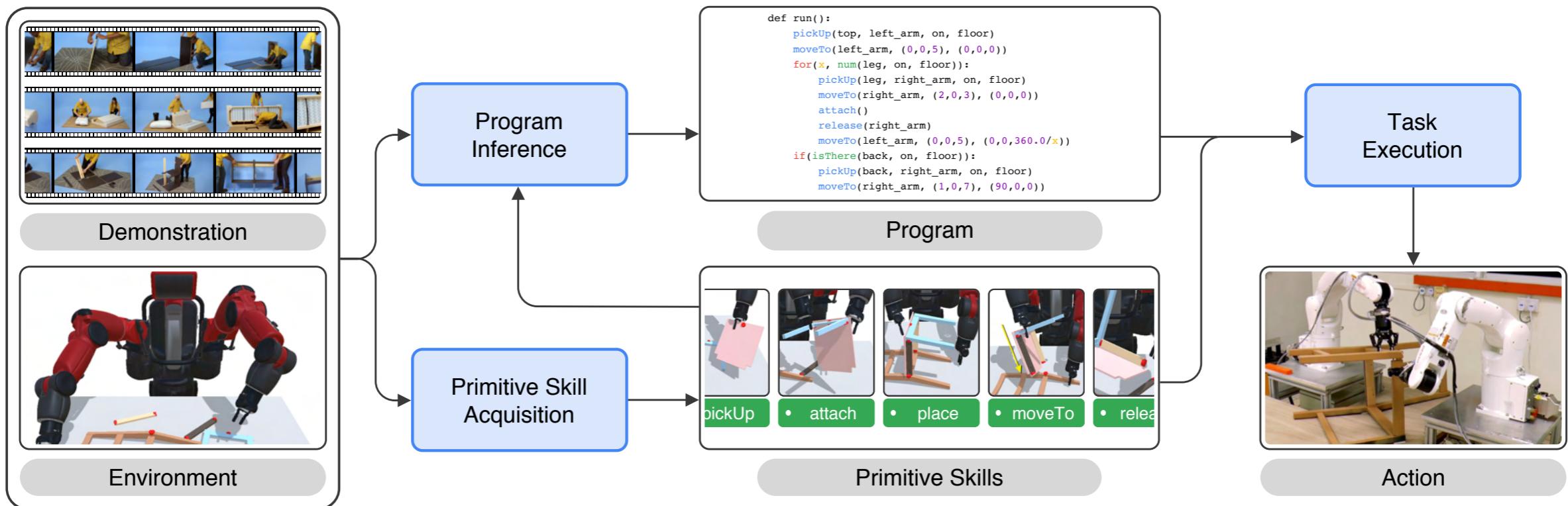


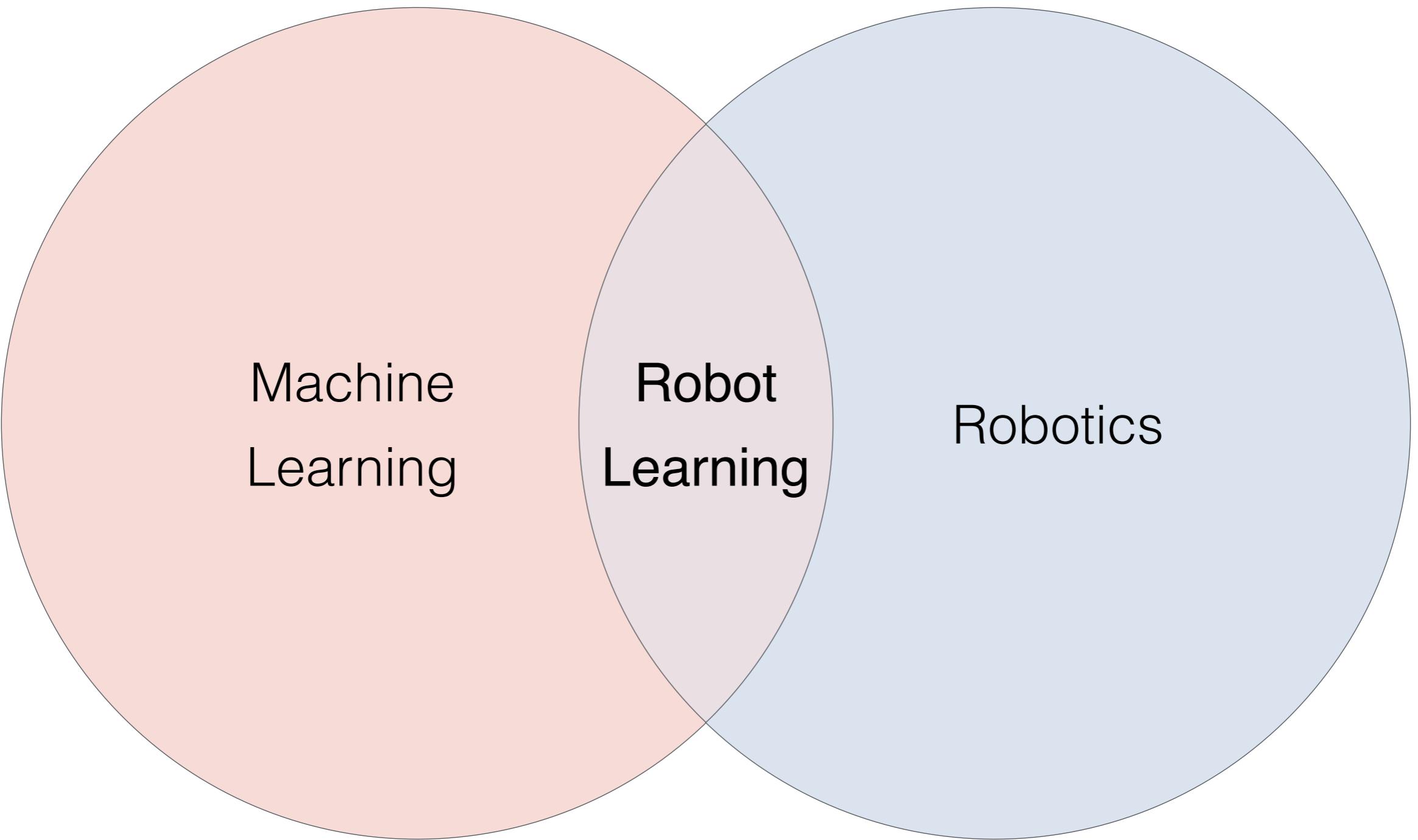
# Program-Guided Framework for Interpreting and Acquiring Complex Skills with Learning Robots



Shao-Hua Sun

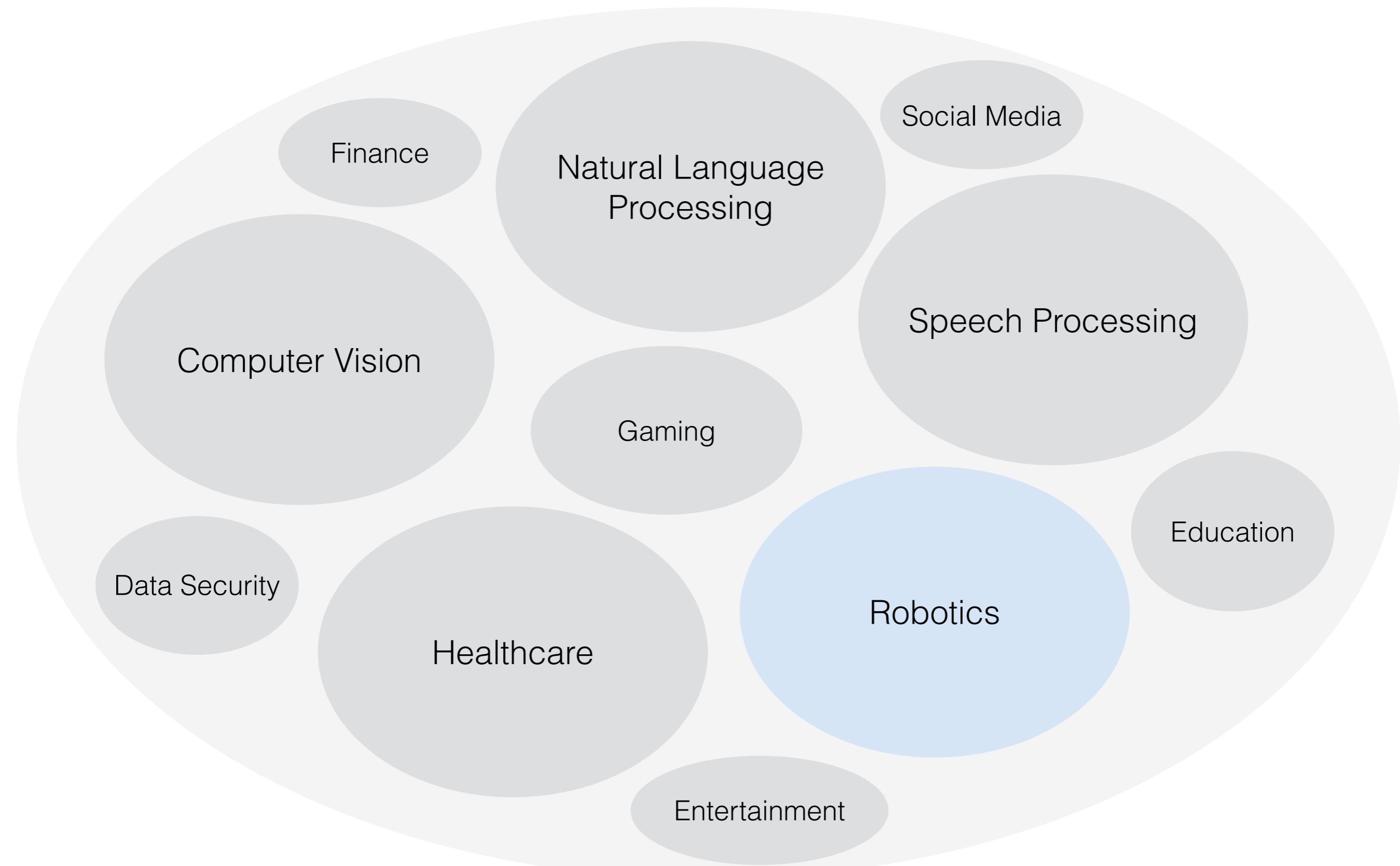
---

Ph.D. candidate in Computer Science  
at the University of Southern California (USC)



