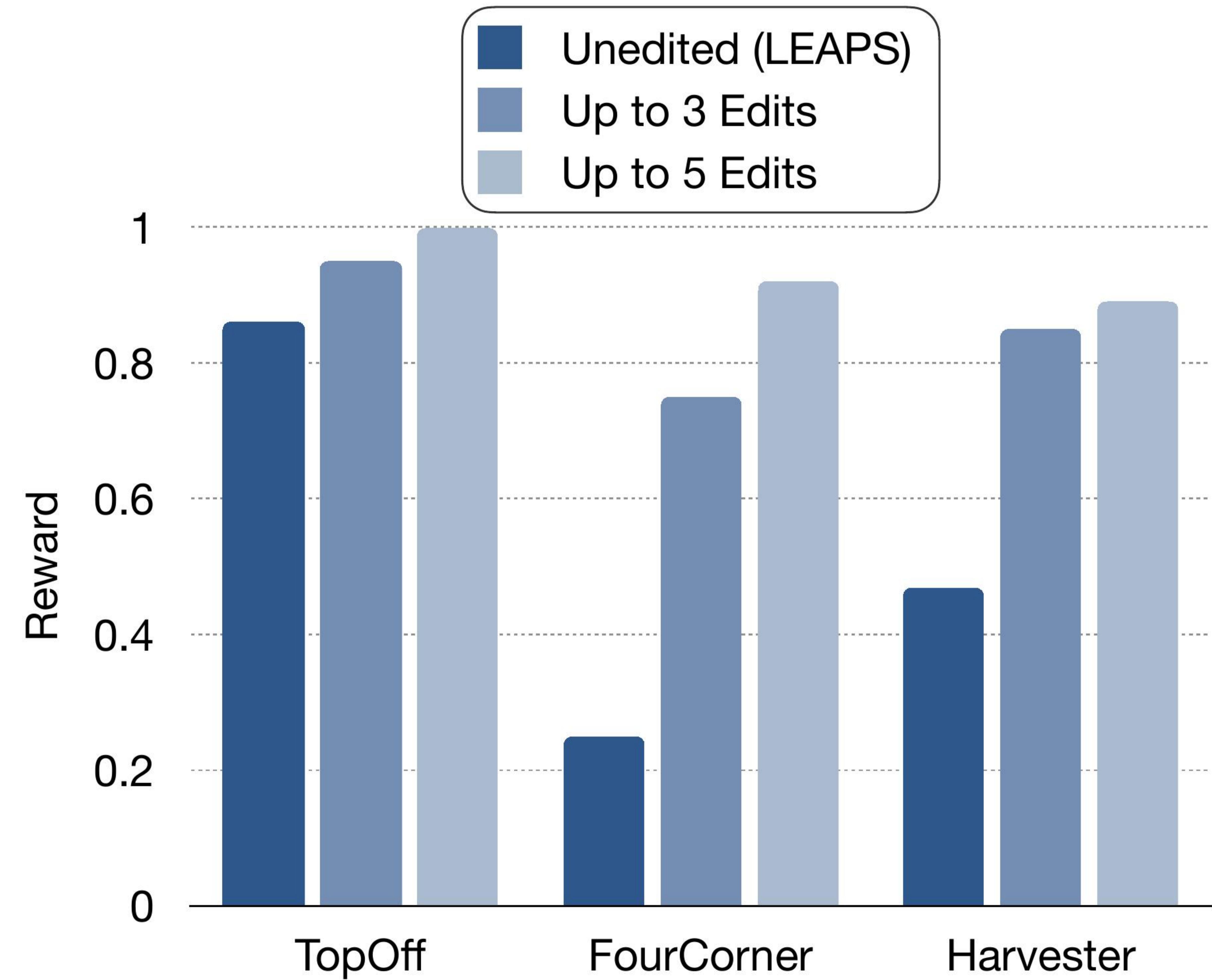


# Interpretability & Interactivity

## Interactive Debugging Interface

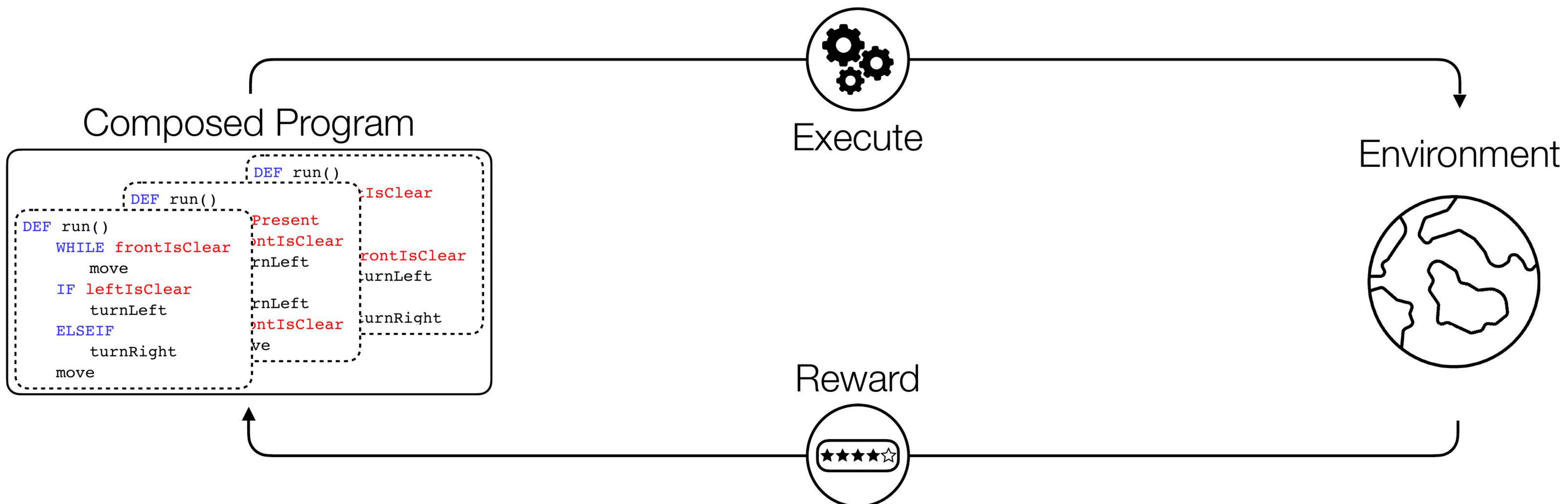


## Performance Improvement



# Hierarchical Programmatic Reinforcement Learning via Learning to Compose Programs

ICML 2023



Guan-Ting Liu



En-Pei Hu



Pu-Jen Cheng



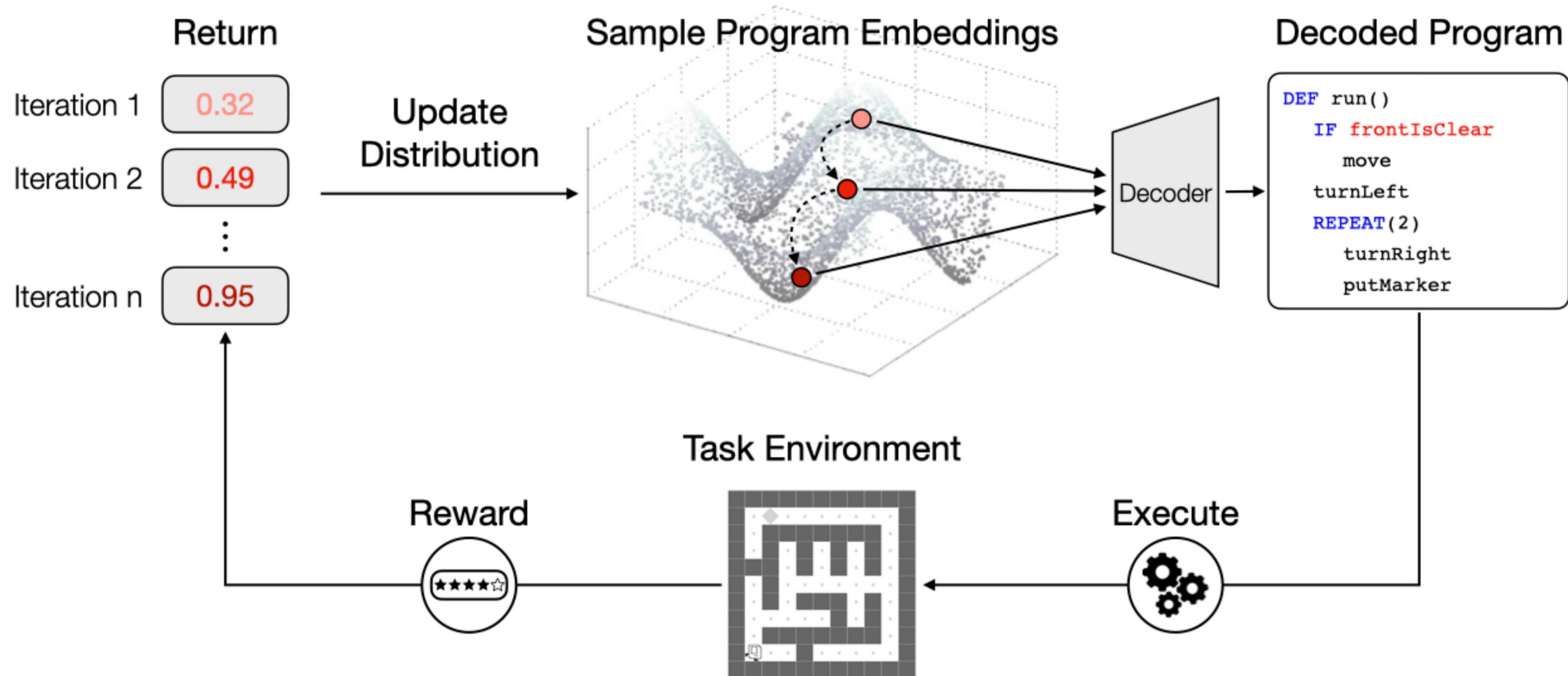
Hung-Yi Lee



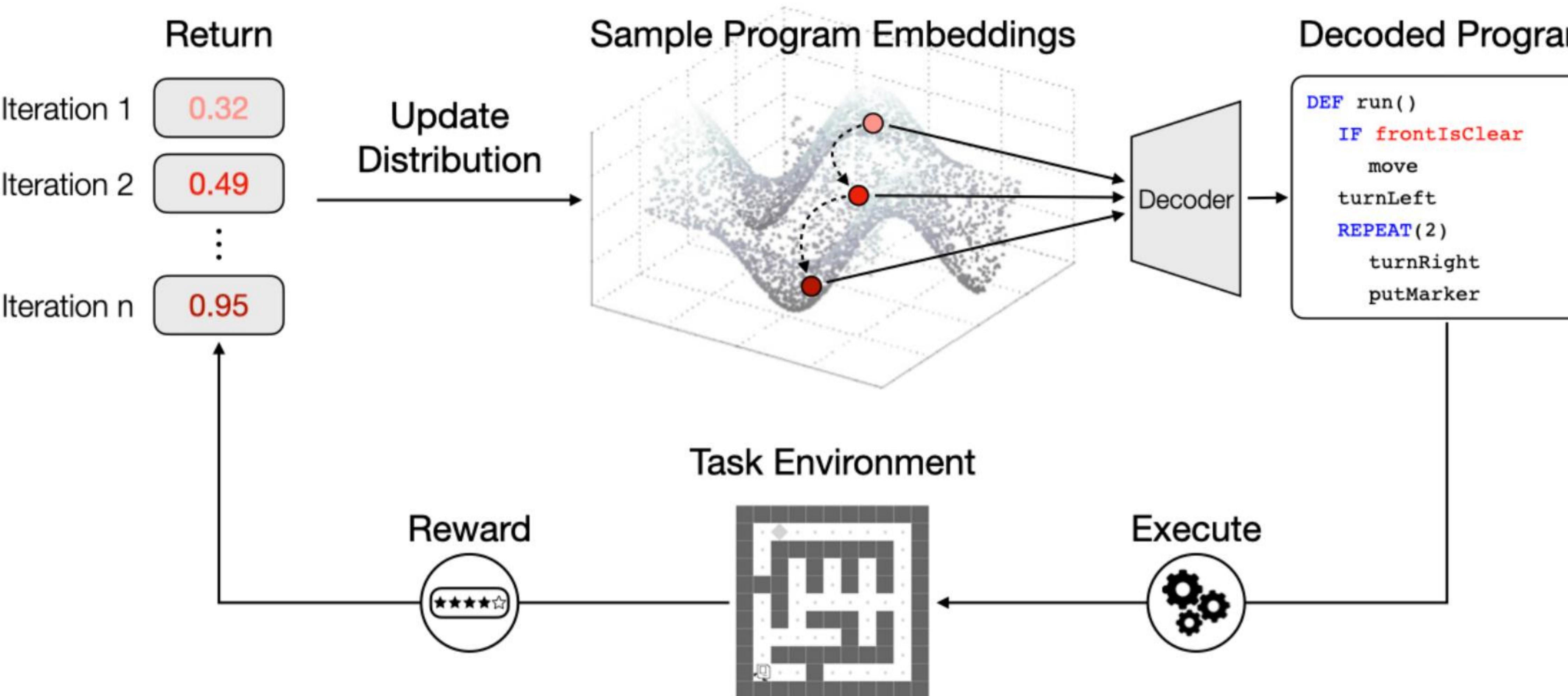
Shao-Hua Sun

# LEAPS: Learning Embeddings for Latent Program Synthesis

Stage 2 Searching for a task-solving program using the cross-entropy method



## Stage 2 Searching for a task-solving program using the cross-entropy method

**Limited program distribution**

Search in the program embedding space spanned  
by the dataset programs



Cannot synthesize longer or  
more complex programs

**Poor credit assignment**

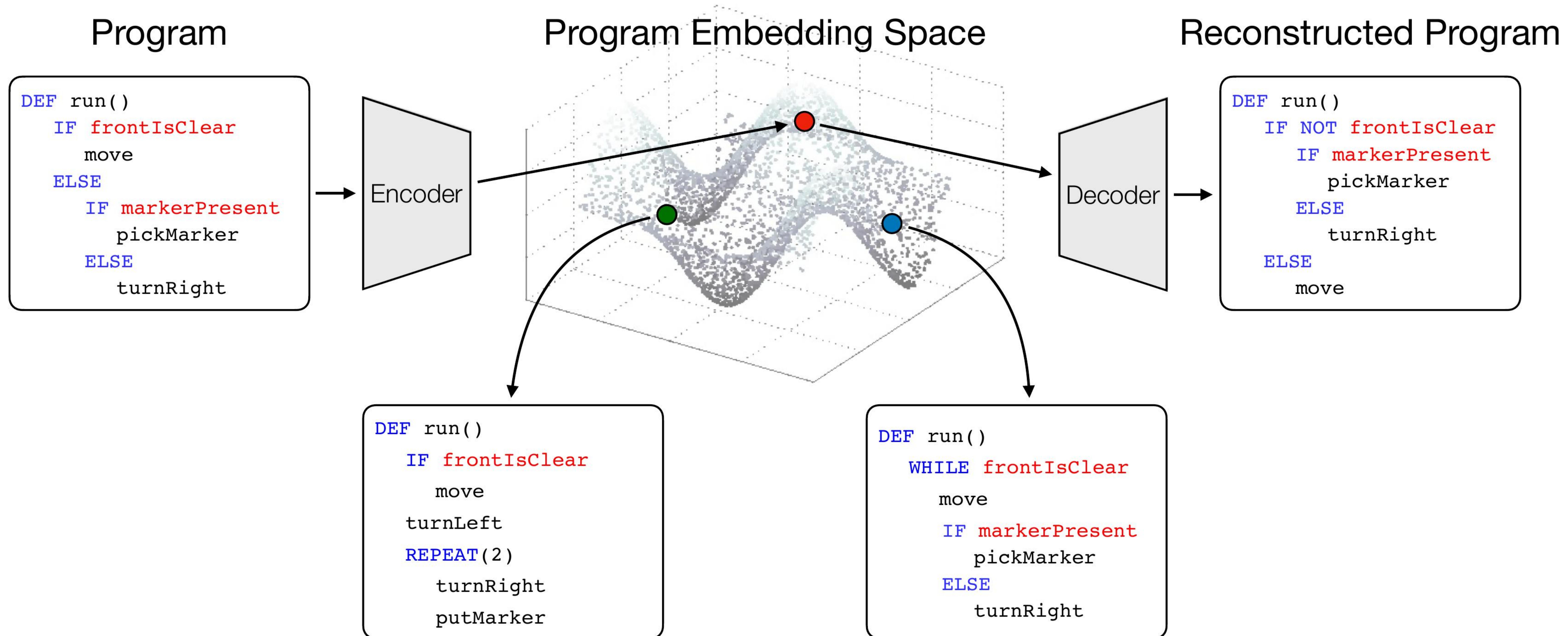
Evaluate each candidate program solely based on  
the **cumulative return** of its execution trace



Cannot accurately attribute rewards to  
corresponding program parts


# HPRL: Hierarchical Programmatic Reinforcement Learning

Stage 1 Learning a **compressed** program embedding space from randomly generated programs

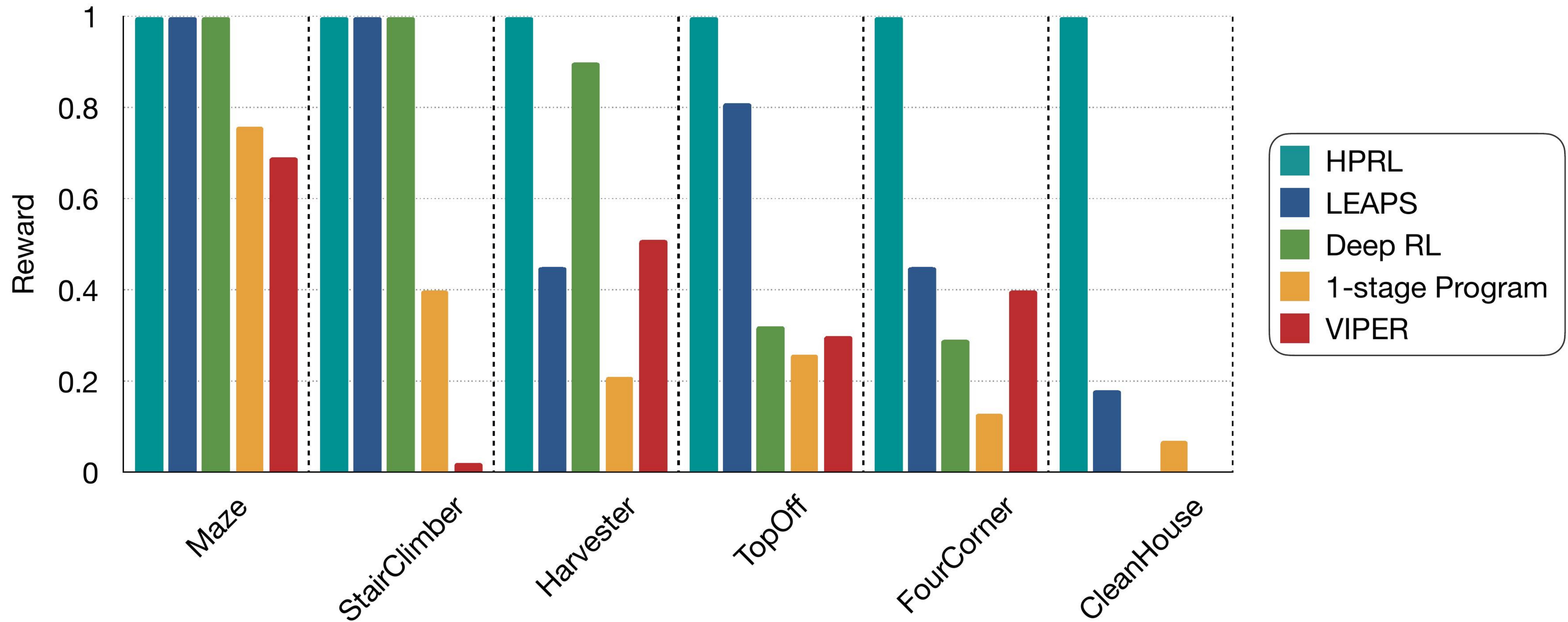


# HPRL: Hierarchical Programmatic Reinforcement Learning

Stage 2 Learning a meta policy to produce a series of programs (i.e., predict a series of actions) to yield a composed task-solving program