# Applications of AI

---



# Learning Problems

---

Supervised Learning

Active Learning

Meta-Learning

Unsupervised Learning

Reinforcement Learning

Meta-Reinforcement Learning

Sparse Dictionary Learning

Semi-supervised Learning

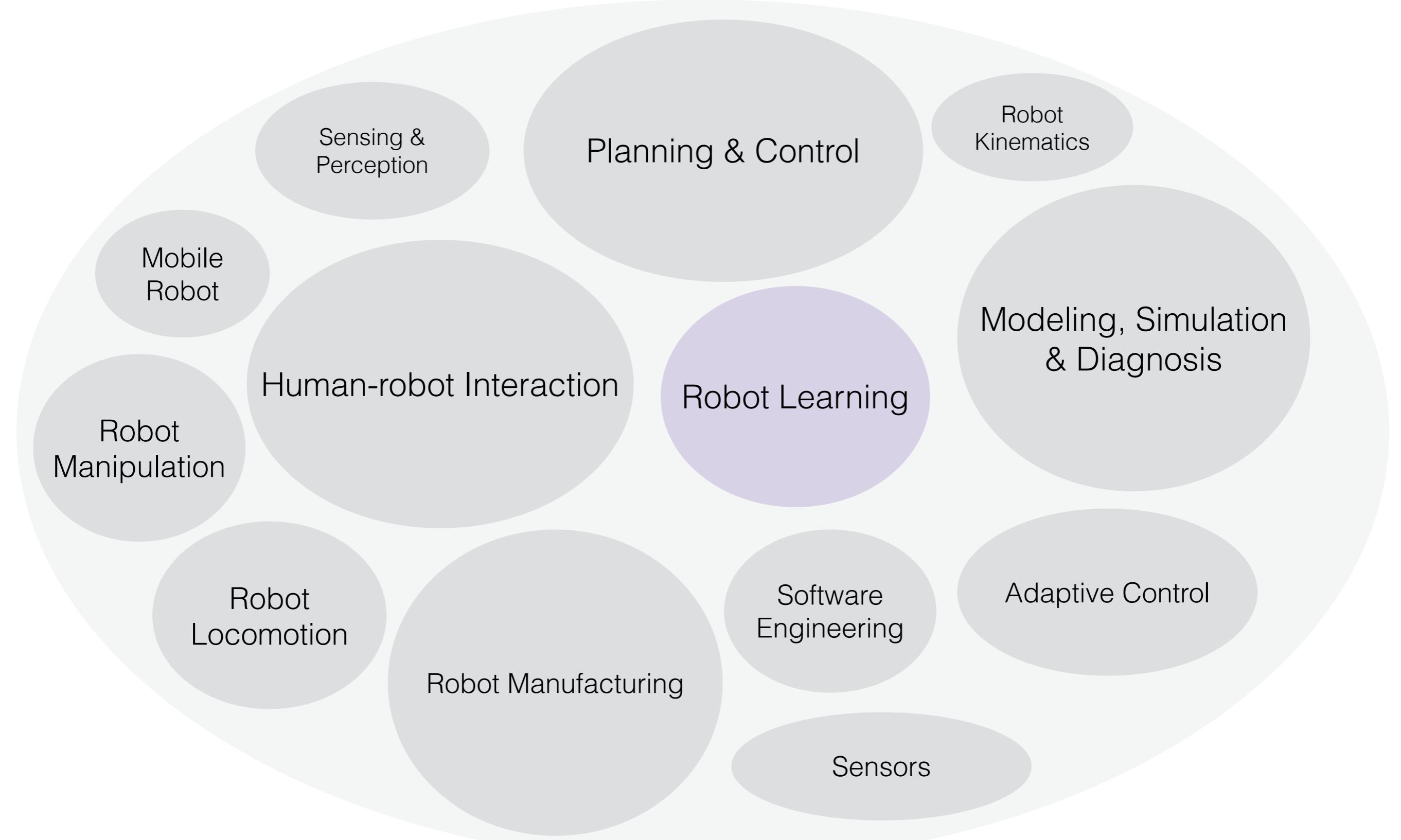
Self Learning

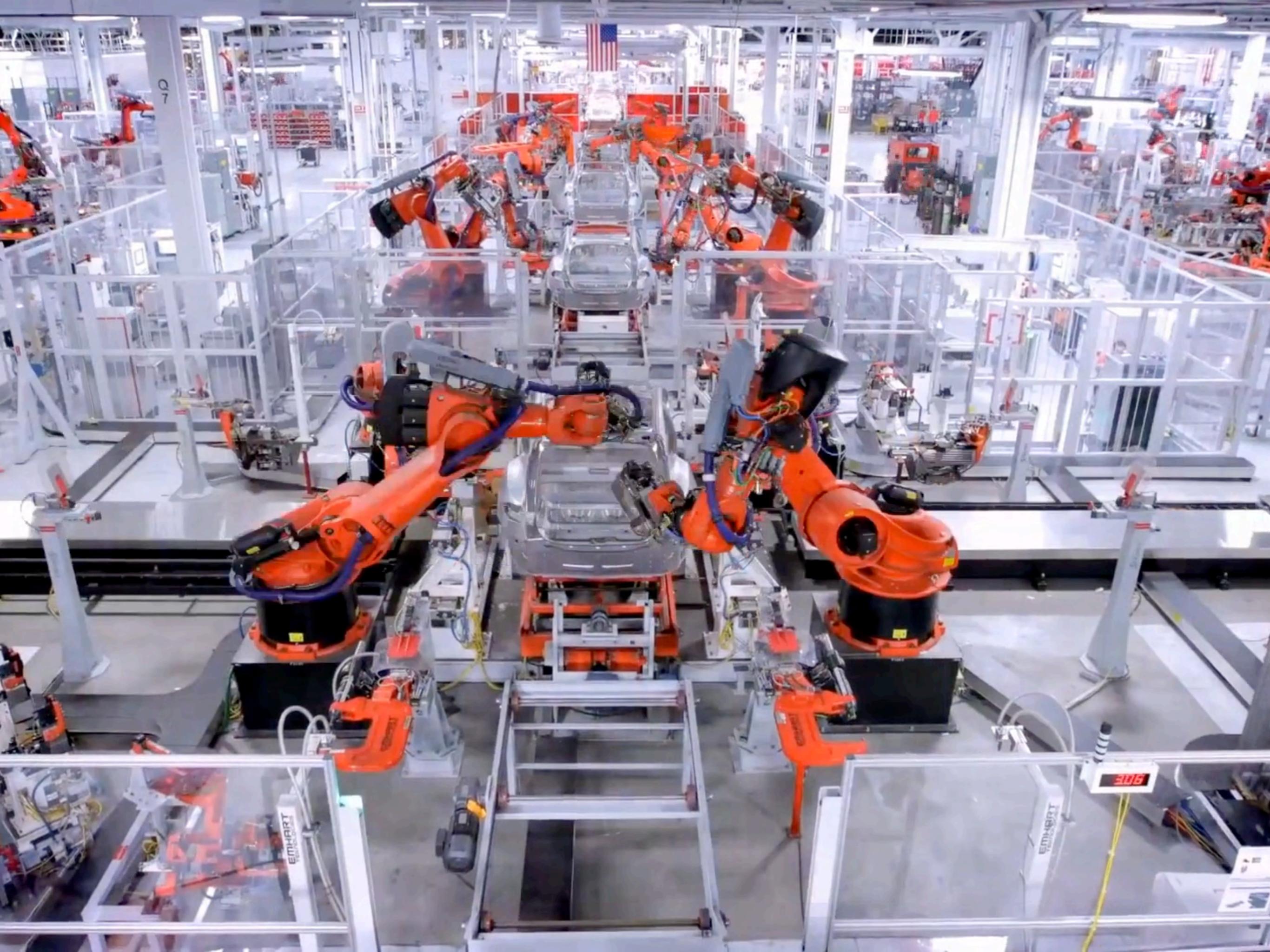
Feature Learning

Multi-Instance Learning

# Robotics

---





# Robot Learning

## Environment

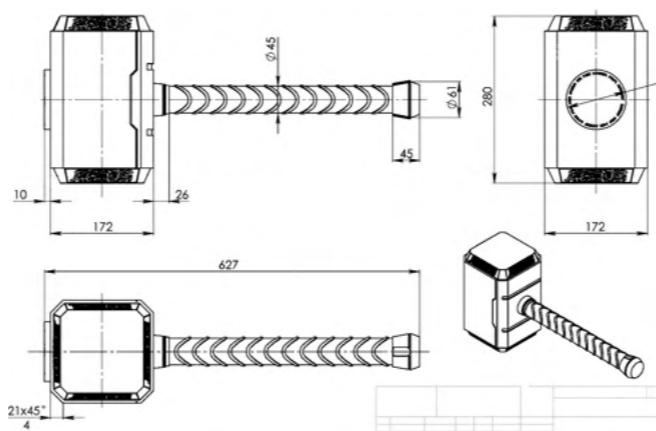


Structured



Unstructured

## Object



Known



Unseen

## Task



Pre-defined / Pre-programmed



Diverse and Novel

# Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224x224	11.2 M	1.94 B	74.8	92.2
<b>NASNet-A (5 @ 1538)</b>	<b>299x299</b>	<b>10.9 M</b>	<b>2.35 B</b>	<b>78.6</b>	<b>94.2</b>
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
<b>NASNet-A (7 @ 1920)</b>	<b>299x299</b>	<b>22.6 M</b>	<b>4.93 B</b>	<b>80.8</b>	<b>95.3</b>
ResNeXt-101 (64 x 4d) [68]	320x320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331x331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320x320	79.5 M	32.0 B	81.5	95.8
<b>SENet [25]</b>	<b>320x320</b>	<b>145.8 M</b>	<b>42.3 B</b>	<b>82.7</b>	<b>96.2</b>
<b>NASNet-A (6 @ 4032)</b>	<b>331x331</b>	<b>88.9 M</b>	<b>23.8 B</b>	<b>82.7</b>	<b>96.2</b>

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Multi-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).

## Visual Question Answering

Method	VQA v2 test-dev				VQA v2 test-mld			
	All	Yes/no	None	Other	All	Yes/no	None	Other
Prior (most common answer in training set) [3]	—	—	—	—	24.08	30.40	31.56	31.17
LSTM Language only (fixed model) [1-2]	—	—	—	—	44.26	47.40	31.55	27.37
Deep LSTM Q ansn. [17] as reported in [3]	—	—	—	—	54.26	73.40	35.18	41.83
MCB [11] as reported in [3]	—	—	—	—	62.27	78.42	38.28	53.36
UPMC-LPNS [3]	—	—	—	—	65.71	82.07	41.06	57.12
LiCoSUS	—	—	—	—	67.43	84.40	41.97	57.97
HCU-USyd-UNiC	—	—	—	—	66.77	81.40	46.29	58.30
Proposed model	—	—	—	—	68.09	84.50	45.39	59.01
ResNet features T=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 T=7, ensemble	66.34	83.38	43.17	57.05	66.73	83.71	43.77	57.39
Image features from bottom-up attention, adaptive K, ensemble	<b>69.87</b>	<b>86.08</b>	<b>45.99</b>	<b>68.01</b>	<b>70.24</b>	<b>86.64</b>	<b>46.64</b>	<b>61.55</b>

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

He et al. Mask R-CNN

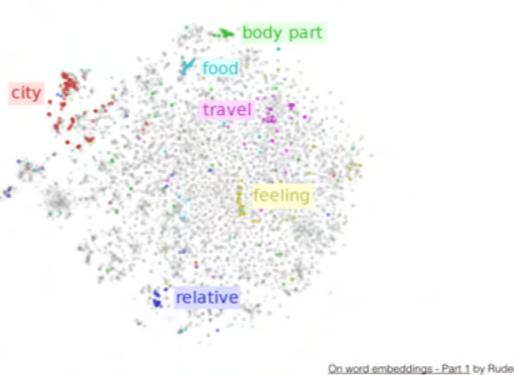
Teney et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

## Machine Translation

Source	"The reason Boeing are 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 sont en train de faire, c'est de concentrer davantage de sièges pour prendre leur avion plus compétitif 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 caser plus de sièges et rendre ses avions plus compétitifs par rapport à nos produits", a déclaré Kevin Keniston, directeur de 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 d'espionnage 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 indexPoliticsSubscribeLogInSubscribeLogInToday's PaperAdvertisementSupported ORG by F.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 Timesby Adam Goldman ORG and Michael S. SchmidtAug PERSON 13 CARDINAL 2016WASHINGTON 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 Russian GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON, a 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 assailing the Russia GPE investigation as a illegitimate "witch hunt." Mr. Strzok PERSON, who now 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 - 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.+TriviaQnA)	<b>84.2</b>	<b>91.5</b>	<b>85.1</b>	<b>91.8</b>
BERT <sub>LARGE</sub> (Ens.+TriviaQnA)	<b>86.2</b>	<b>92.2</b>	<b>87.4</b>	<b>93.2</b>

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

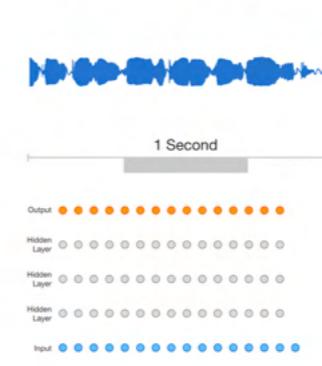
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## Speech Recognition

Sennor set	Model/combination step	WER		WER	
		devset	test	devset	test
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.2	6.3
27k-puhpum	BLSTM	11.3	8.0	9.2	5.4
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]	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 × 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	95.5
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.5
SENet [25]	320×320	145.8 M	42.3 B	82.7	95.6
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	95.2

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



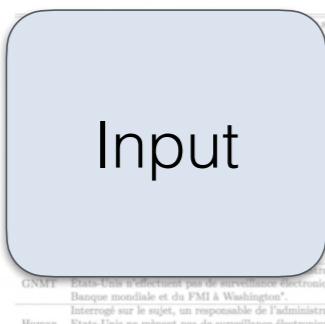
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Method	VQA v2 test dev			VQA v2 test val		
	All	Youth	Novice	All	Youth	Novice
Prior most common answer in training set [1]	-	-	-	23.08	23.26	23.17
LSTM Language only fitted model [1-2]	-	-	-	44.26	47.40	33.55
Deep LSTM Q matrix [1-2] as reported in [3]	-	-	-	54.22	73.40	35.18
MCB [4] as reported in [3]	-	-	-	42.27	78.42	38.28
UPMC-LPBN [1]	-	-	-	65.71	82.01	43.06
NASNet-A [5]	-	-	-	67.74	83.40	43.97
UNIC [6]	-	-	-	66.77	84.40	46.26
HDC-UNID-UNIC	-	-	-	68.09	84.30	45.39
Proposed model	-	-	-	68.09	84.30	50.01
ResNet features T=7, single network	62.07	79.20	39.46	52.42	42.27	79.32
Image features from bottom-up attention, adaptive $K$ , single network	65.32	81.82	44.21	56.07	65.67	82.20
ResNet features T=7, ensemble	66.47	83.59	43.45	66.47	83.59	43.29
Image features from bottom-up attention, adaptive $K$ , ensemble	69.87	86.08	48.00	76.34	90.60	61.83

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

## Machine Translation



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

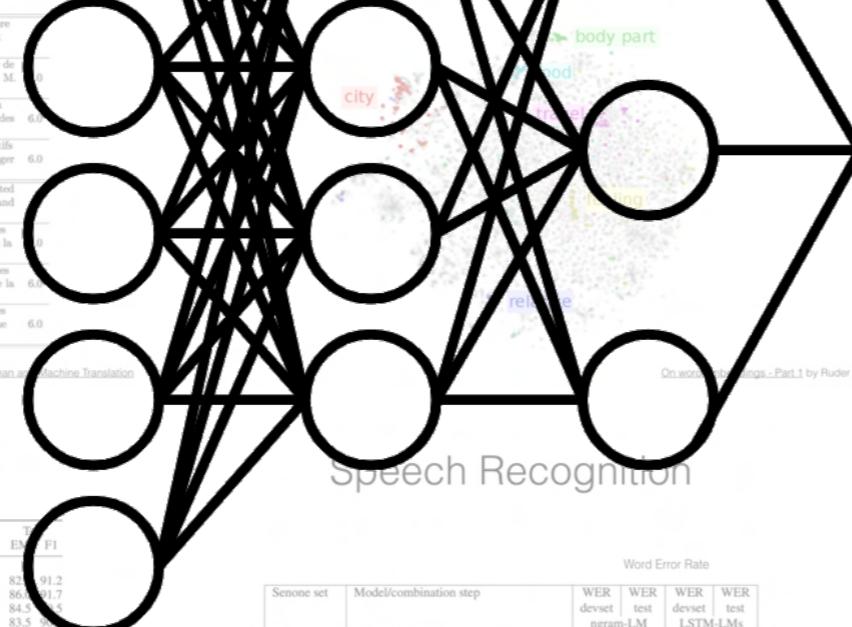
## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.1	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#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	91.5	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

## Word Embeddings



On word embeddings - Part 1 by Ruder

## Named Entity Recognition



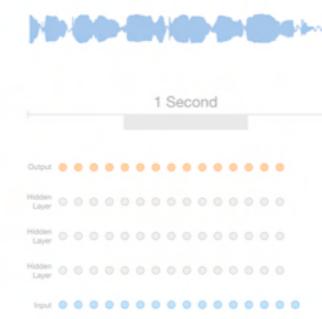
Named Entity Recognition and Classification with Scikit-Learn by Susan Li  
Esteves et al. Named Entity Recognition in Twitter using Images and Text

## Speech Recognition

Senone set	Model/combination step	Word Error Rate	
		WER devset	WER test
9k	BLSTM	11.5	8.3
27k	BLSTM	11.4	8.0
9k-puhpum	BLSTM	11.3	8.0
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4
27k	BLSTM+ResNet+LACE	9.7	7.3
	Confusion network combination + LSTM rescoring + ngram rescoring + backchannel penalty	7.4	5.2

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

# Image Classification

## 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 × 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	95.5
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.5
SENet [25]	320×320	145.8 M	42.3 B	82.7	95.6
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	95.2

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 of 299×299. Model size is [17] calculated from open-source implementation.

## Instance Segmentation



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

## Visual Question Answering

Method	VQA v2 test dev			VQA v2 test val		
	All	Youku	Youtube	All	Youku	Youtube
Prior most common answer in training set [1]	-	-	-	23.08	23.08	23.07
LSTM Language only fitted model [1-2]	-	-	-	44.26	47.40	33.55
Deep LSTM Q matrix [2] as reported in [1]	-	-	-	54.22	73.40	35.18
MCB [3] as reported in [1]	-	-	-	42.27	78.42	38.28
UPMC [4,6,9]	-	-	-	65.71	82.01	43.06
UNIC [5]	-	-	-	67.44	82.01	43.07
UNIC	-	-	-	66.77	81.80	46.26
HDC-UNIC-UNIC	-	-	-	68.09	81.90	45.39
Proposed model	-	-	-	68.09	81.90	50.01
RoCo features T×T, single network	62.07	79.20	39.46	52.42	42.27	79.32
Image features from bottom-up attention, adaptive $K$ , single network	65.32	81.82	44.21	56.01	65.67	82.20
RoCo features T×T, ensemble	66.47	83.59	43.45	66.47	83.59	37.29
Image features from bottom-up attention, adaptive $K$ , ensemble	69.87	86.08	48.00	76.34	96.60	61.83

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

Image

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

He et al. Mask R-CNN

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

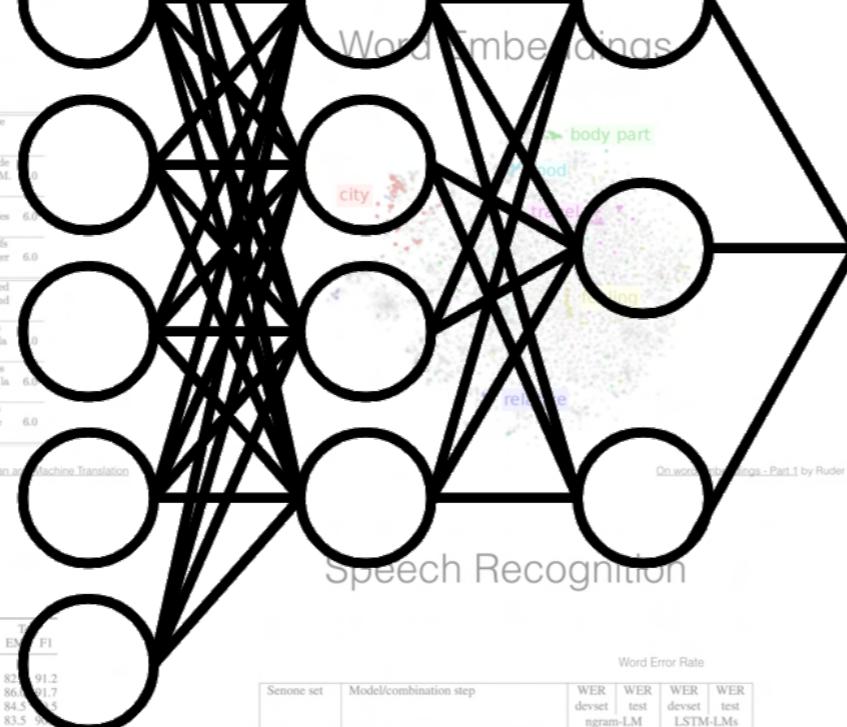


## Question Answering

System	Dev EM	Test EM	Dev F1	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	-	91.2
#1 Ensemble - nInet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nInet	-	-	83.5	90.3
#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	91.5	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.

## Word Embeddings



## Speech Recognition

Senone set	Model/combination step	WER		WER	
		devset	test	devset	test
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.2	6.3
9k-puhpum	BLSTM	11.3	8.0	9.2	6.3
9k-puhpum	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
27k	BLSTM+ResNet+LACE	9.7	7.3	7.8	5.5
-	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

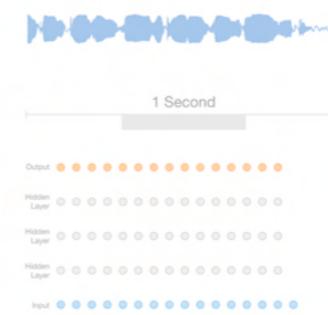
Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

## Named Entity Recognition



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

## Speech Synthesis (text-to-speech)



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

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

# Machine Translation

## 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 × 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	95.7
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.7
SENet [25]	320×320	145.8 M	42.3 B	82.7	95.8
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	95.8

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



Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

	VQA v2 test dev	VQA v2 test val
All	25.00	26.26
Youku	25.00	27.17
YouTube	25.00	25.55
Others	25.00	27.27
Proposed model	44.26	47.40
ResNet features T=7, single network	44.22	47.40
Image features from bottom-up attention, adaptive $K$ , single network	42.27	48.40
ResNet features T=7, ensemble	45.73	47.06
Image features from bottom-up attention, adaptive $K$ , ensemble	47.47	49.47
HICO-USYD-UNNC	50.77	51.80
Proposed model	68.09	69.30
ResNet features T=7, single network	62.07	79.20
Image features from bottom-up attention, adaptive $K$ , single network	65.32	81.82
ResNet features T=7, ensemble	66.43	83.59
Image features from bottom-up attention, adaptive $K$ , ensemble	69.87	86.08

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

## English sentence

France is never cold  
in September

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

## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)	-	-	-	-
Human	-	-	82.0	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#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

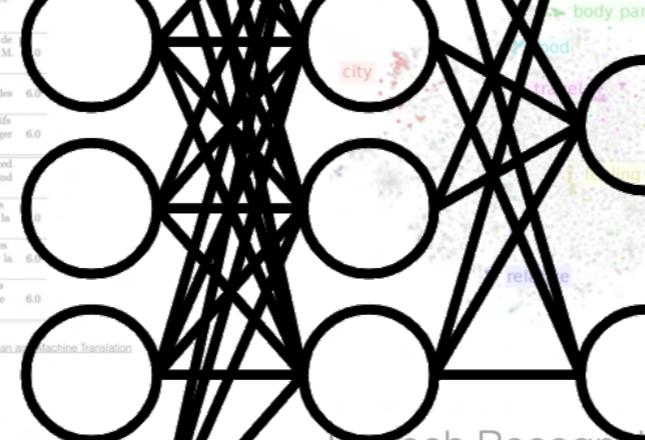
BERT <sub>BASE</sub> (Single)	80.8	88.5	-	-
BERT <sub>LARGE</sub> (Single)	84.1	90.9	-	-
BERT <sub>LARGE</sub> (Ensemble)	85.8	91.8	-	-
BERT <sub>LARGE</sub> (Sgl.+TriviaQnA)	84.2	91.5	85.1	91.8
BERT <sub>LARGE</sub> (Ens.+TriviaQnA)	86.2	92.2	87.4	93.2

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

## Word Embeddings

### city



On word embeddings - Part 1 by Ruder

## French sentence

la france est jamais  
froid en septembre

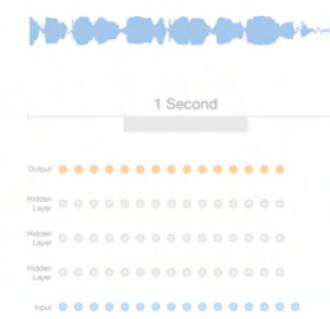
Named Entity Recognition and Classification with Scikit-Learn by Susan Li  
Esteves et al. Named Entity Recognition in Twitter using Images and Text

## Speech Recognition

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

Word Error Rate

## Speech Synthesis (text-to-speech)



Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

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

# Automatic Speech Recognition

## 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 × 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	95.5
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.5
SENet [25]	320×320	145.8 M	42.3 B	82.7	95.6
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	95.2

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



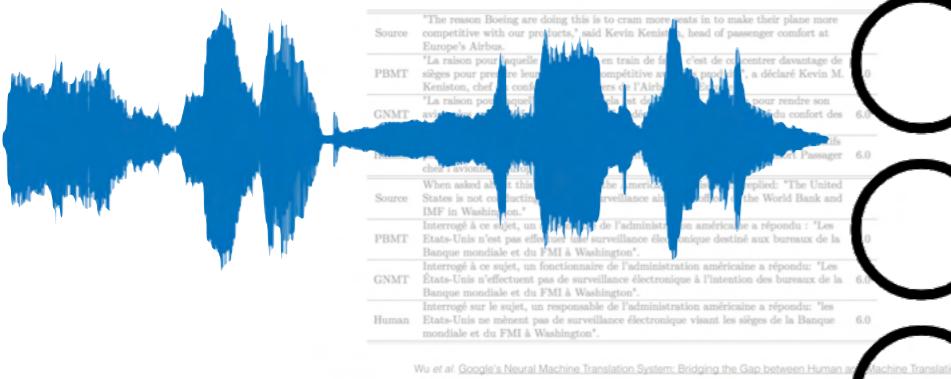
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

	VQA v2 test dev	VQA v2 test val
All	25.08	26.26
Yes/no	31.55	31.57
Multiple choice	44.26	47.40
LSTM Language only (fixed model) [1]	-	-
Deep LSTM Q ansn. [27] as reported in [3]	-	-
MCB [10] as reported in [3]	-	-
UPMC-LPBN [1]	-	-
Proposed model	65.71	62.01
UCM [33]	-	-
HDC-UNSD-UNNC	-	-
Proposed model	68.09	64.30
RoCoN features T×T, single network	62.07	79.20
Image features from bottom-up attention, adaptive K, single network	65.32	81.82
RoCoN features T×T, ensemble	66.43	81.11
Image features from bottom-up attention, adaptive K, ensemble	69.87	86.08

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

## Waveform Translation



## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)	-	-	-	-
Human	-	-	82.4	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#2 Single - QANet	-	-	82.5	89.3

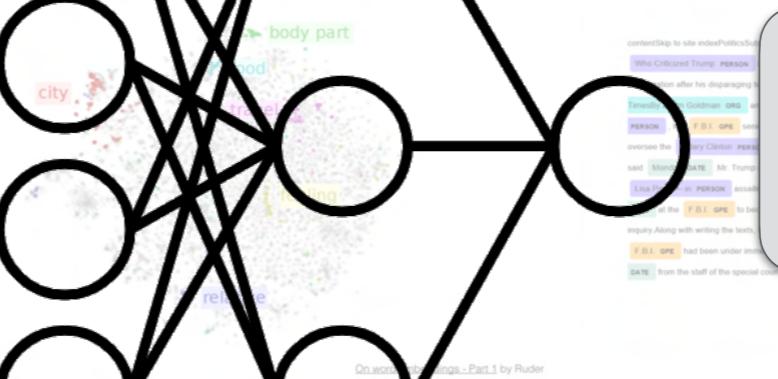
Published

BIDAF+ELMo (Single)	85.8	-
R.M. Reader (Single)	78.9	86.3
R.M. Reader (Ensemble)	81.2	87.9
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.+TriviaQnA)	84.2	91.5
BERT <sub>LARGE</sub> (Ens.+TriviaQnA)	86.2	92.3
BERT <sub>LARGE</sub> (Ens.+TriviaQnA)	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

## Word Embeddings



## Text

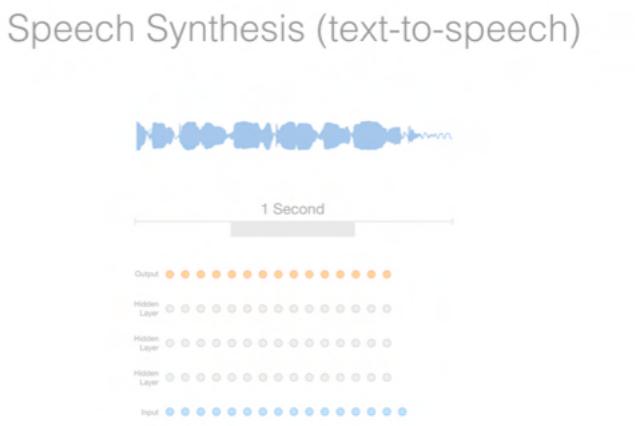
This is a supervised learning method

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

## 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	BLSTM	11.4	8.0	9.3	6.3
9k-puhpum	BLSTM	11.3	8.0	9.2	6.3
9k-puhpum	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
27k	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
	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	

## Word Error Rate



Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

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	11.2 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 × 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	95.5
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.5
SENet [25]	320×320	145.8 M	42.3 B	82.7	95.6
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	95.2

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



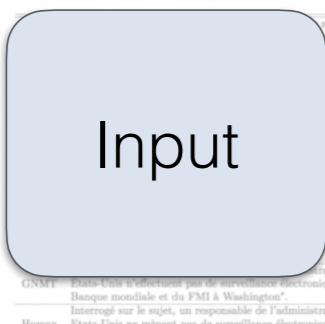
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Method	VQA v2 test dev			VQA v2 test val		
	All	Youth	Novice	All	Youth	Novice
Prior most common answer in training set [1]	-	-	-	23.08	23.26	23.17
LSTM Language only fitted model [1-2]	-	-	-	44.26	47.40	33.55
Deep LSTM Q matrix [1-2] as reported in [3]	-	-	-	54.22	73.40	35.18
MCB [4] as reported in [3]	-	-	-	42.27	78.42	38.28
UPMC-LPBN [1]	-	-	-	65.71	82.01	43.06
FCOS [5]	-	-	-	67.74	81.20	45.97
UNet [6]	-	-	-	56.77	81.40	46.26
HDC-UNet-UNet [7]	-	-	-	68.09	81.30	45.39
Proposed model	-	-	-	68.09	81.30	50.01
ResNet features T=7, single network	62.07	79.20	39.46	52.42	42.27	79.32
Image features from bottom-up attention, adaptive $K$ , single network	65.32	81.82	44.21	56.07	65.87	82.20
ResNet features T=7, ensemble	66.47	83.59	43.45	66.47	83.59	37.29
Image features from bottom-up attention, adaptive $K$ , ensemble	69.87	86.08	48.00	76.34	86.64	61.83