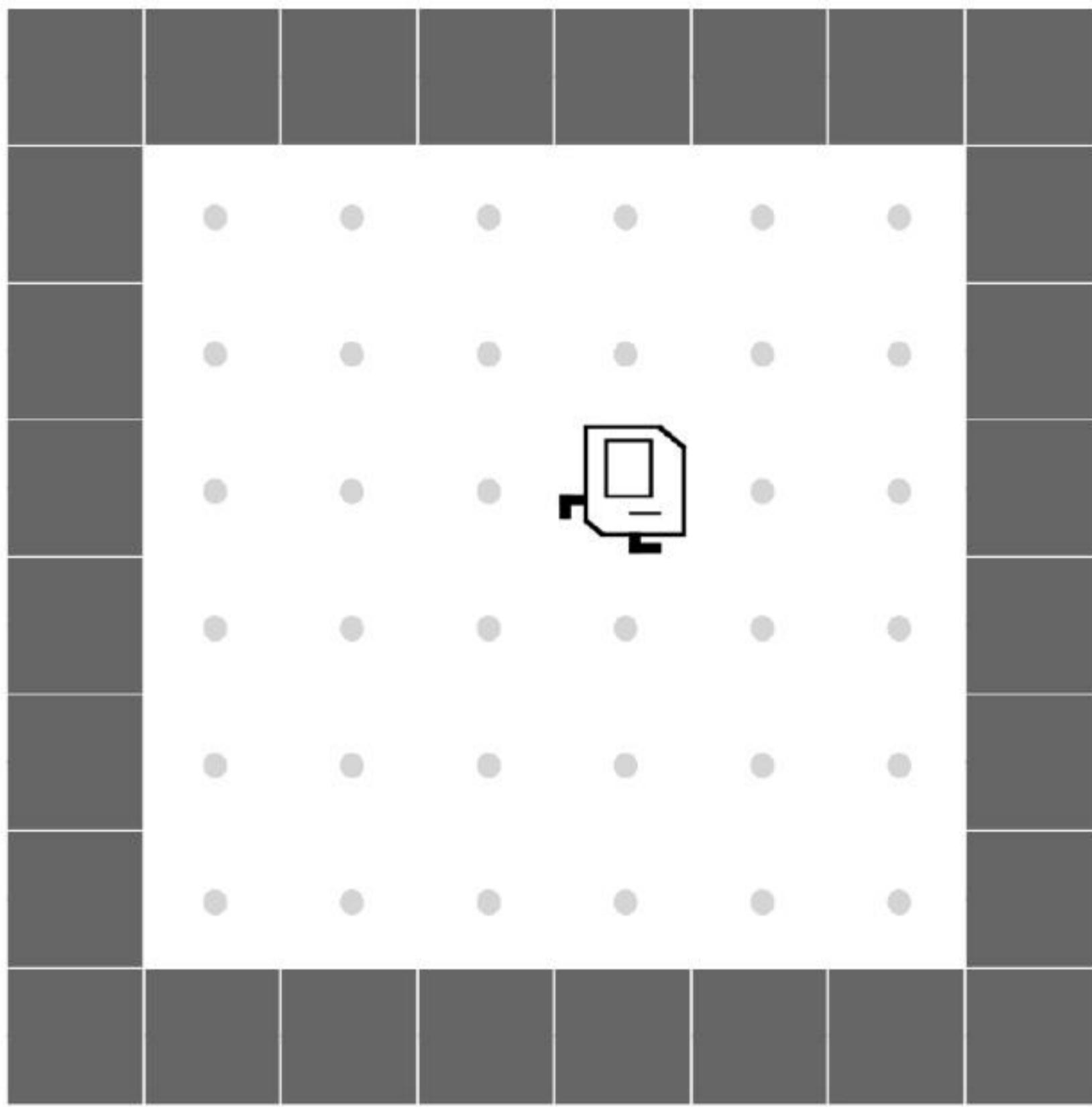


# Quantitative Results - Karel Tasks

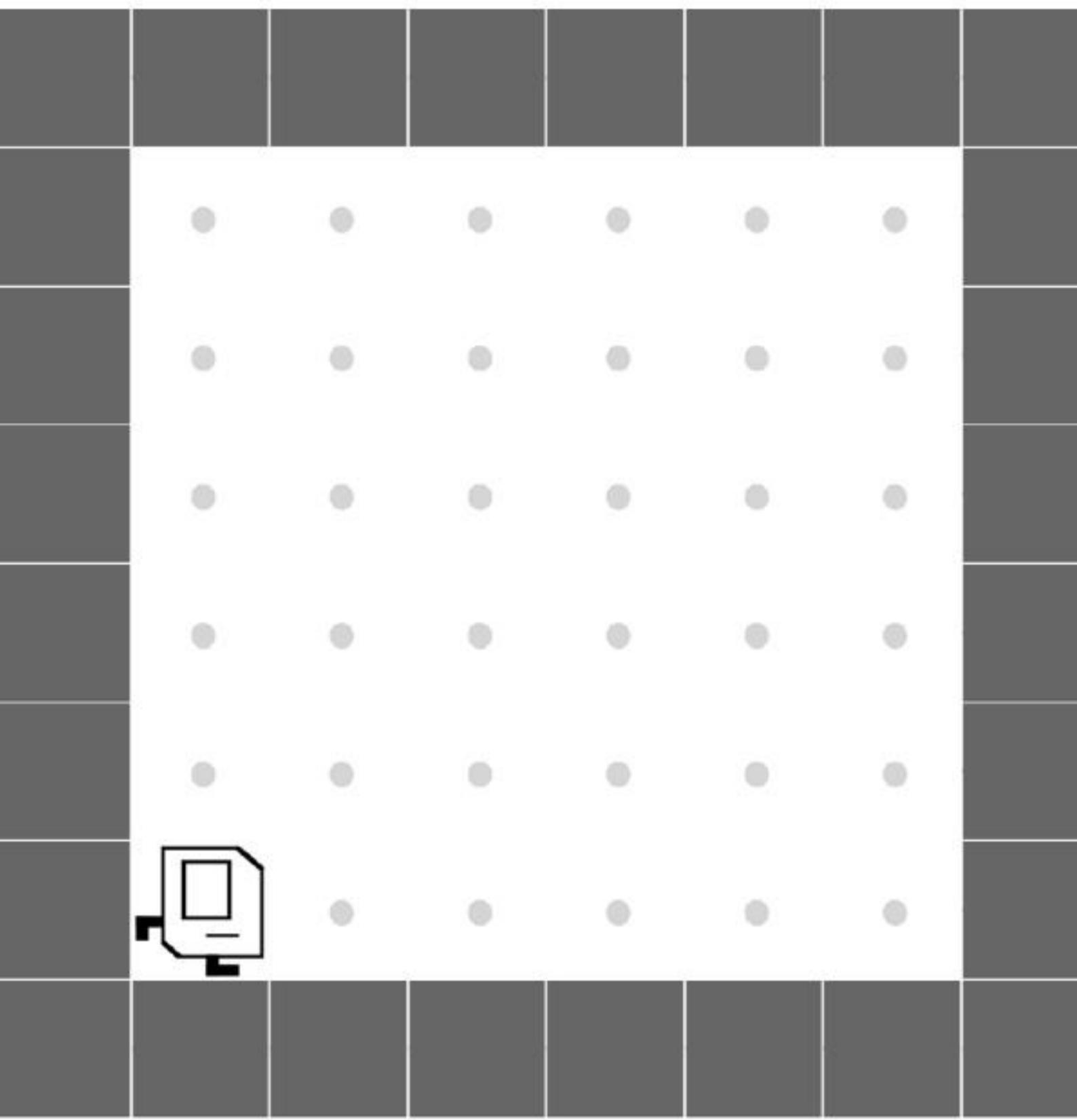


# Karel-Hard Tasks

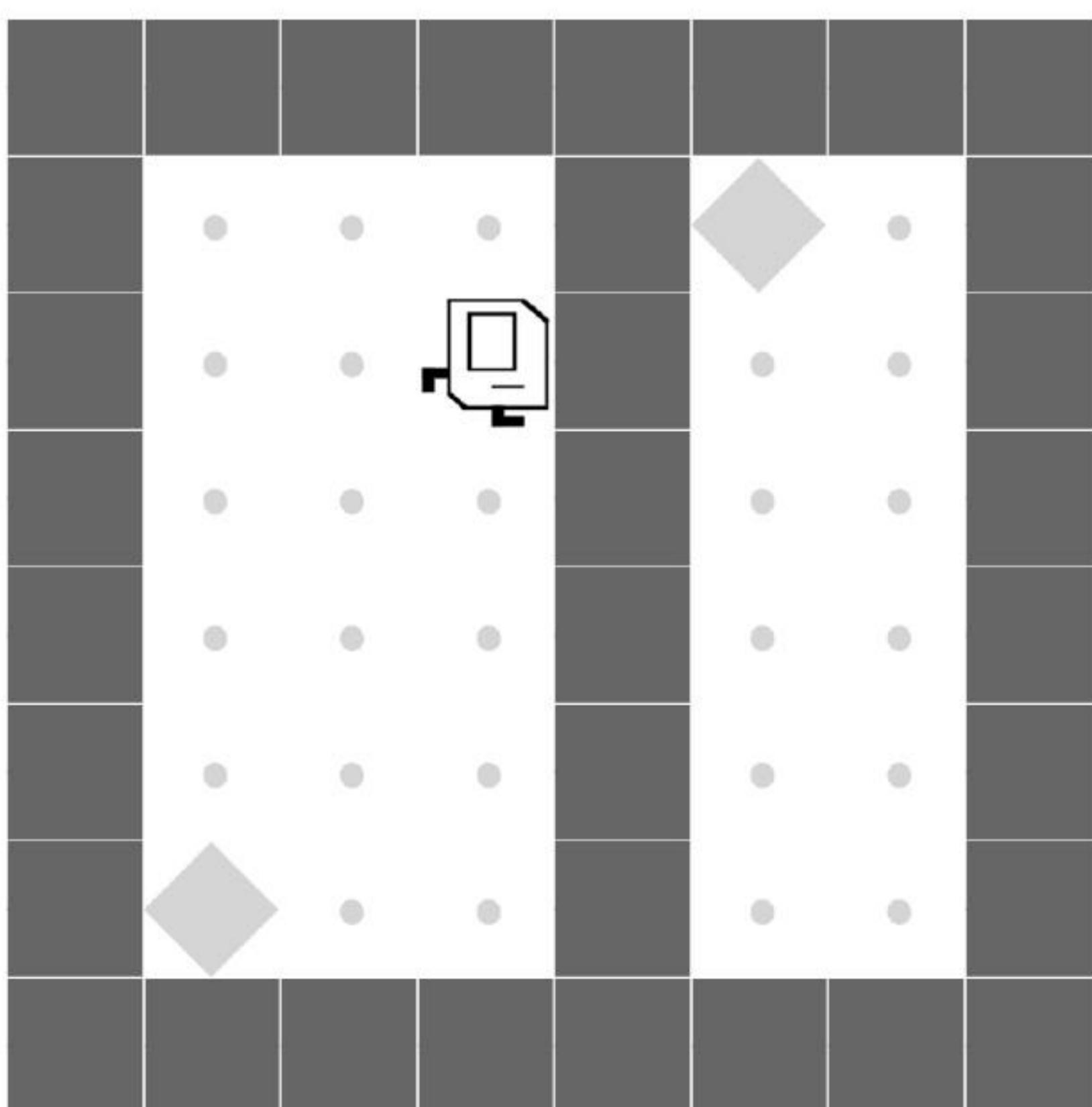
OneStroke



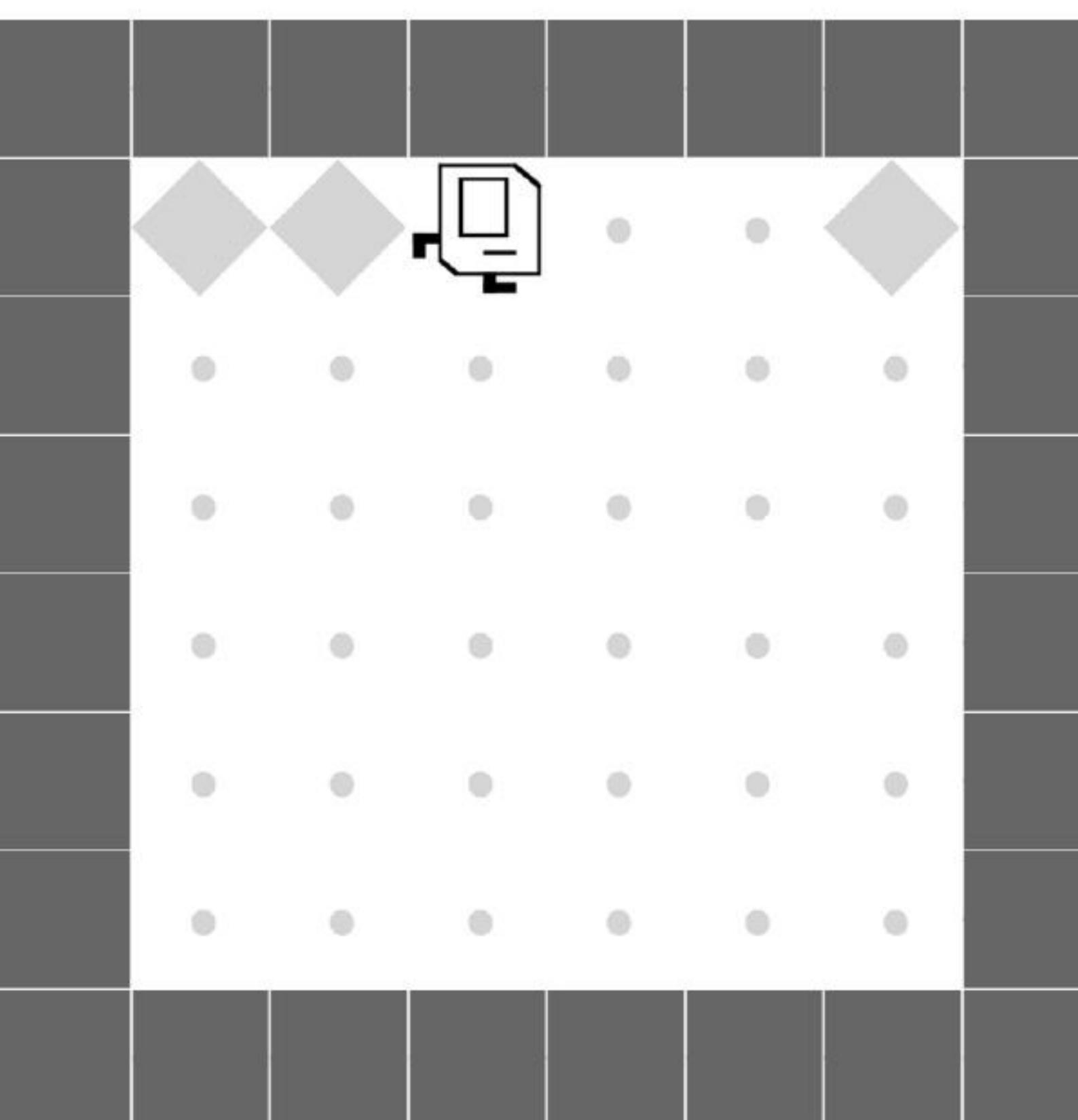
Seeder