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

## Machine Translation



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

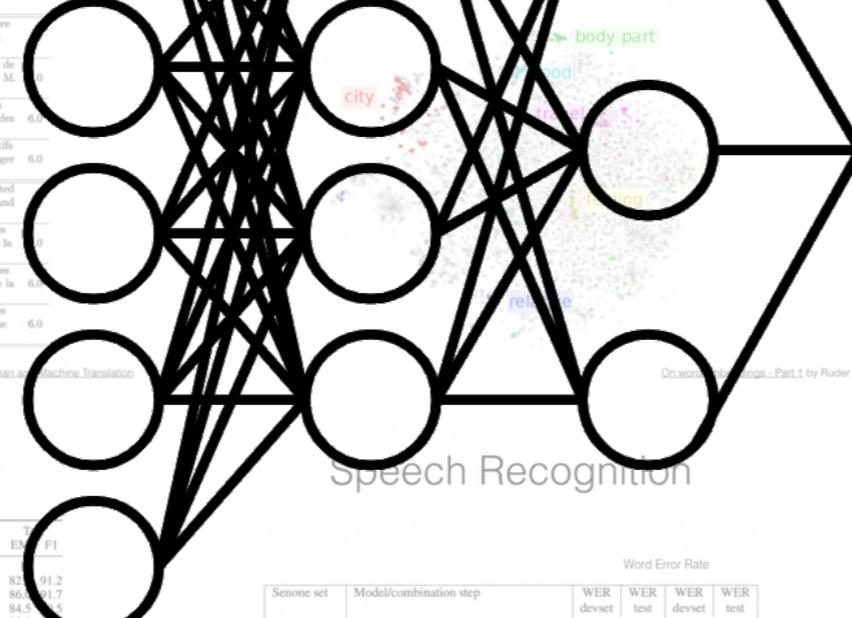
## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.1	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#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	91.5	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

## Word Embeddings



## Named Entity Recognition



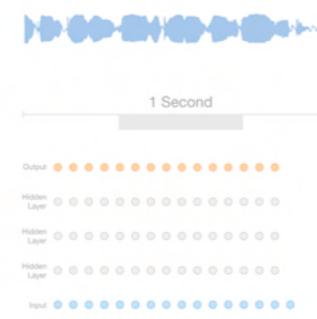
Named Entity Recognition and Classification with Scikit-Learn by Susan Li Esteves et al. Named Entity Recognition in Twitter using Images and Text

## Speech Recognition

Senone set	Model/combination step	Word Error Rate	
		WER devset	WER test
9k	BLSTM	11.5	8.3
27k	BLSTM	11.4	8.0
9k-puhpum	BLSTM	11.3	8.0
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4
27k	BLSTM+ResNet+LACE	9.7	7.3
	Confusion network combination	-	7.4
	+ LSTM rescoring	-	7.3
	+ ngram rescoring	-	7.2
	+ backchannel penalty	-	5.8

WER devset	WER test
ngram-LM	6.3
BLSTM	6.3
BLSTM	6.3
BLSTM	5.4
BLSTM	5.4
BLSTM	5.5
BLSTM	5.8
BLSTM	5.8
BLSTM	5.2
BLSTM	5.2
BLSTM	5.1

## 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 V3 [70]	299×299	11.5M	1.64B	74.8	92.2
NASNet-A (5@ 15.58)	299×299	8.9M	2.35 B	78.6	94.2
Inception V3 [80]	299×299	10.5M	5.72 B	78.8	94.4
Xception [91]	299×299	10.5M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	9M	13.2 B	80.1	95.1
NASNet-A (7@ 19.20)	299×299	10.5M	4.93 B	80.8	95.3
MobileNet V2 [69]	320×320	8.3M	0.72 B	80.9	95.6
PolyNet [69]	331×331	92.5M	34.7 B	81.3	95.5
DPN-131 [8]	320×320	79.5M	32.0 B	81.5	95.5
SENet [25]	320×320	145.8M	42.3 B	82.7	95.6
NASNet-A (6@ 40.32)	331×331	88.9M	23.8 B	82.7	95.2

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of compute operations required to process one input image. Note that some models contain multiple accurate operations are calculated for the image size reported in the table. Model size for 144F is calculated from open-source implementation.

## Instance Segmentation



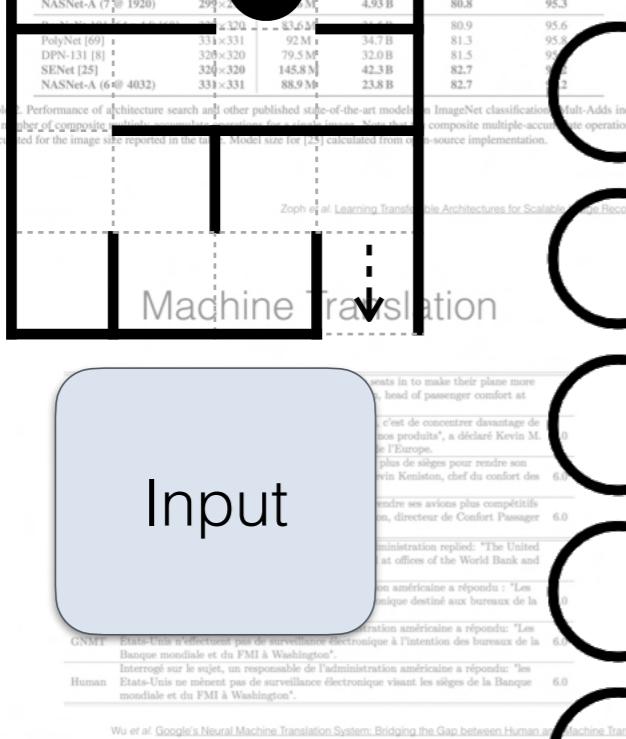
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

	VQA v2 test dev	VQA v2 test val
All	25.00	26.26
Youth	25.00	27.17
Novice	-	-
Other	-	-
Proposed model	-	-
Prior most common answer in training set [1]	-	-
LSTM Language only (fixed model) [1-2]	-	-
Deep LSTM Q ansin [1] as reported in [1]	-	-
MCR [10] as reported in [1]	-	-
UPMC-LPBN [1]	-	-
65.73 82.40 43.06 37.22	65.73 82.40 43.06 37.22	65.73 82.40 43.06 37.22
67.47 84.39 46.26 36.30	67.47 84.39 46.26 36.30	67.47 84.39 46.26 36.30
ESONIC	-	-
HDC-USyd-UNNC	-	-
Proposed model	-	-
ResNet features T=7, single network	62.07	79.20 39.46 52.42
Image features from bottom-up attention, adaptive $K$ , single network	65.32	81.82 44.21 56.01
ResNet features T=7, ensemble	66.47	83.59 43.49 57.45
Image features from bottom-up attention, adaptive $K$ , ensemble	69.87	86.08 48.99 60.80
76.24 96.60 64.64 61.83	76.24 96.60 64.64 61.83	76.24 96.60 64.64 61.83

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

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



## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.4	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#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	91.5	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.

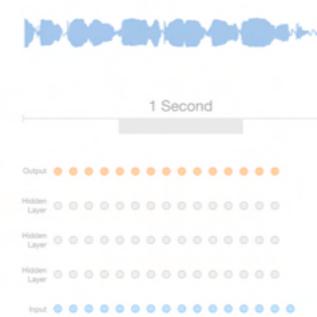
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## Speech Recognition

Senone set	Model/combination step	Word Error Rate			
		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	6.3
9k-puhpum	BLSTM	11.3	8.0	9.2	6.3
9k-puhpum	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
27k	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
	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 V3 [70]	299×299	11.5M	1.64B	74.8	92.2
NASNet-A (5@ 15.58)	299×299	8.9M	2.35 B	78.6	94.2
Inception V3 [80]	299×299	10.5M	5.72 B	78.8	94.4
Xception [91]	299×299	10.5M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	9M	13.2 B	80.1	95.1
NASNet-A (7@ 19.20)	299×299	10.5M	4.93 B	80.8	95.3
MobileNet V2 [69]	320×320	8.3M	0.70 B	80.9	95.6
PolyNet [69]	331×331	92.5M	34.7 B	81.3	95.5
DPN-131 [8]	320×320	79.5M	32.0 B	81.5	95.5
SENet [25]	320×320	145.8M	42.3 B	82.7	95.6
NASNet-A (6@ 40.32)	331×331	88.9M	23.8 B	82.7	95.2

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of compute operations required to process one input image. Note that some models contain multiple accurate operations are calculated for the image size reported in the table. Model size for 144F is calculated from open-source implementation.

## Instance Segmentation



Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

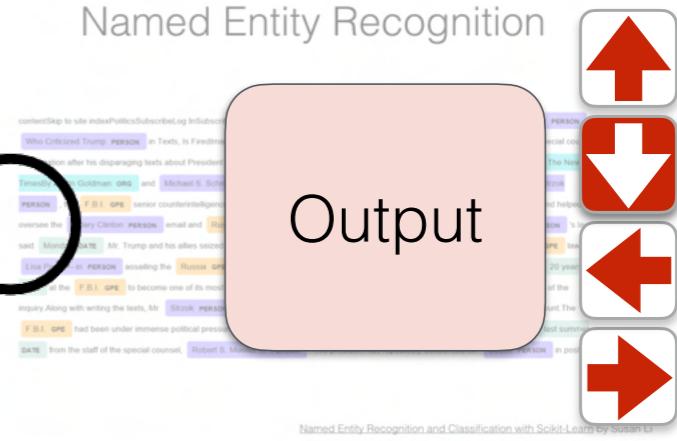
	VQA v2 test dev	VQA v2 test val
All	25.00	26.26
Youth	25.00	27.17
Novice	-	-
Other	-	-
Proposed model	-	-
Prior most common answer in training set [1]	-	-
LSTM Language only (fixed model) [1-2]	-	-
Deep LSTM Q ansin [1] as reported in [1]	-	-
MCR [10] as reported in [1]	-	-
UPMC-LPBN [1]	-	-
65.73 82.0 43.06 37.22	-	-
67.45 83.2 43.06 37.27	-	-
68.77 81.86 45.26 38.36	-	-
68.09 84.39 45.39 39.01	-	-
HDC-UVSD-UNNC	-	-
Proposed model	-	-
ResNet features T=7, single network	62.07	79.20 39.46 52.42
Image features from bottom-up attention, adaptive $K$ , single network	65.32	81.82 44.21 36.01
ResNet features T=7, ensemble	66.42	83.59 43.49 37.45
Image features from bottom-up attention, adaptive $K$ , ensemble	69.87	86.08 48.99 48.80
76.24 96.60 64.64 61.83	-	-

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

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



## Named Entity Recognition



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

## Input

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

GNMT - 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".

## Question Answering

System	Dev EM	Test EM	Dev F1	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	-	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#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	91.5	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.

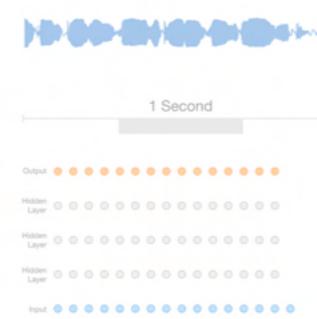
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## Speech Recognition

Senone set	Model/combination step	Word Error Rate	
		WER devset	WER test
9k	BLSTM	11.5	8.3
27k	BLSTM	11.4	8.0
27k-puhpum	BLSTM	11.3	8.0
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3
27k	BLSTM+ResNet+LACE	10.0	7.5
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 V3 [70]	299×299	11.5M	1.61B	74.8	92.2
NASNet-A (5 $\oplus$ 15.58)	299×299	10.9 M $\oplus$	2.35 B	78.6	94.2
Inception V3 [60]	299×299	23.8 M $\oplus$	5.72 B	78.8	94.4
Xception [91]	299×299	22.8 M $\oplus$	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M $\oplus$	13.2 B	80.1	95.1
NASNet-A (7 $\oplus$ 1920)	299×299	22.6 M $\oplus$	4.93 B	80.8	95.3
<b>Proposed model</b>	331×331	83.6 M $\oplus$	2.07 B	80.9	95.6
PolyNet [69]	331×331	3.4 M $\oplus$	34.7 B	81.3	95.5
DPN-131 [8]	320×320	3.4 M $\oplus$	32.0 B	81.5	95.6
SENet [25]	320×320	3.4 M $\oplus$	42.3 B	82.7	95.6
NASNet-A (6 $\oplus$ 4032)	331×331	3.4 M $\oplus$	23.8 B	82.7	95.2

Table 2. Performance of architecture search and other proposed state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiplicative operations. Note that some of the reported composite multiple-accuracy operations are calculated for the image size reported in the paper. Model size for 1.45 $\oplus$  is calculated from open-source implementation.

## Instance Segmentation



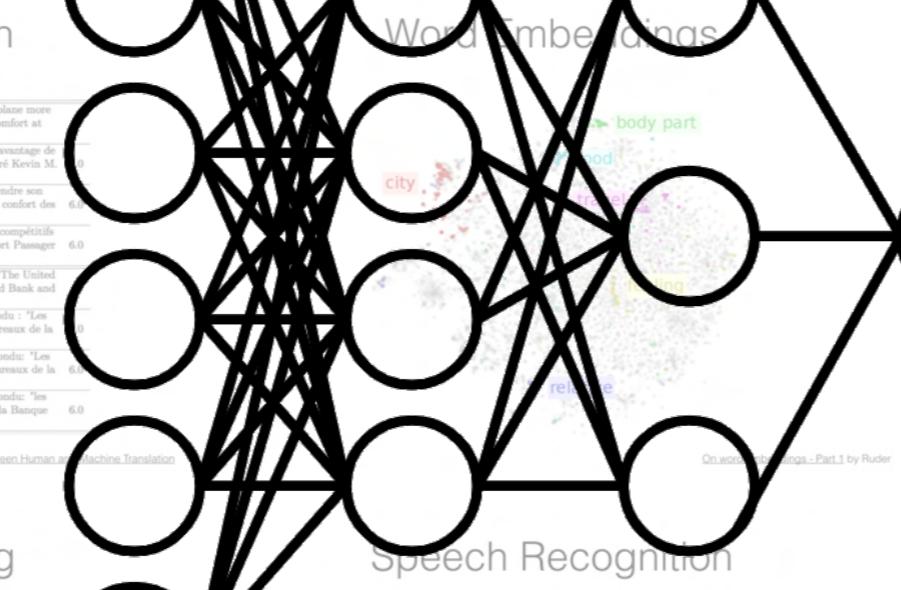
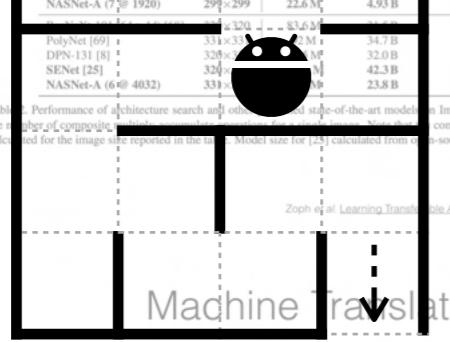
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

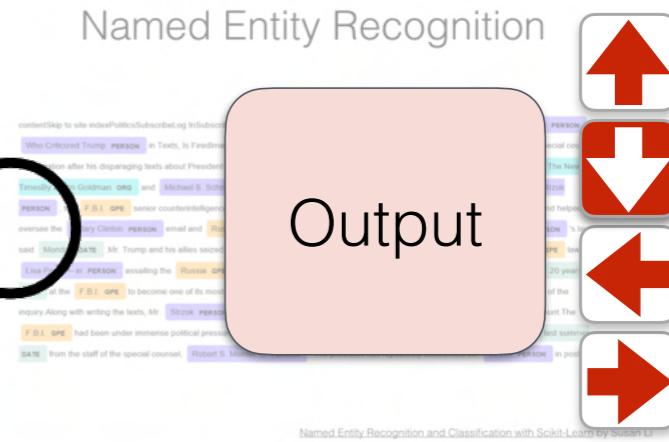
	VQA v2 test dev	VQA v2 test val
All	25.00	26.26
Youth	25.00	27.17
Novice	24.26	27.37
Other	24.26	27.37
Proposed model	44.22	35.18
DRNN-LSTM Q net [1] as reported in [1]	54.22	41.83
DRNN-LSTM Q net [1] as reported in [1]	54.22	41.83
MCR [10] as reported in [1]	42.27	38.28
UPMC-LPBN [1]	45.71	41.06
UNIC [1]	47.77	37.97
HCU-USyd-UNIC	50.77	48.26
Proposed model	68.09	45.39
ResNet features T $\times$ T, single network	62.07	79.20
Image features from bottom-up attention, adaptive K, single network	65.32	81.82
ResNet features T $\times$ T, ensemble	66.47	83.58
Image features from bottom-up attention, adaptive K, ensemble	69.87	86.08

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

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



## Named Entity Recognition



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

Input

Output

## Question Answering

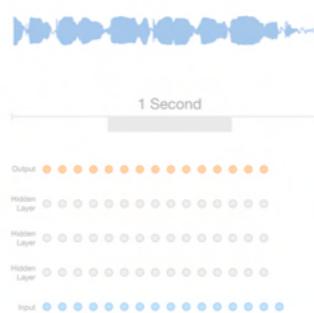
System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)	-	-	-	-
Human	-	-	82.4	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#2 Single - QANet	-	-	82.5	89.3
Published	-	-	-	-
BIDAF+ELMo (Single)	-	-	85.8	-
R.M. Reader (Single)	78.9	86.3	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	91.5	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

Senone set	Model/combination step	Word Error Rate	
		WER devset	WER test
9k	BLSTM	11.5	8.3
27k	BLSTM	11.4	8.0
27k-puhpum	BLSTM	11.3	8.0
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3
27k	BLSTM+ResNet+LACE	10.0	7.5
	Confusion network combination		7.4
	+ LSTM rescoring		7.3
	+ ngram rescoring		7.2
	+ backchannel penalty		7.2
			5.1

Xiong et al. The Microsoft 2017 Conversational Speech Recognition System



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

## Speech Recognition

## Speech Synthesis (text-to-speech)

# Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V3 [70]	299×299	11.5M	1.61B	74.8	92.2
NASNet-A (5 $\times$ 15.58)	299×299	10.9 M $\downarrow$	2.35 B	78.6	94.2
Inception V3 [60]	299×299	23.8 M $\downarrow$	5.72 B	78.8	94.4
Xception [91]	299×299	22.8 M $\downarrow$	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M $\downarrow$	13.2 B	80.1	95.1
NASNet-A (7 $\times$ 1920)	299×299	22.6 M $\downarrow$	4.93 B	80.8	95.3
<b>Proposed model</b>	331×331	83.6 M $\downarrow$	2.07 B	80.9	95.6
PolyNet [69]	331×331	83.6 M $\downarrow$	34.7 B	81.3	95.5
DPN-131 [8]	320×320	32.0 M $\downarrow$	32.0 B	81.5	95.6
SENet [25]	320×320	32.0 M $\downarrow$	42.3 B	82.7	95.6
NASNet-A (6 $\times$ 4032)	331×331	23.8 M $\downarrow$	23.8 B	82.7	95.6

Table 2. Performance of architecture search and other state-of-the-art models in ImageNet classification. Mult-Adds indicate the number of compute operations required to process one input image. Note that some composite mult-adds are operations are calculated for the image size reported in the table. Model size for 1.45 $\times$  calculated from open-source implementation.

## Instance Segmentation



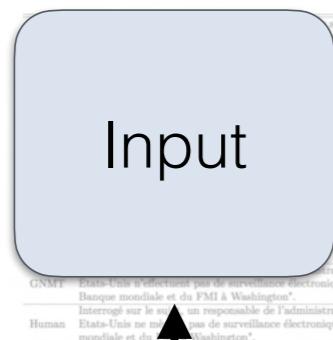
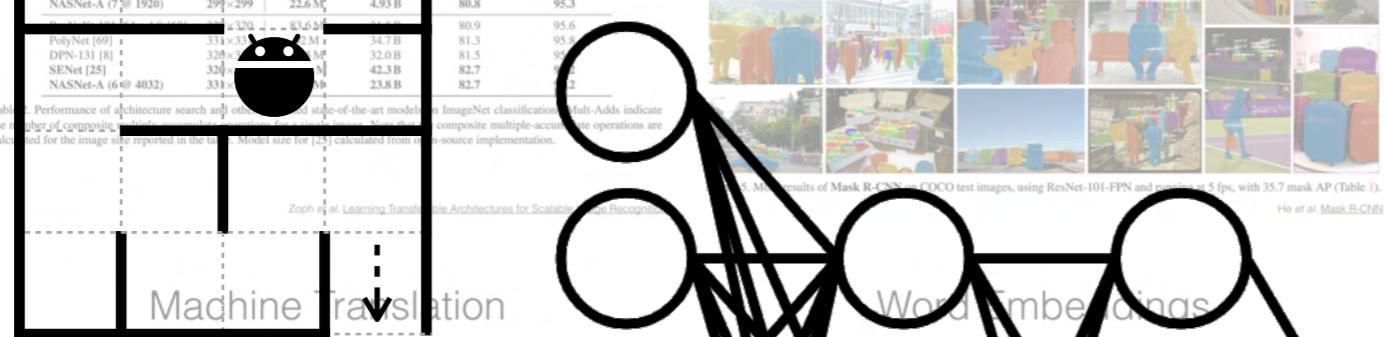
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Method	VQA v2 test dev		
	All	Youth	Novice
Prior most common answer in training set [1]	25.00	26.36	27.17
LSTM Language only fitted model [1-2]	44.26	47.40	51.55
Deep LSTM Q matrix [1] as reported in [1]	54.22	73.40	35.18
MCR [1] as reported in [1]	42.27	78.62	38.28
UPMC-LPBN [1]	65.71	82.07	43.06
UNIC [1]	67.77	81.40	46.26
HCU-USyd-UNNC	66.77	81.40	46.26
<b>Proposed model</b>	68.09	84.30	45.39

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

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



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

## Question Answering

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)	-	-	-	-
Human	-	-	82.9	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#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)	85.8	91.8	-	-
BERT <sub>BASE</sub> (Ensemble)	85.8	91.8	-	-
BERT <sub>BASE</sub> (Sgl+TriviaQA)	84.2	91.5	85.1	91.8
BERT <sub>BASE</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

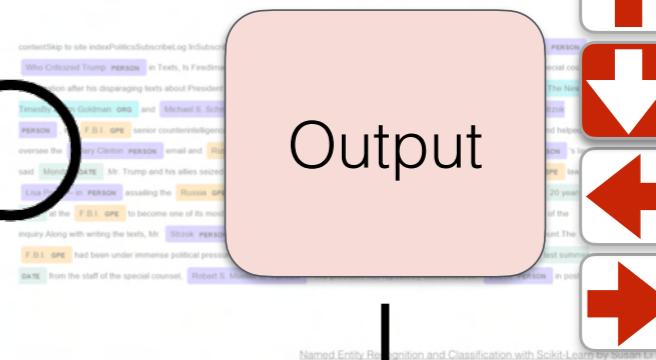
Senone set	Model/combination step	Word Error Rate	
		WER devset ngram-LM	WER test LSTM-LMs
9k	BLSTM	11.5	8.3
27k	BLSTM	11.4	8.0
9k	BLSTM	11.3	8.0
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.8
27k	BLSTM+ResNet+LACE	10.0	7.5
27k	BLSTM+ResNet+LACE+CNN-BLSTM	7.5	5.7

+ 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

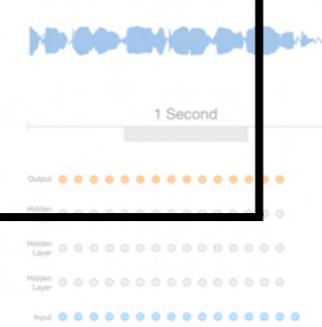
## Speech Recognition

## Named Entity Recognition



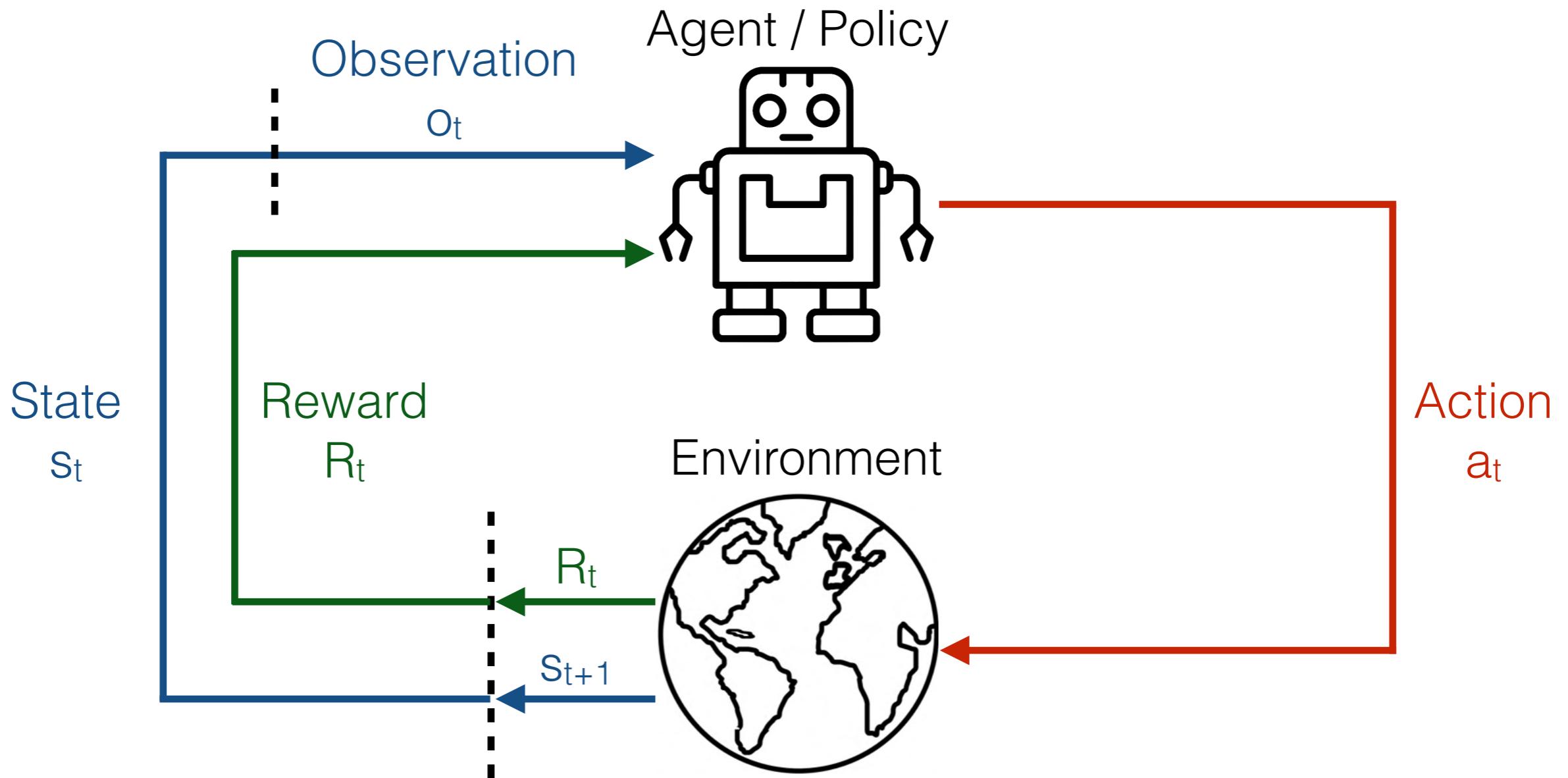
Named Entity Recognition and Classification with Scikit-Learn by Susan Li Esteves et al. Named Entity Recognition in Twitter using Images and Text