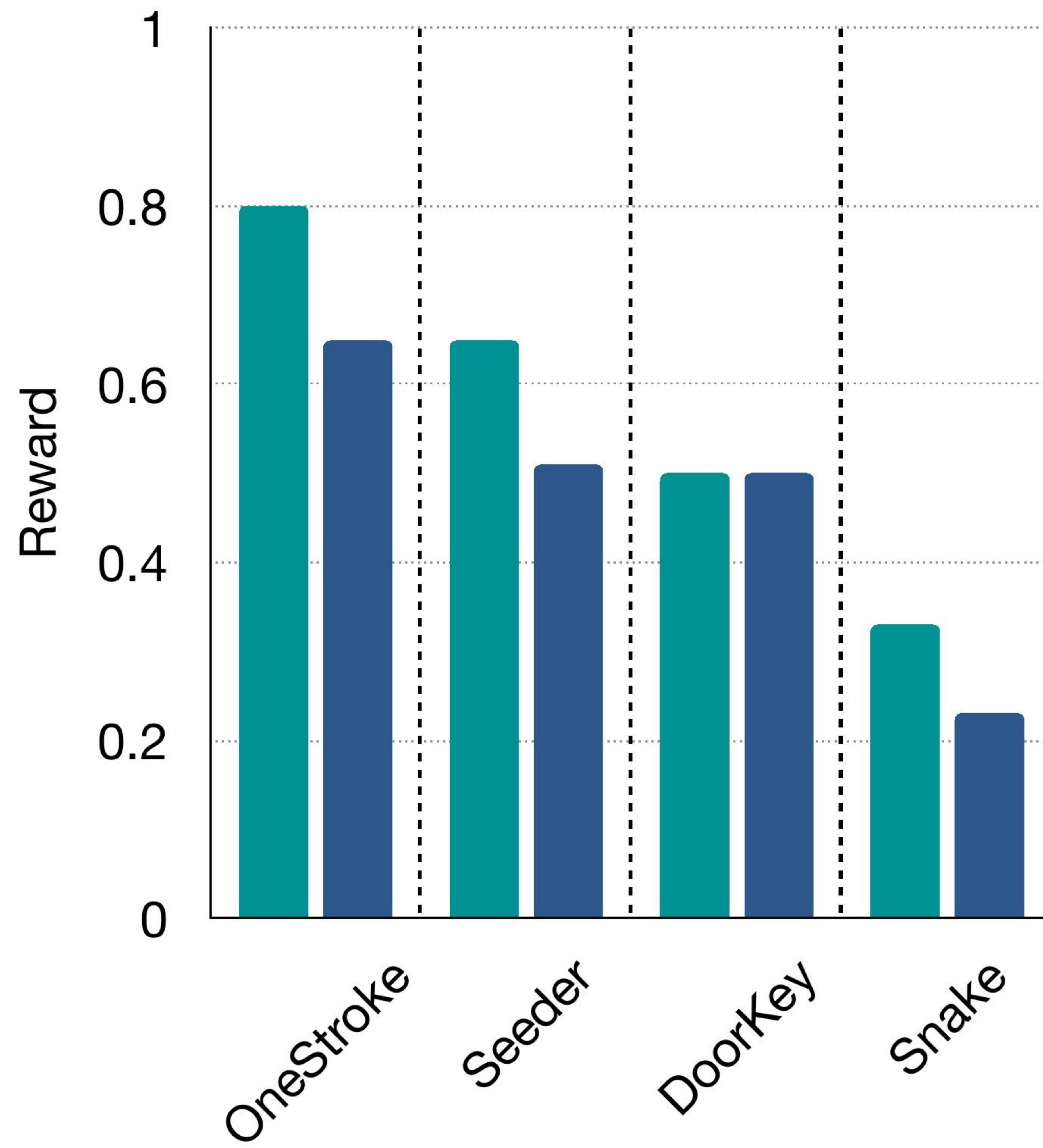
DoorKey



Snake



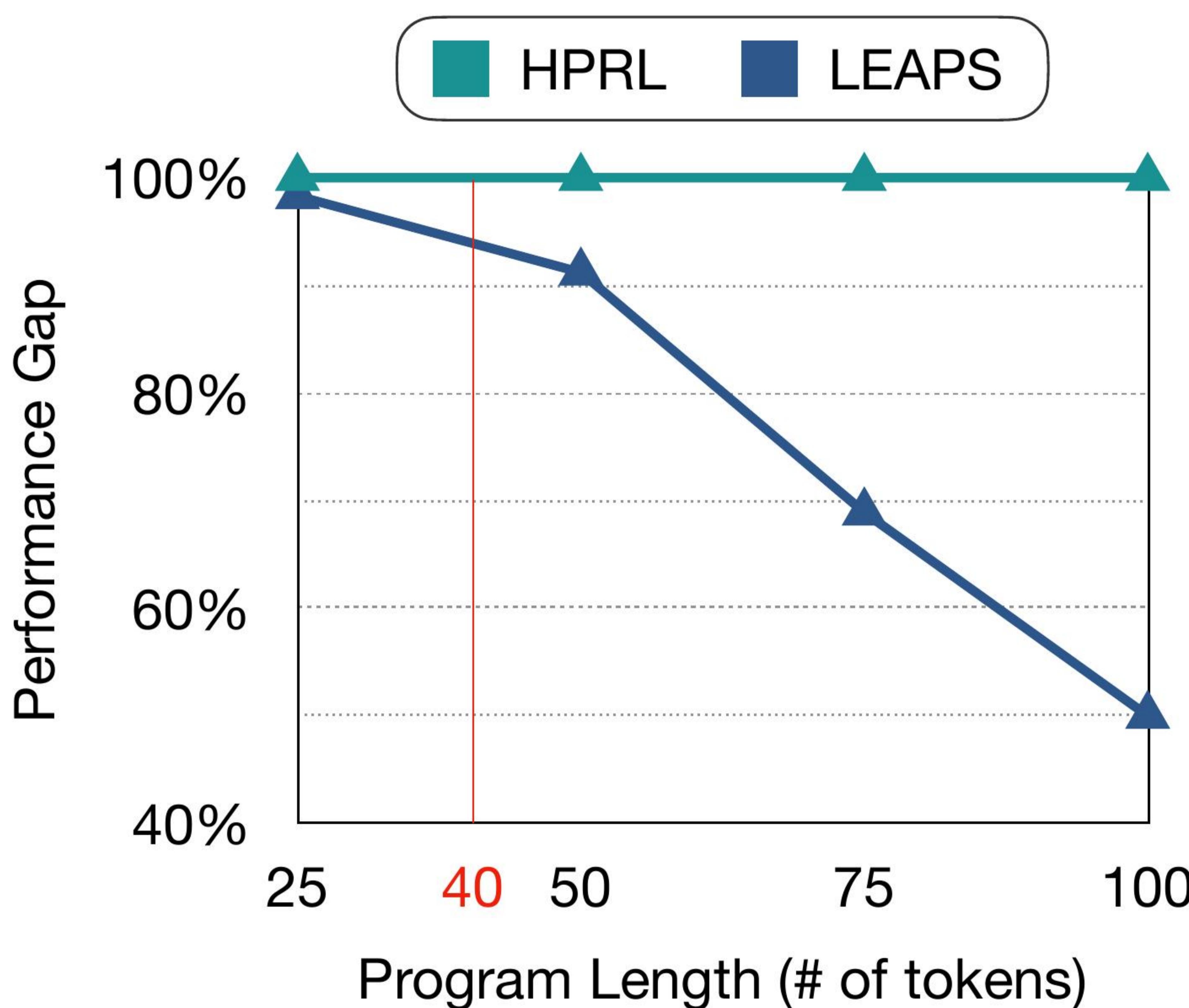
HPRL LEAPS



# Additional Experiments

## Limited program distribution

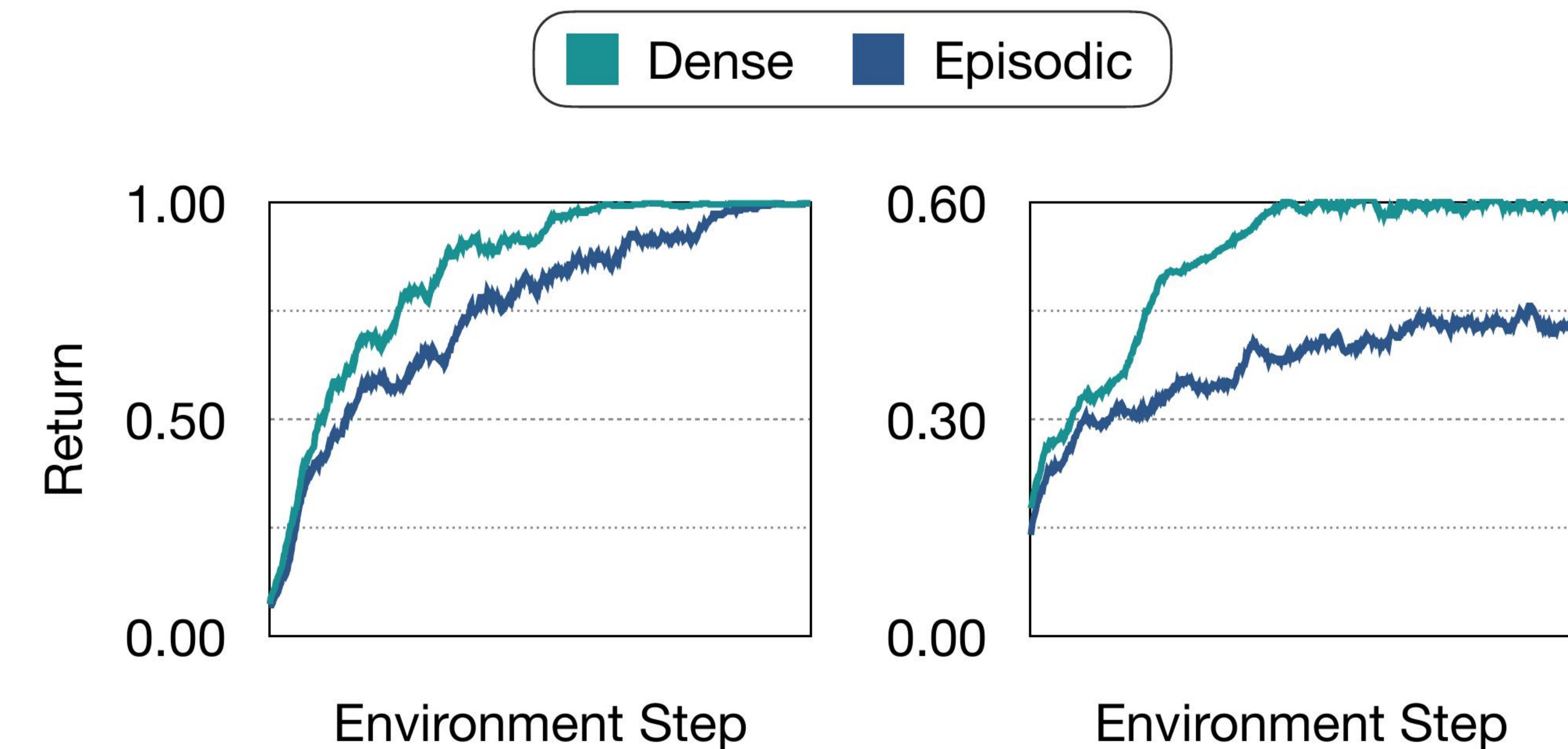
Synthesize out-of-distributionally long programs