## Speech Synthesis (text-to-speech)



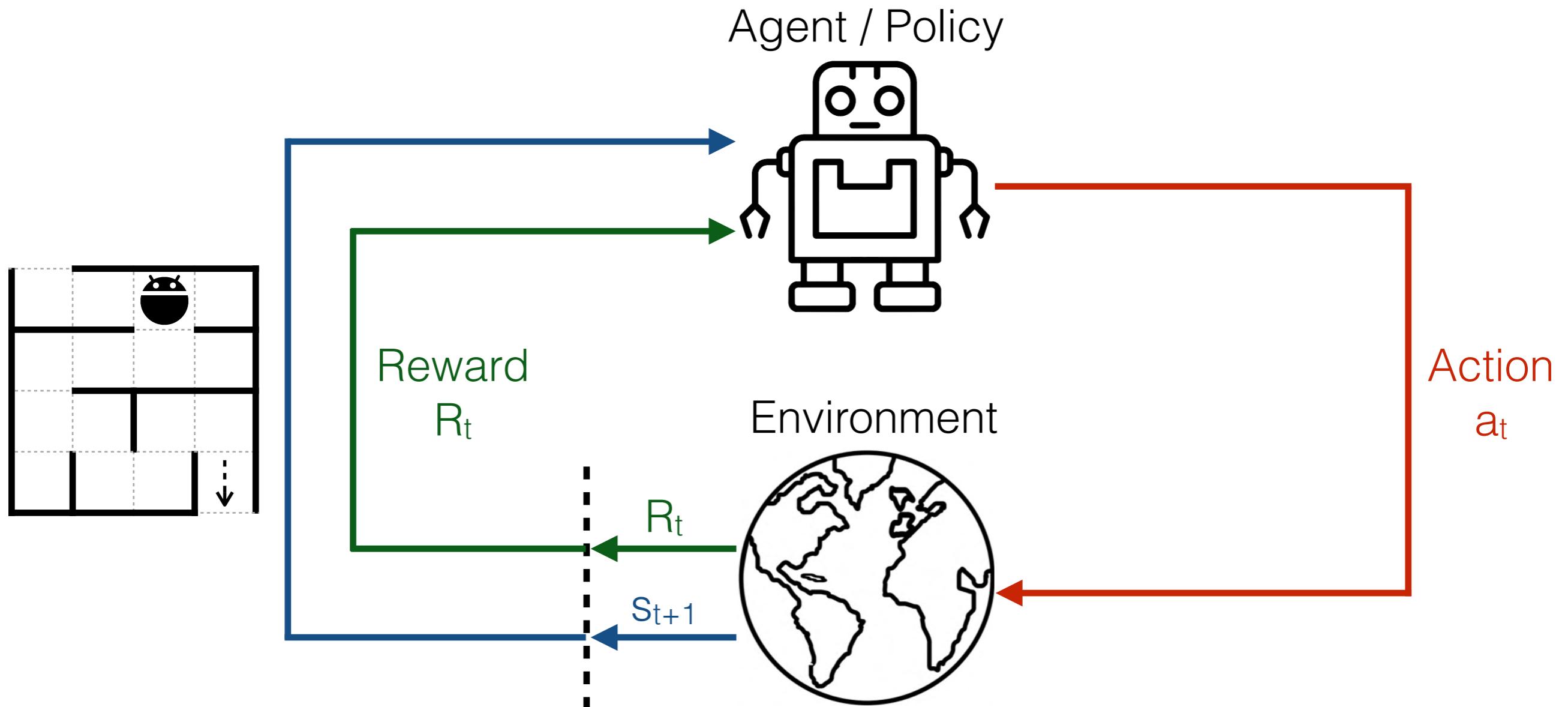
Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

# Robot Learning via Reinforcement Learning



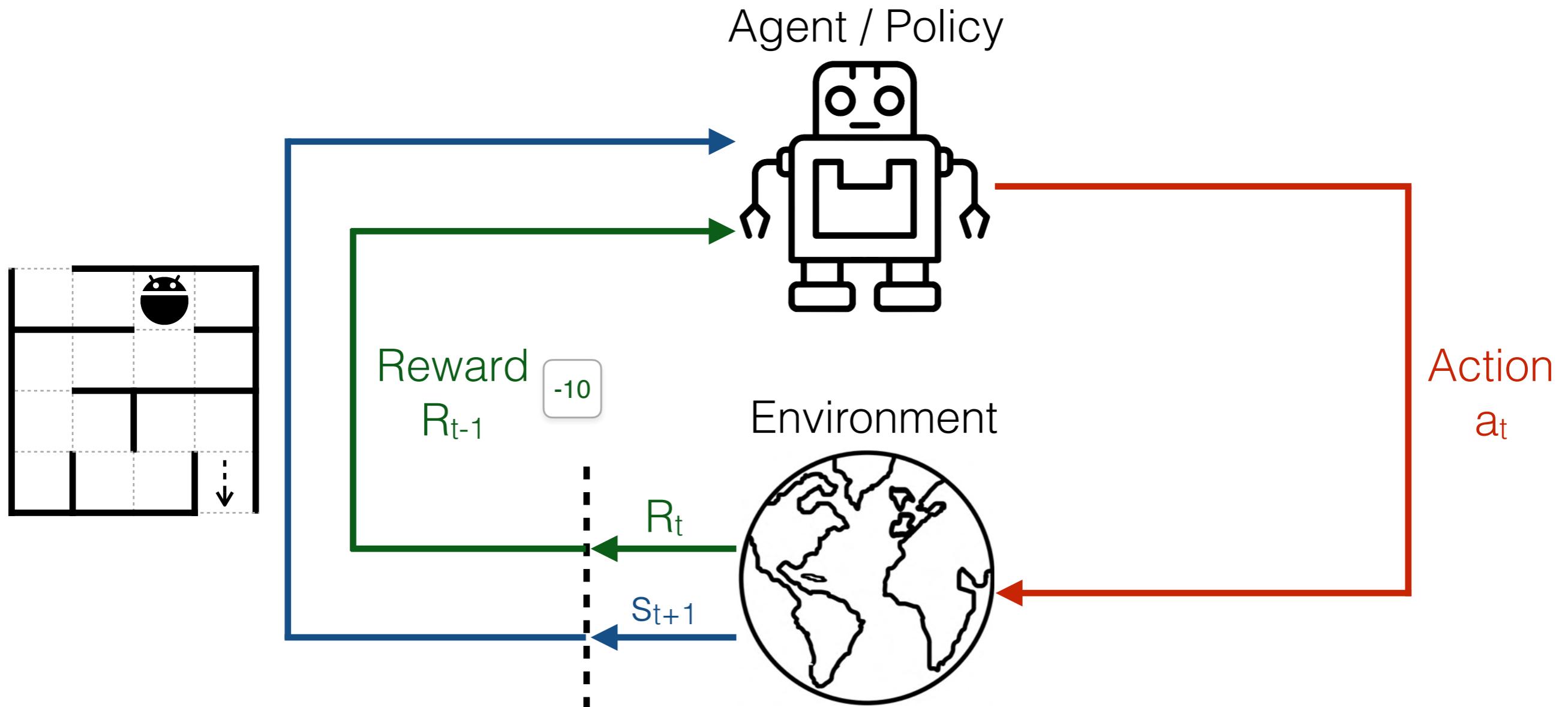
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



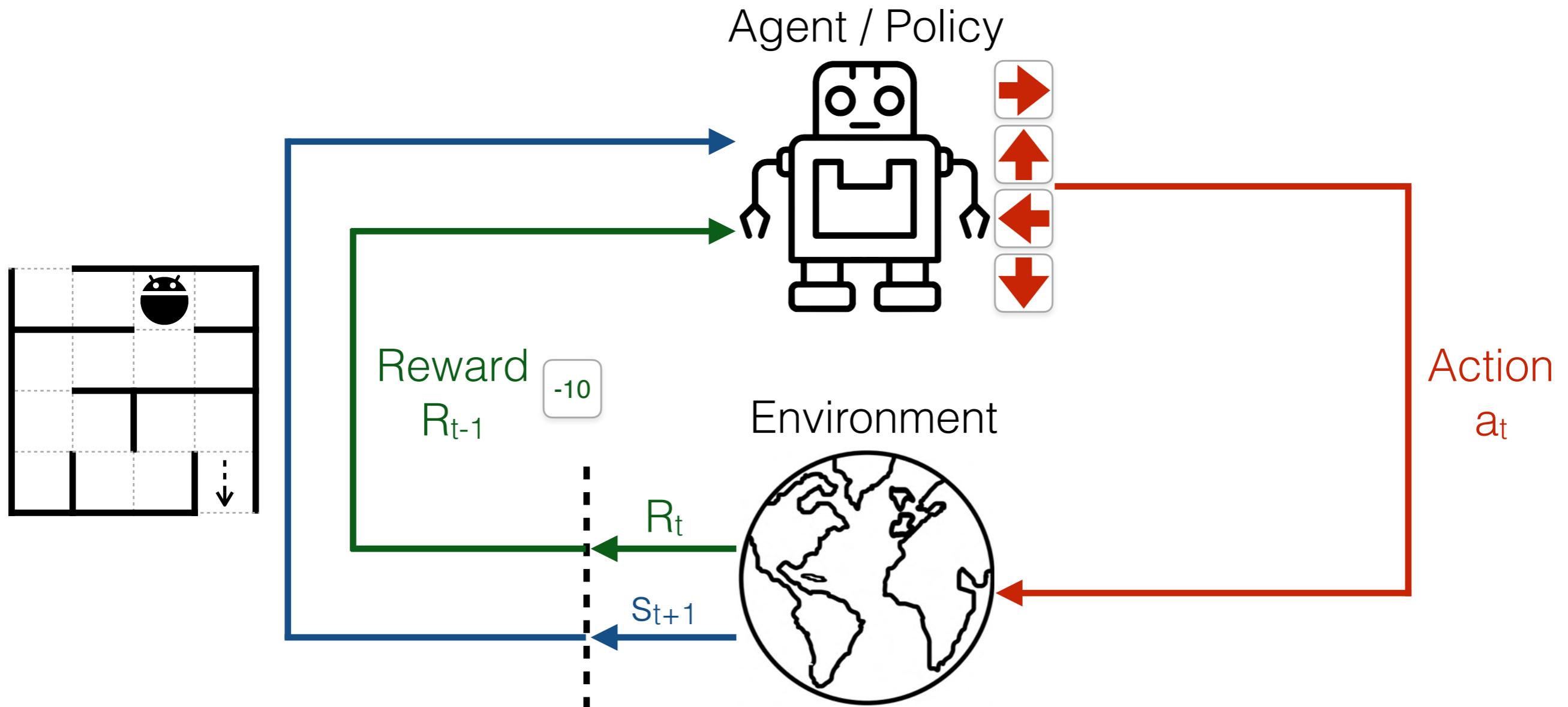
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



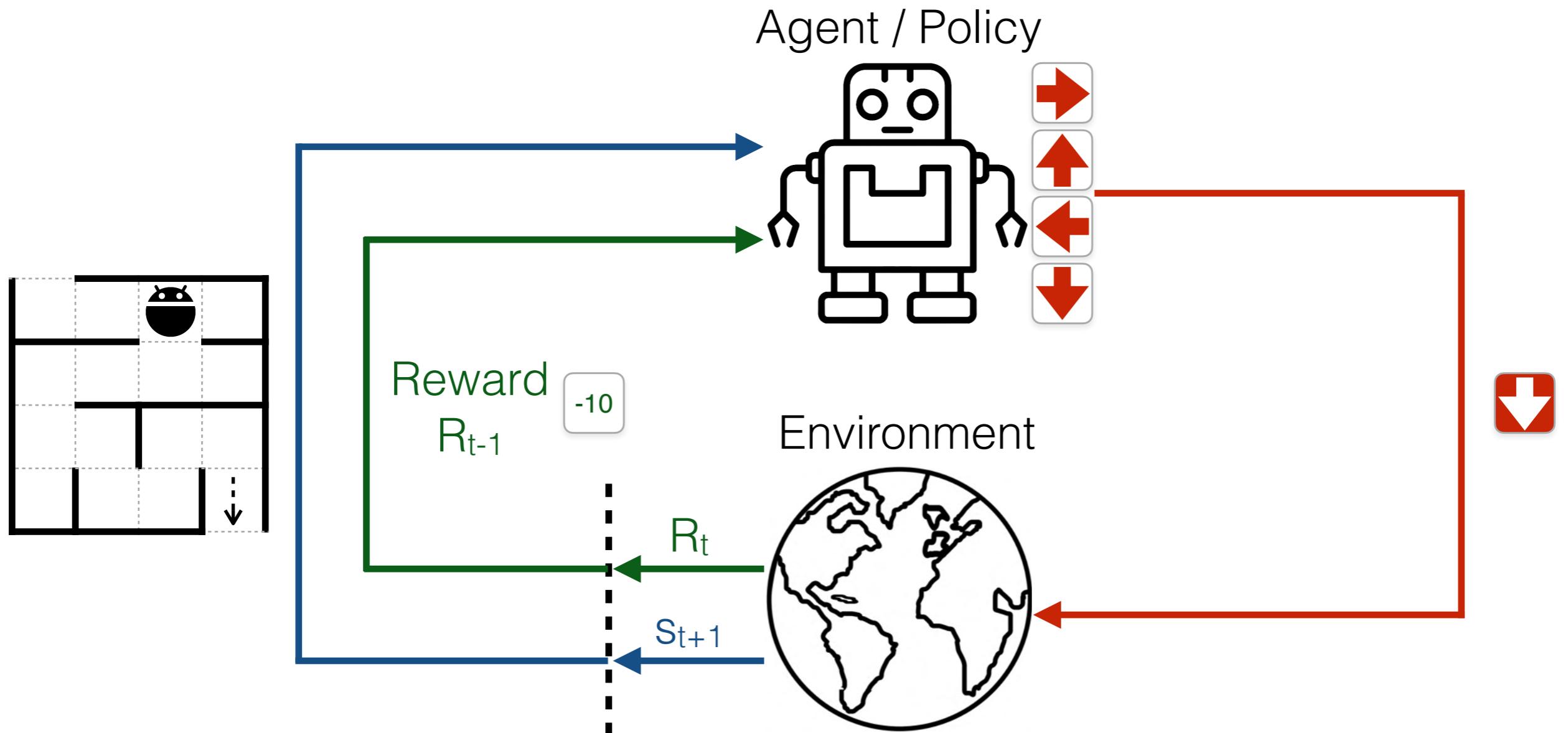
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



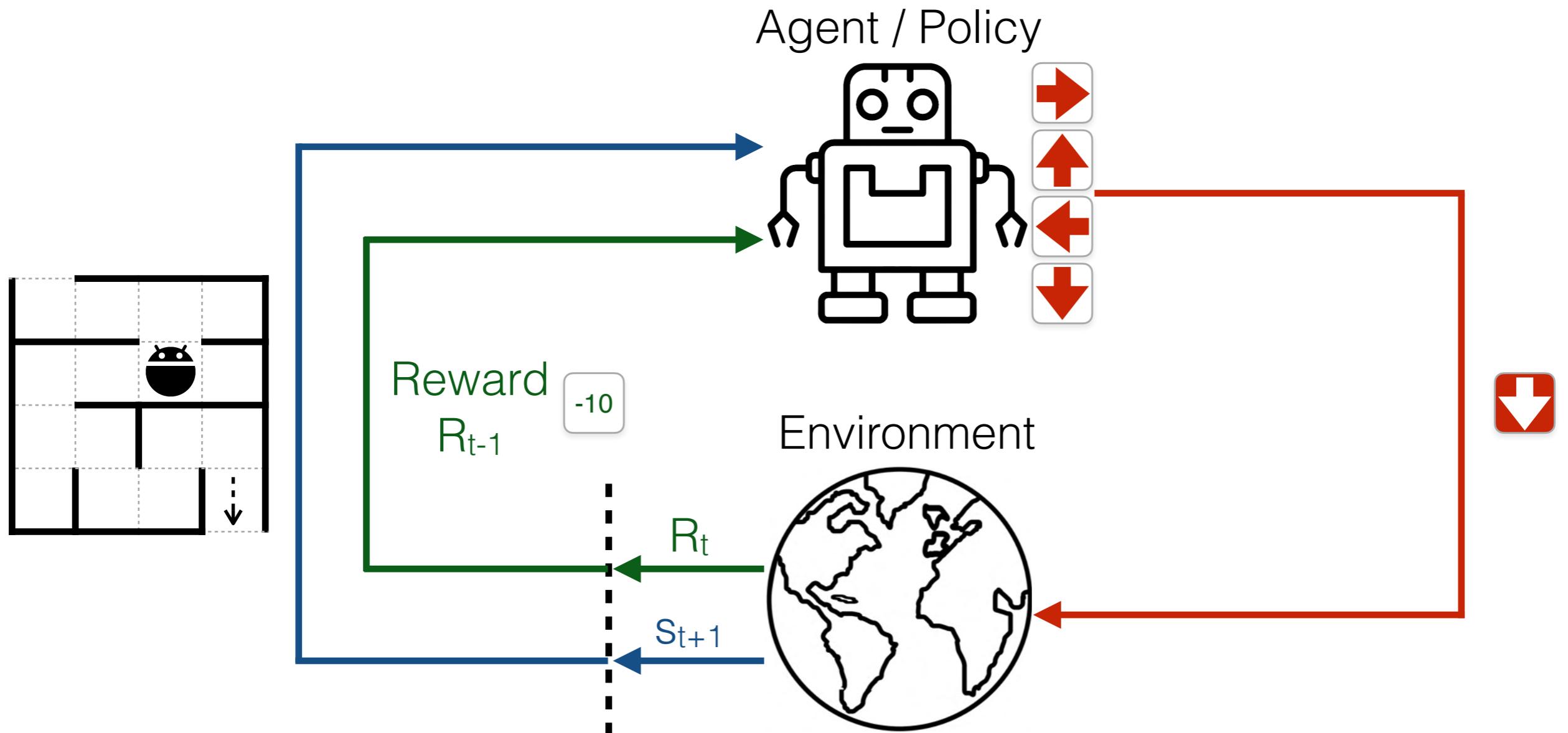
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



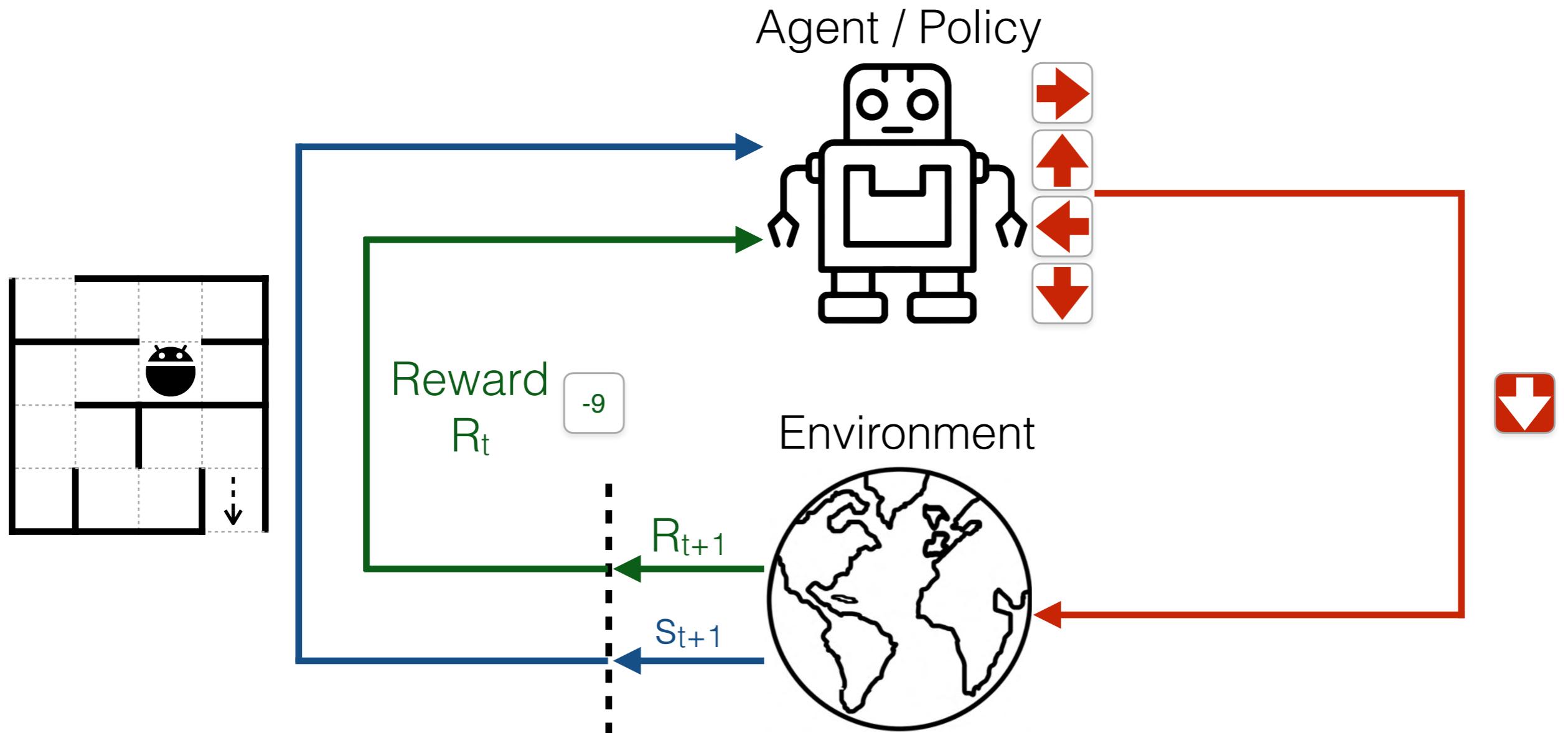
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



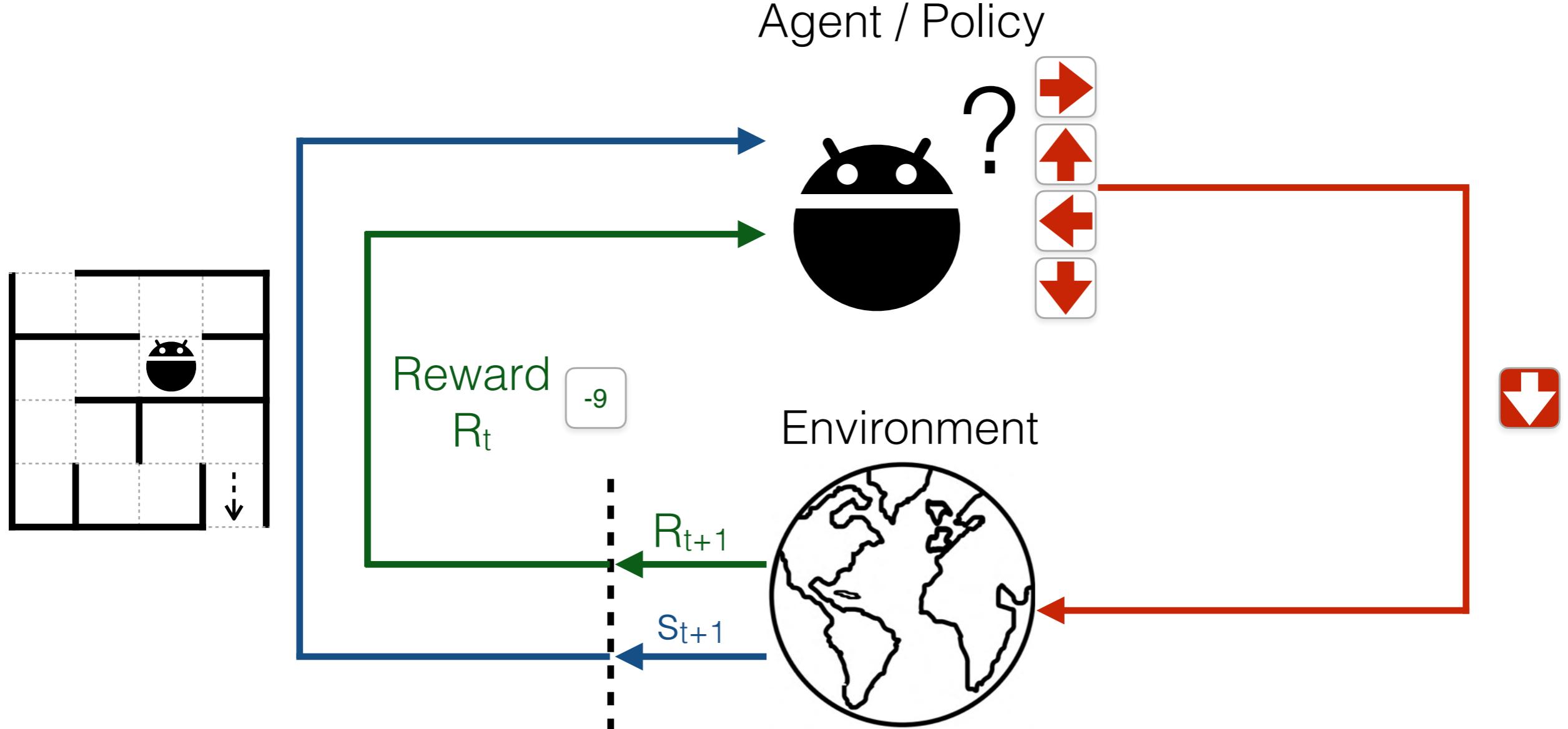
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



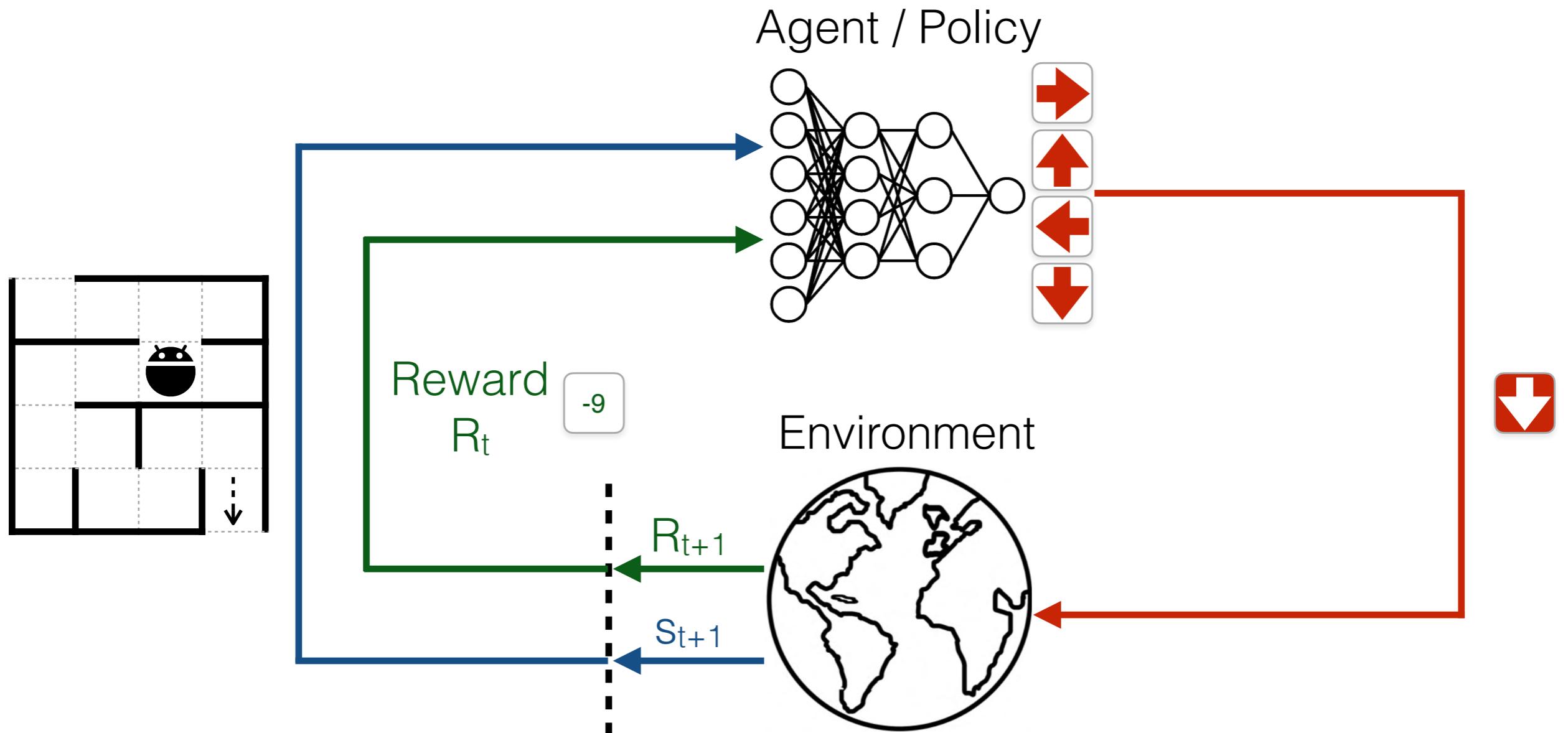
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