## Poor credit assignment

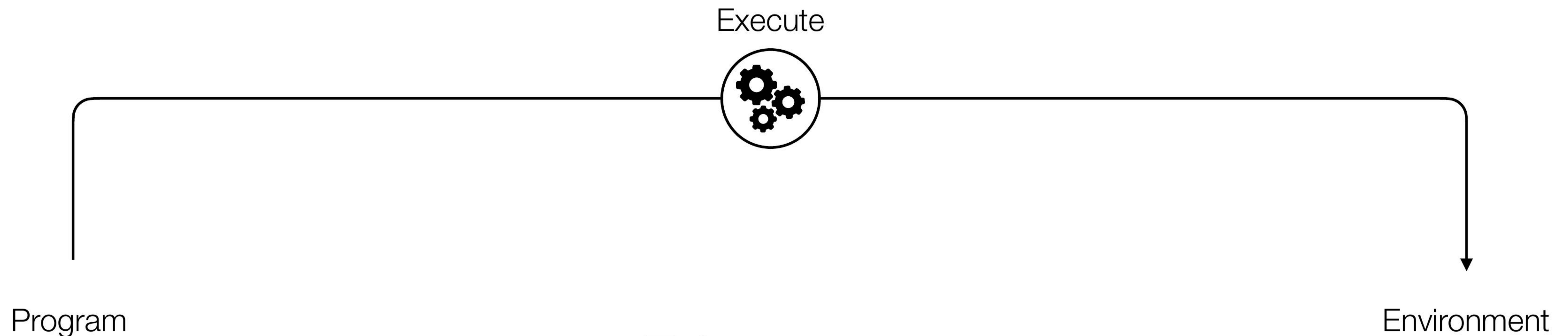
Learning from episodic reward

- **Dense:** Reward each subprogram based on its execution trace
- **Episodic:** Reward the entire composed program at the end



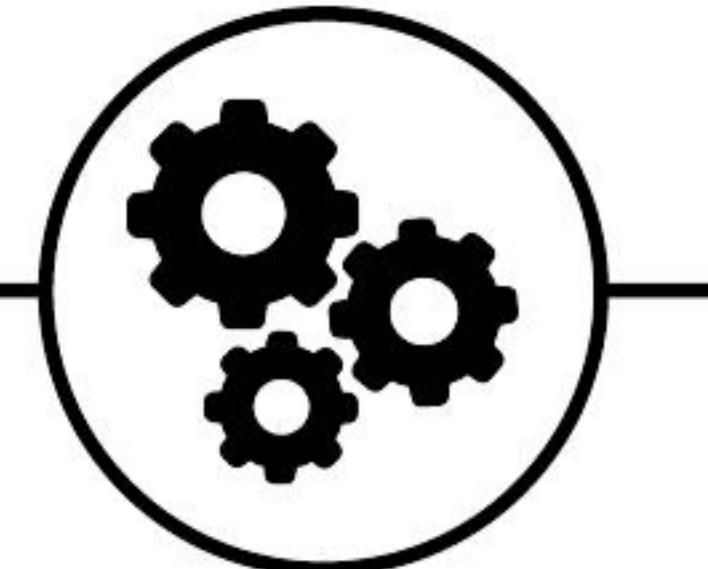
- **HPRL** can synthesize programs longer than the dataset programs (< 40 tokens) better than **LEAPS**

- The hierarchical design of **HPRL** allows for better credit assignment with dense rewards, facilitating the learning progress



```
DEF run() m(
  WHILE c( markerPresent c) w(
    WHILE c( markerPresent c) w(
      pickMarker
      move w)
    turnRight
    move
    turnLeft
    WHILE c( markerPresent c) w(
      pickMarker
      move w)
    turnLeft
    move
    turnRight w) m)
```

Execute

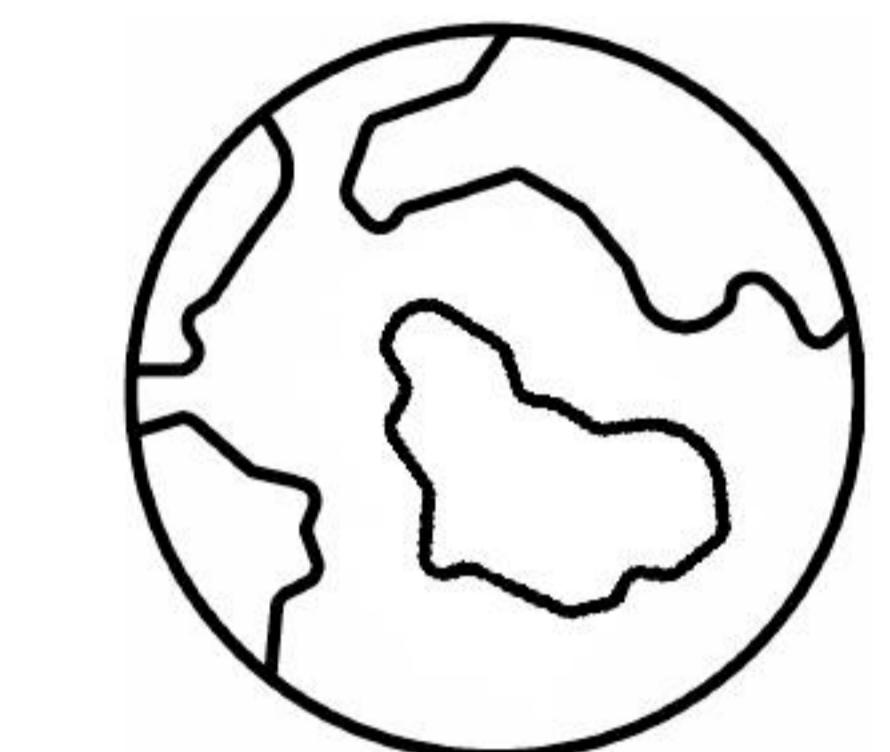


## Takeaways

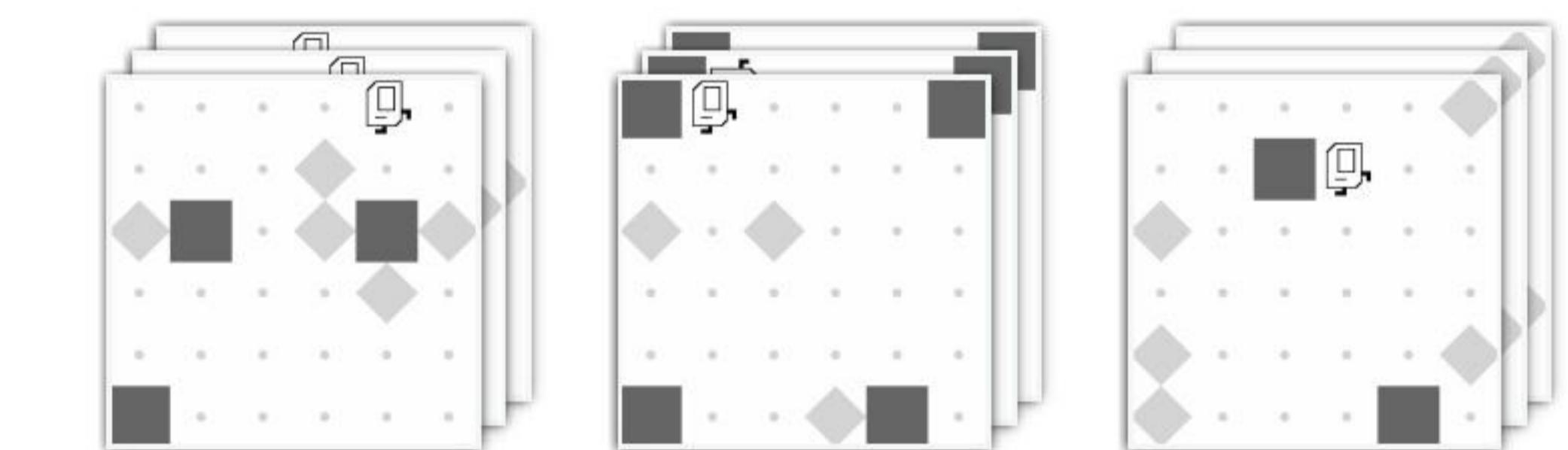
Program Synthesis  $\times$  Reinforcement Learning

= Interpretable and Generalizable Policies

Environment

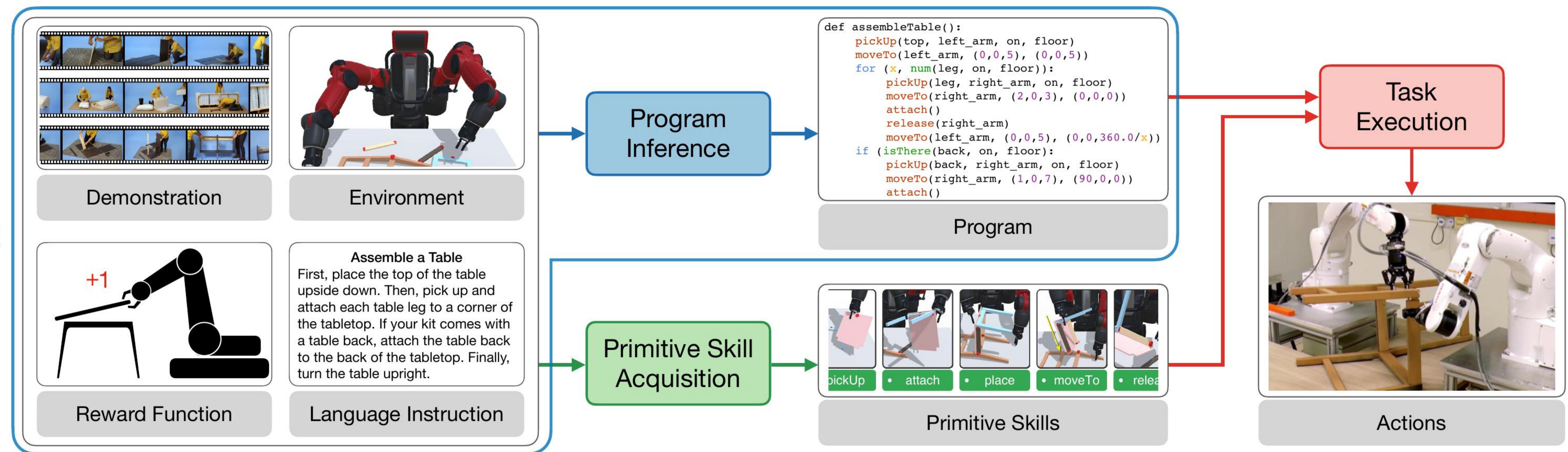


Demonstrations

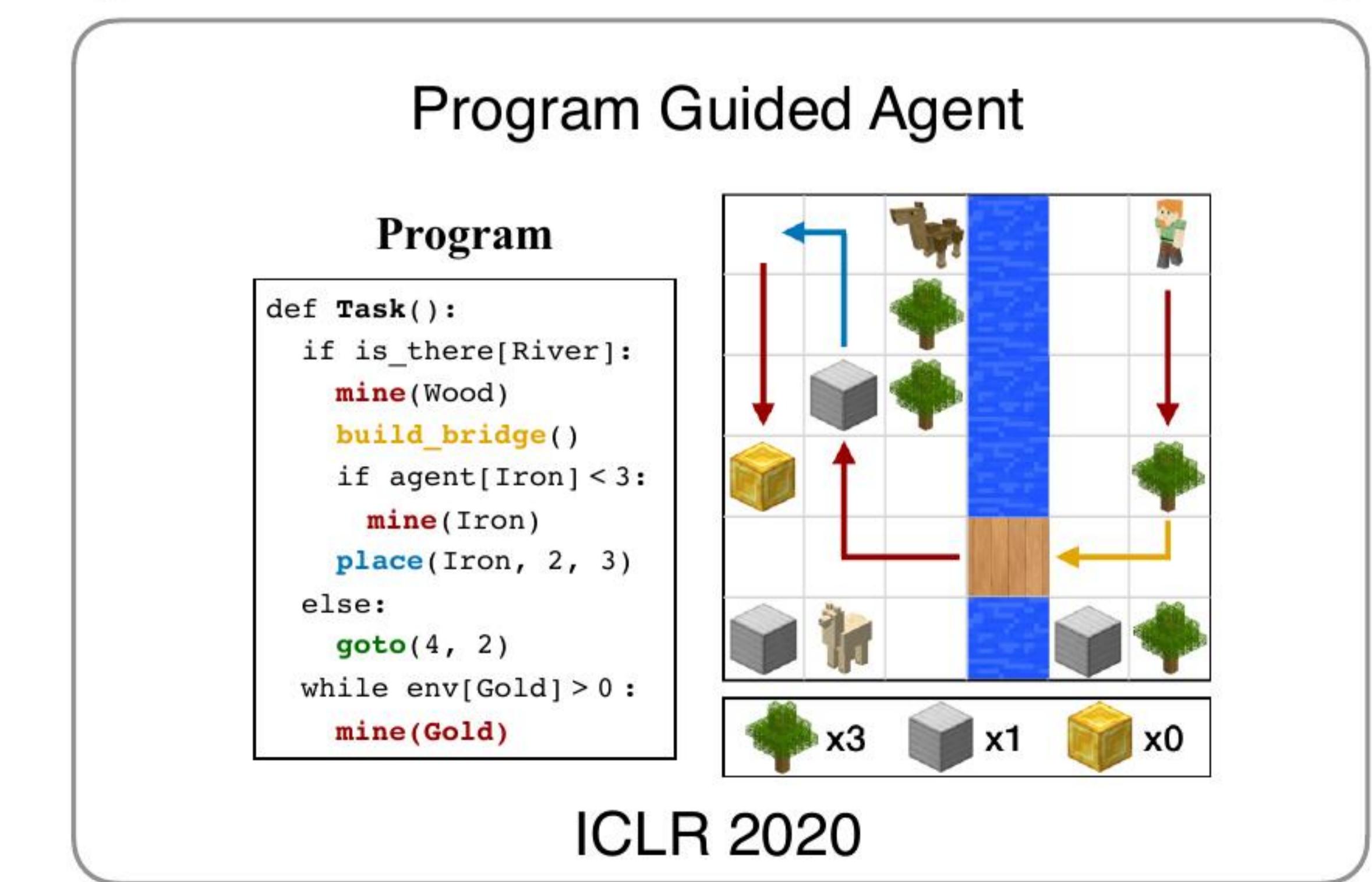
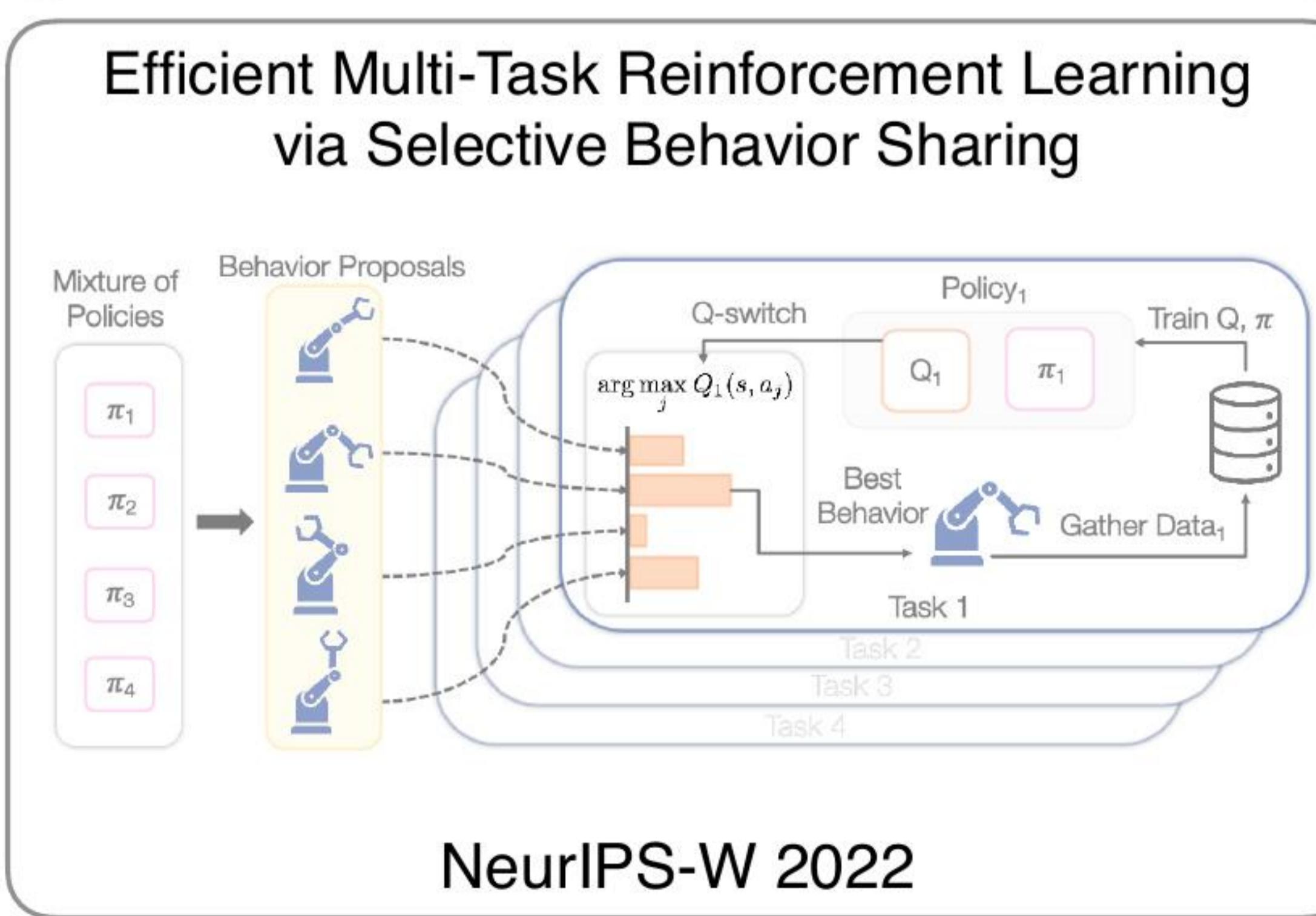
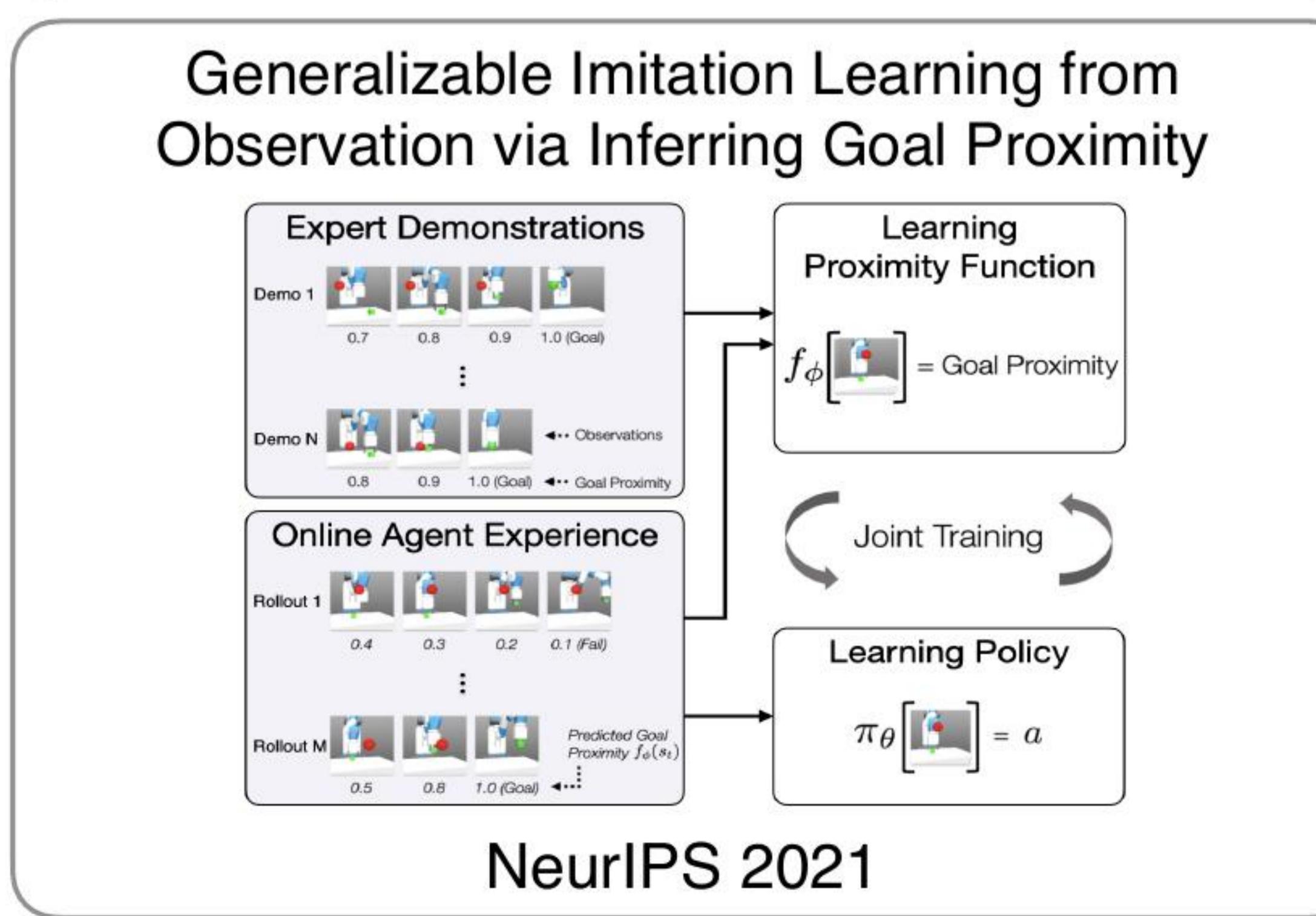
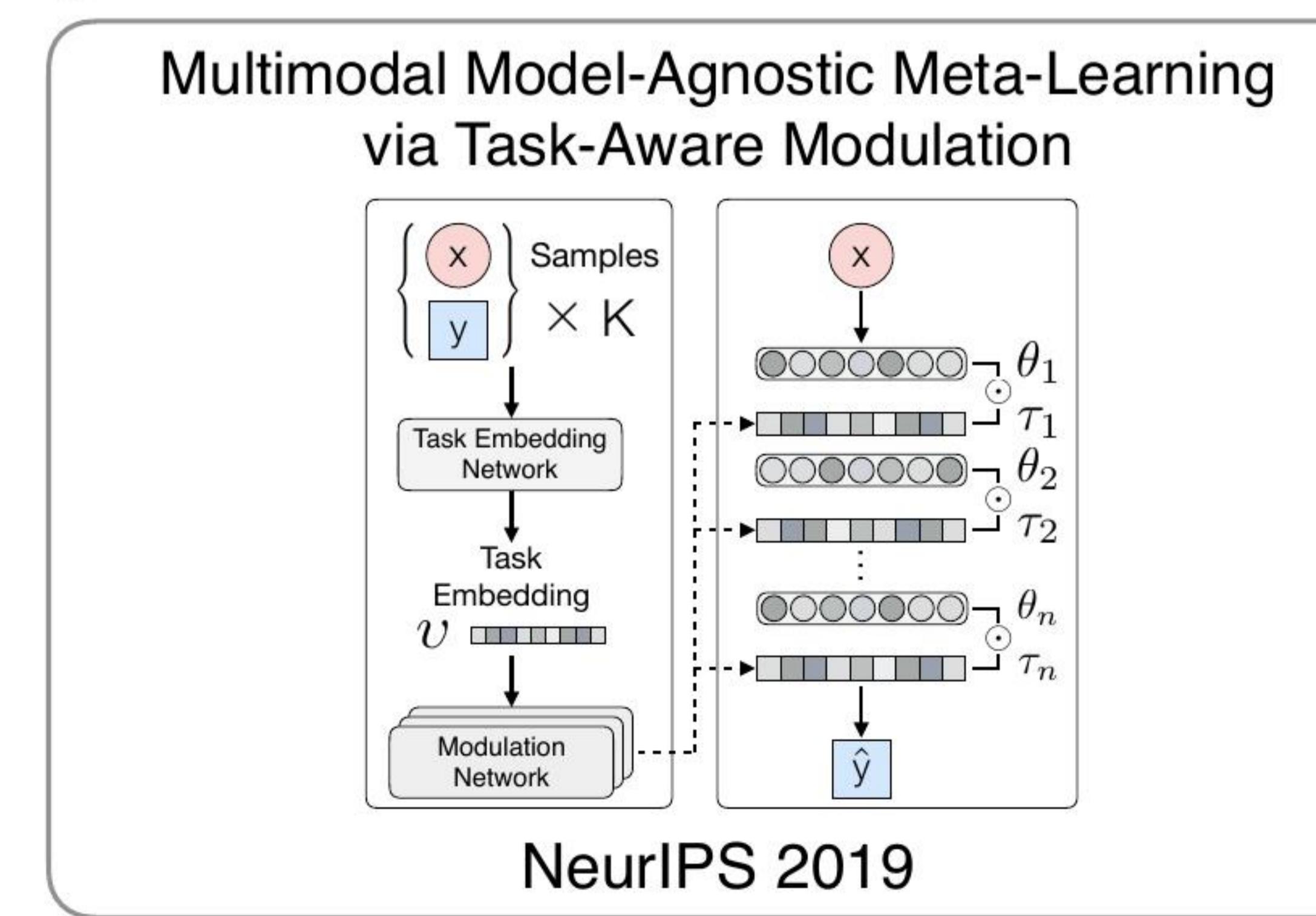
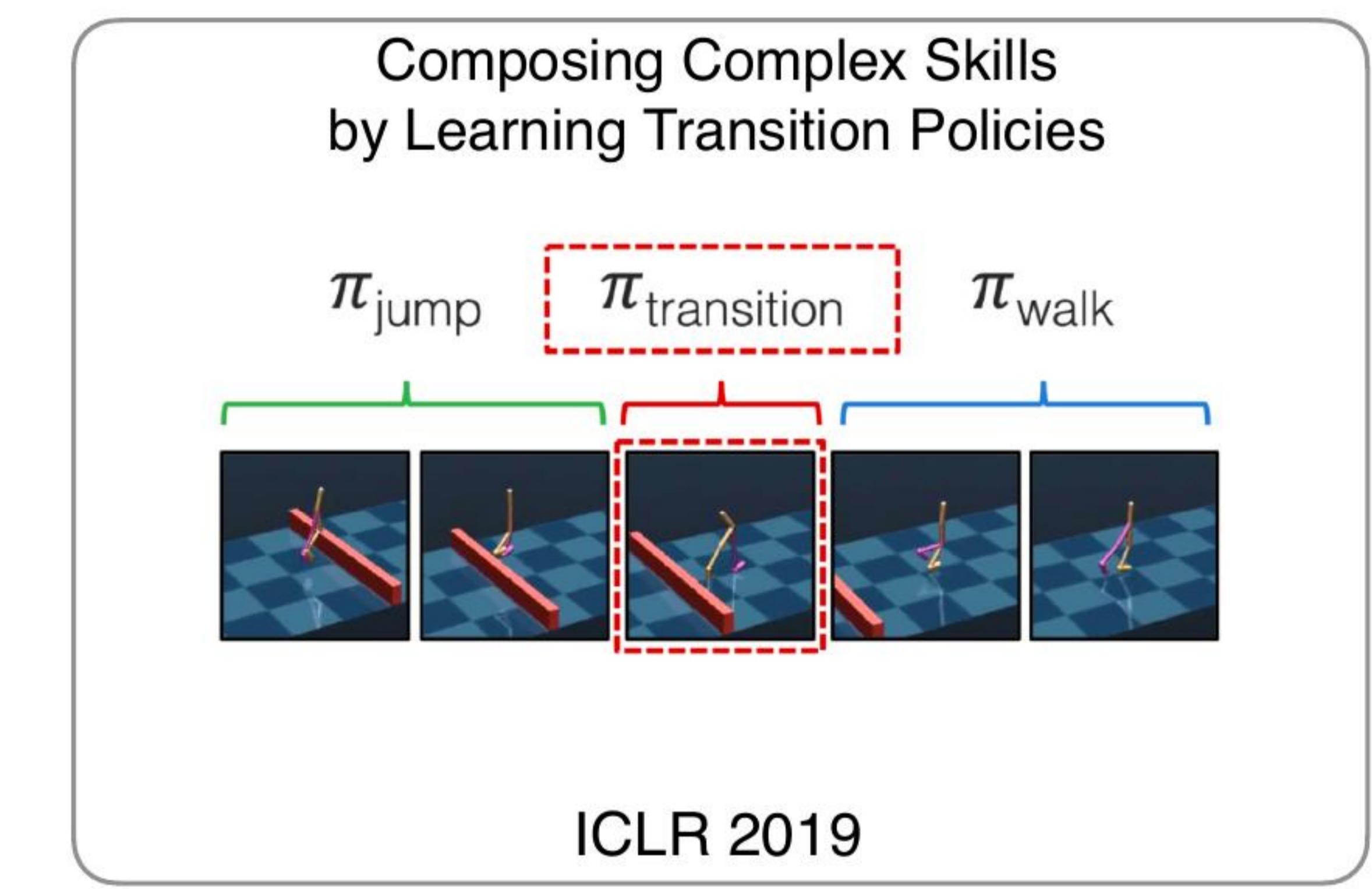
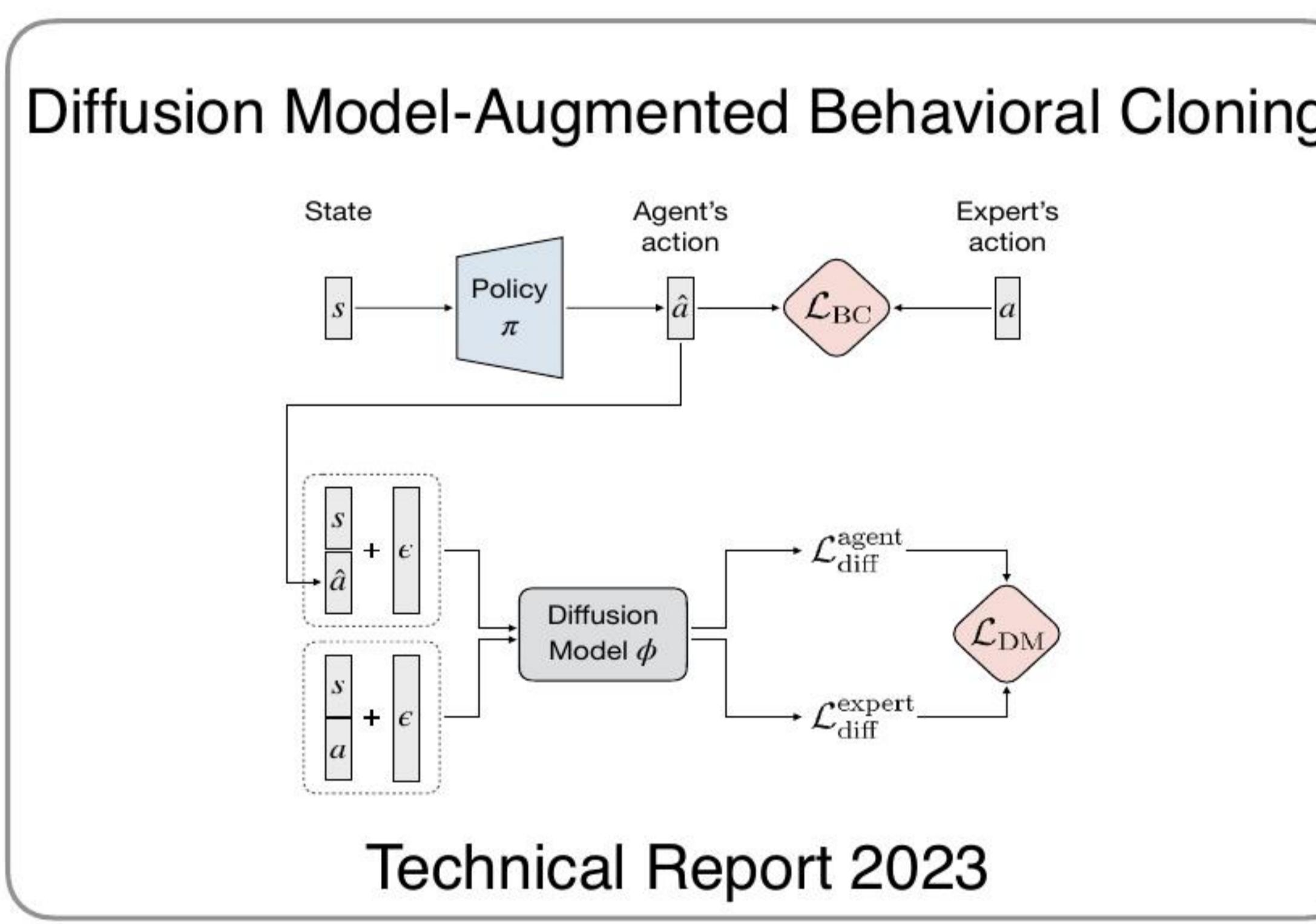
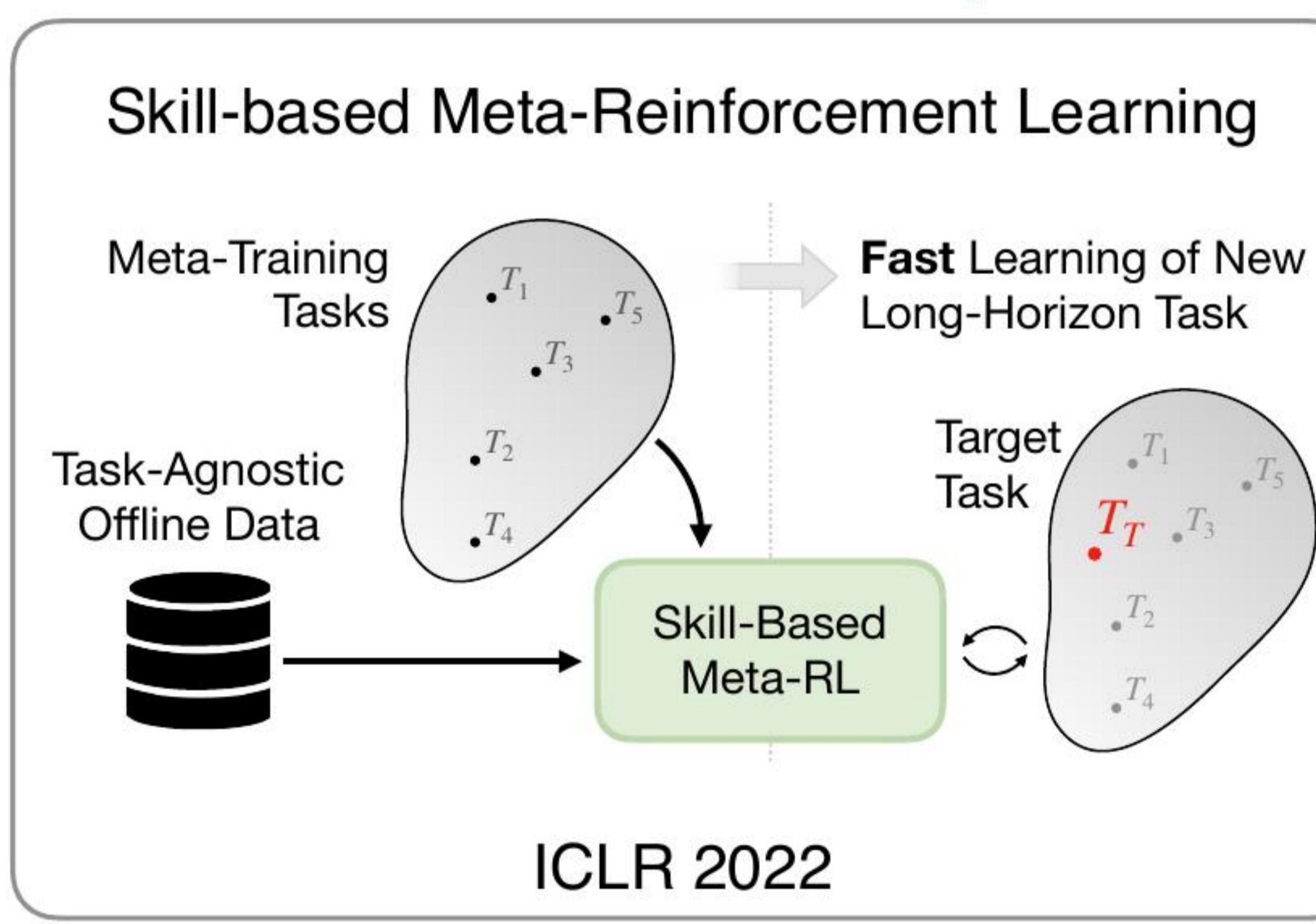
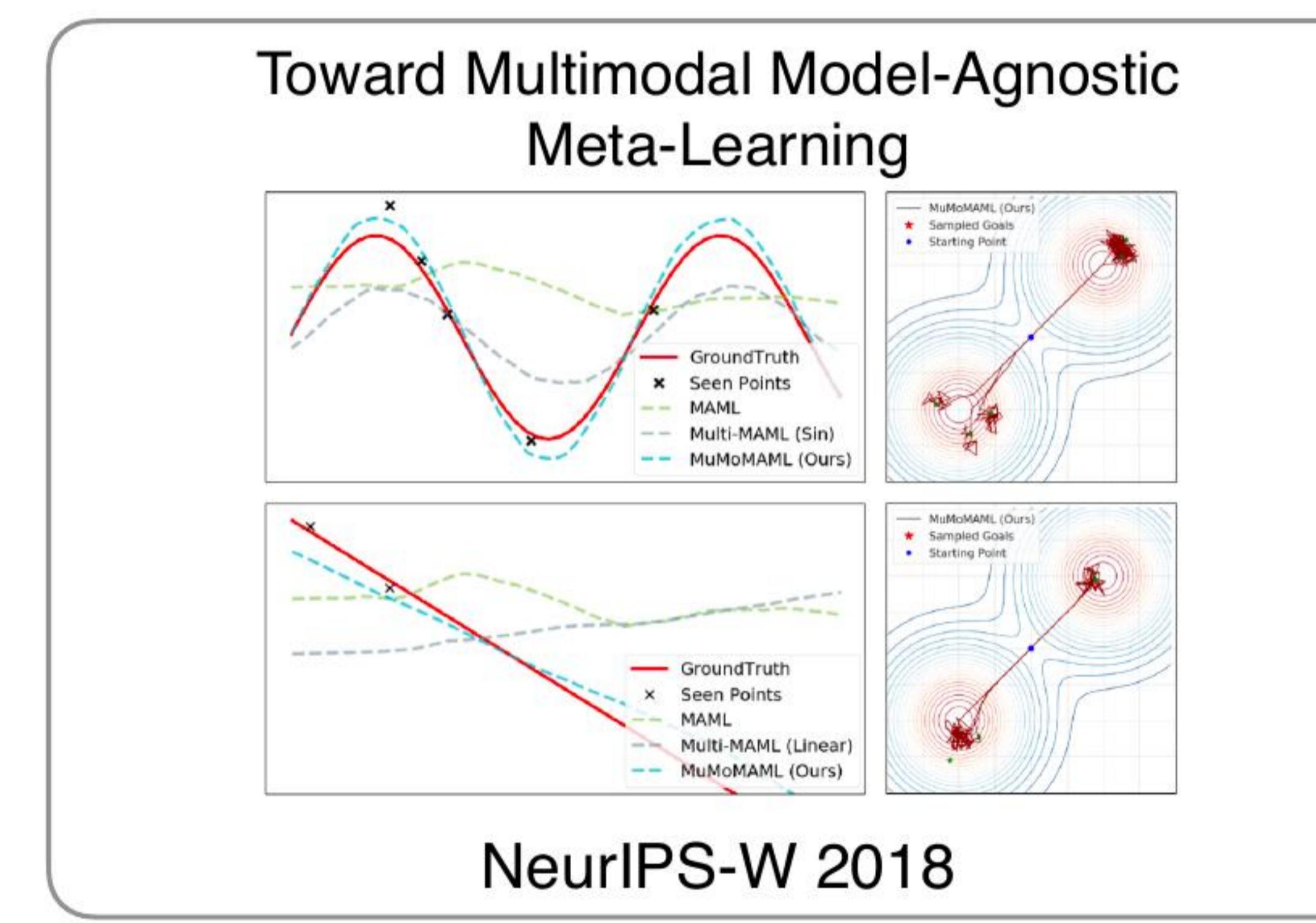


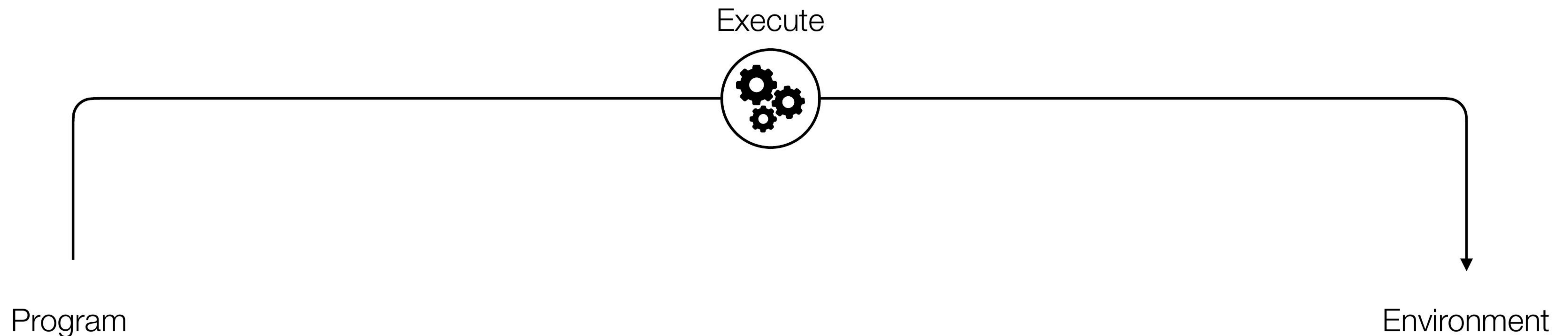
Reward

# This Talk



## Primitive Skill Acquisition



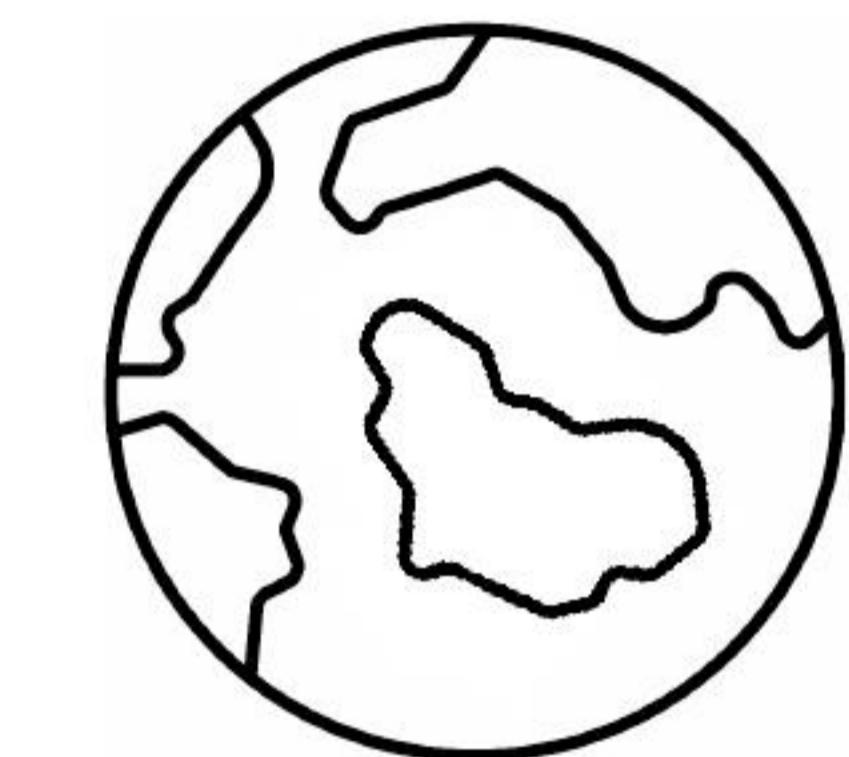


# Thank You

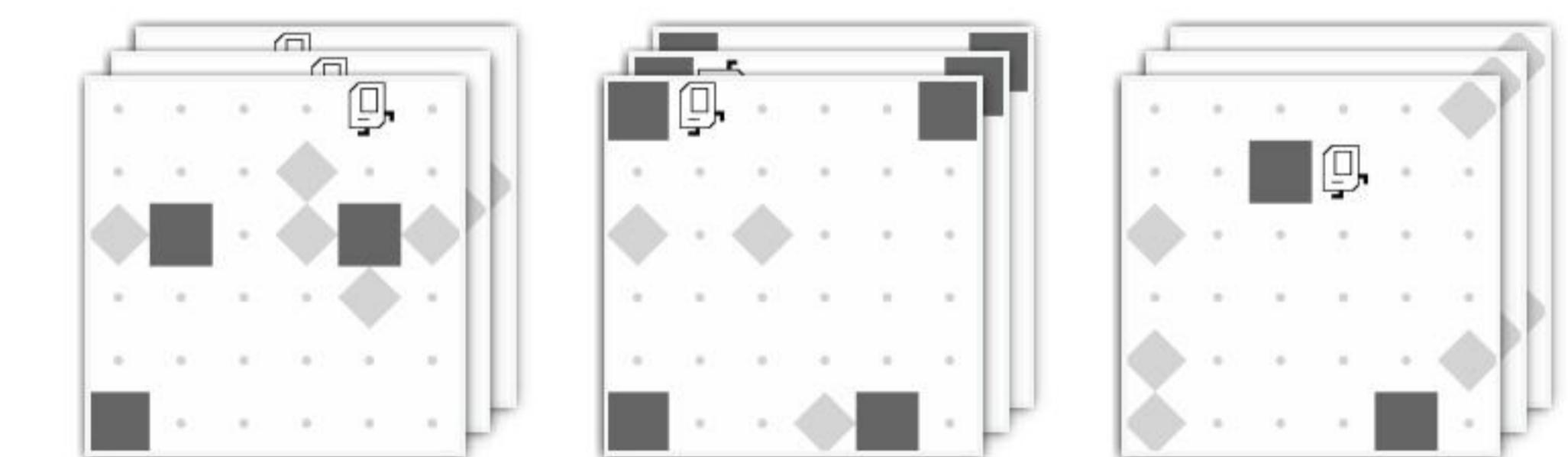


# Questions?

Environment



Demonstrations



Reward

```

DEF run() m(
  WHILE c( markerPresent c) w(
    WHILE c( markerPresent c) w(
      pickMarker
      move w)
    turnRight
    move
    turnLeft
    WHILE c( markerPresent c) w(
      pickMarker
      move w)
    turnLeft
    move
    turnRight w) m)
  
```