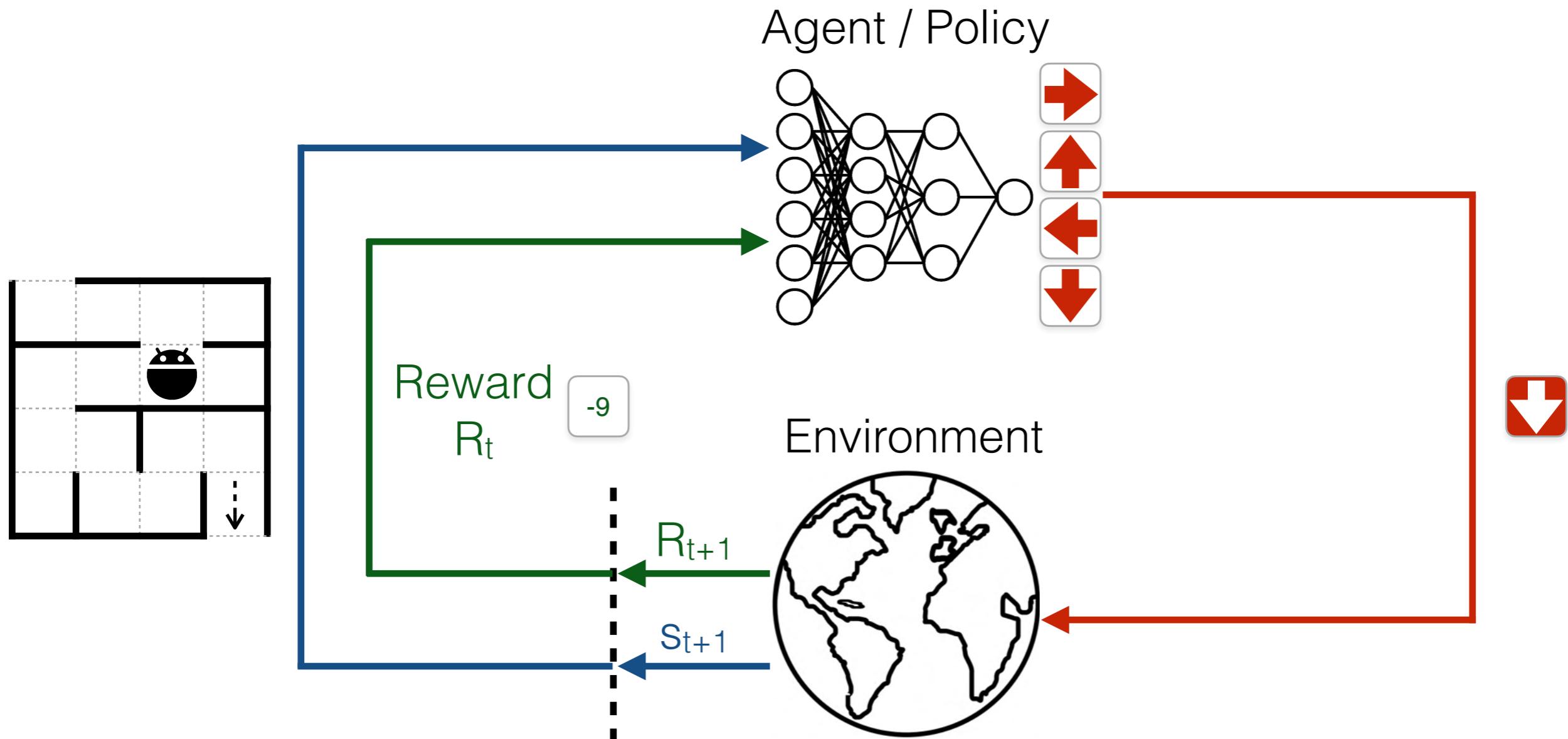
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# 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



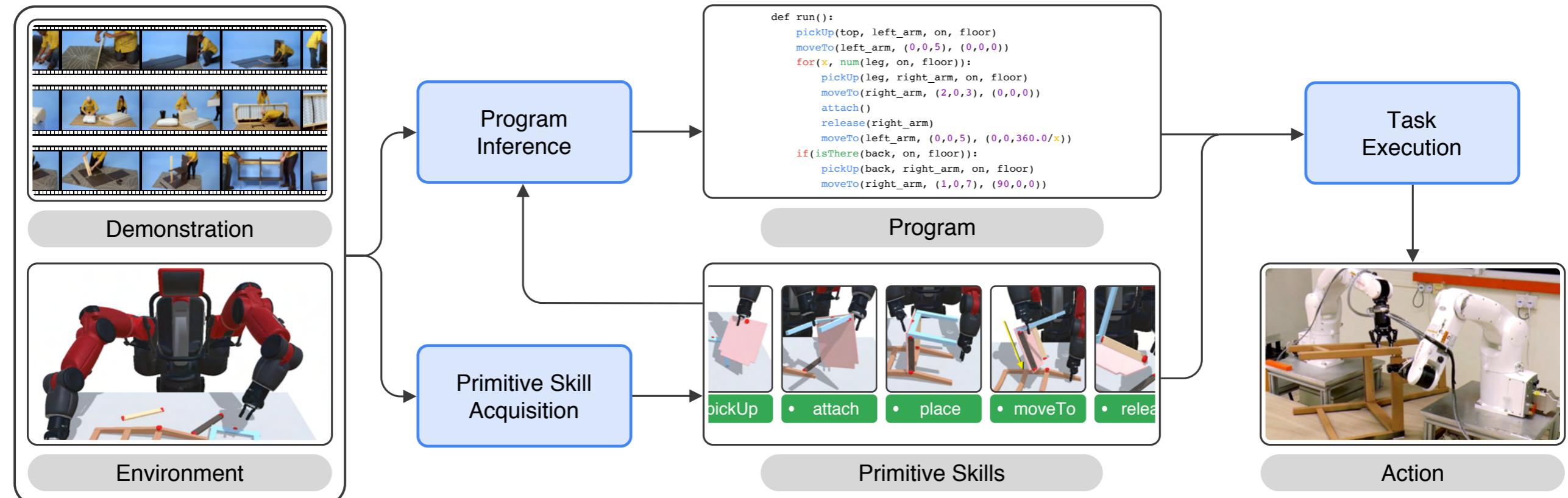
Uninterpretable

Not generalizable

Limited to short-horizon tasks

No skill-reuse

# Program-Guided Framework for Interpreting and Acquiring Complex Skills with Learning Robots



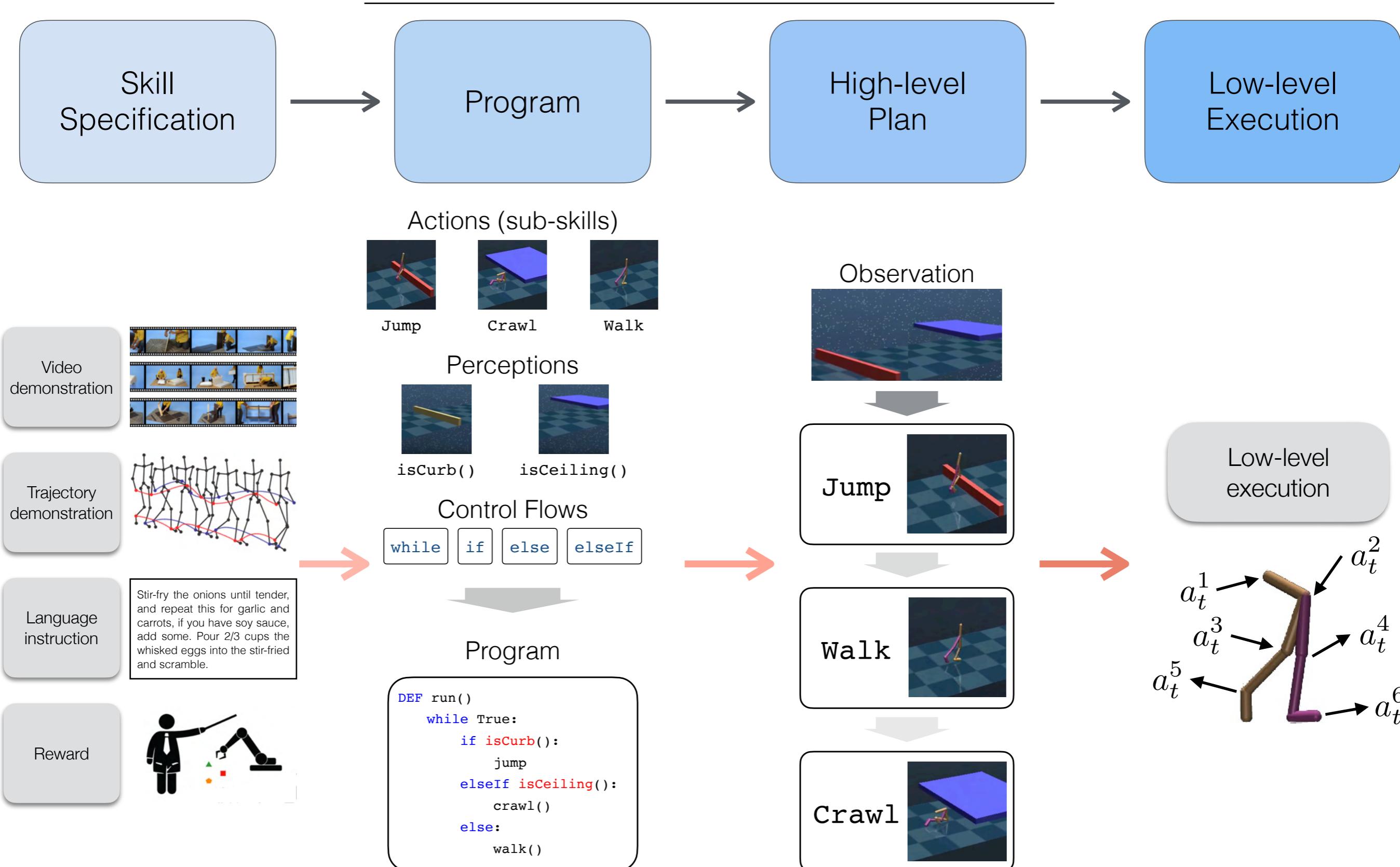
Interpretable

Programmatic /  
Generalizable

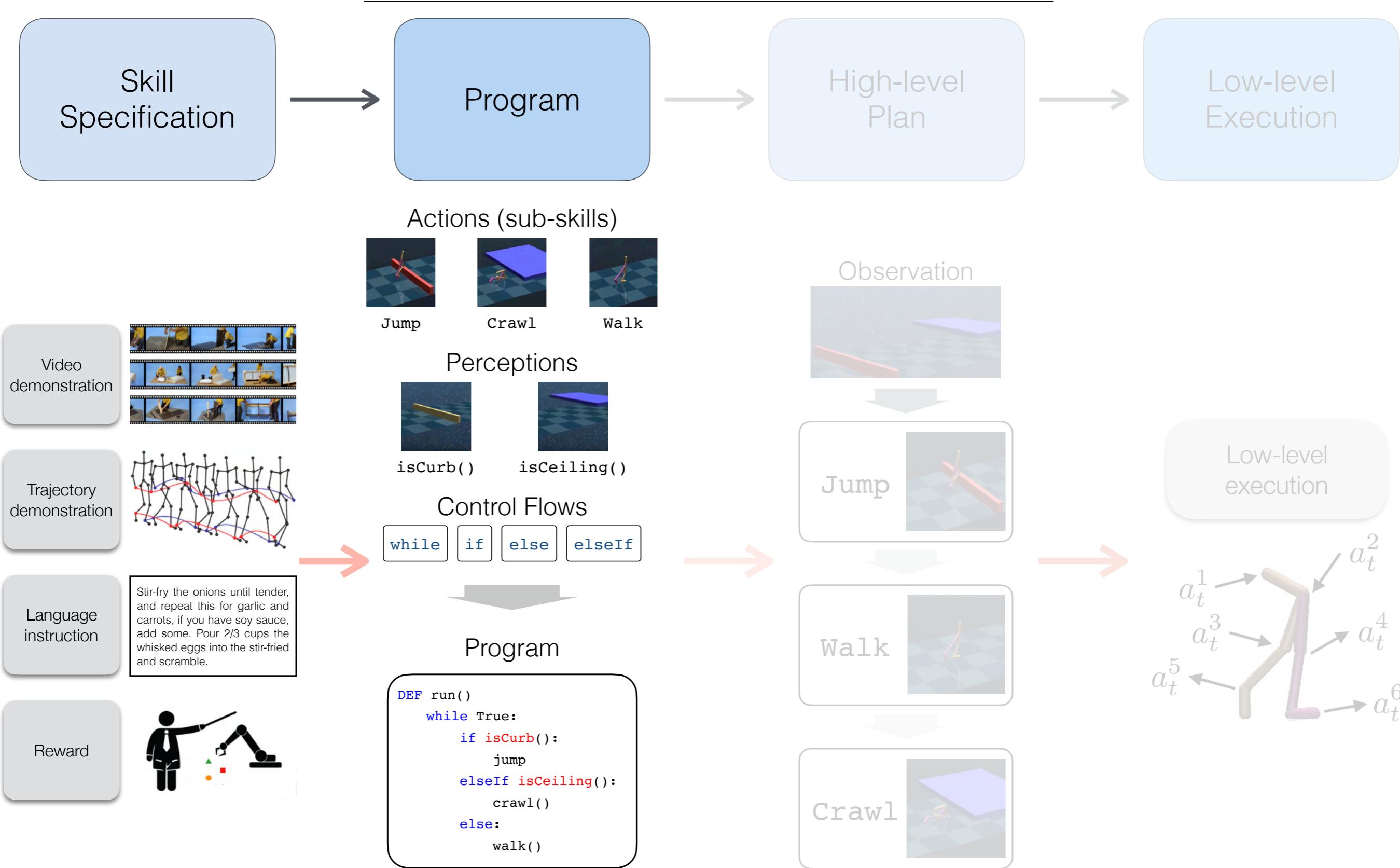
Hierarchical

Modular

# Program-Guided Framework for Interpreting and Acquiring Complex Skills with Learning Robots

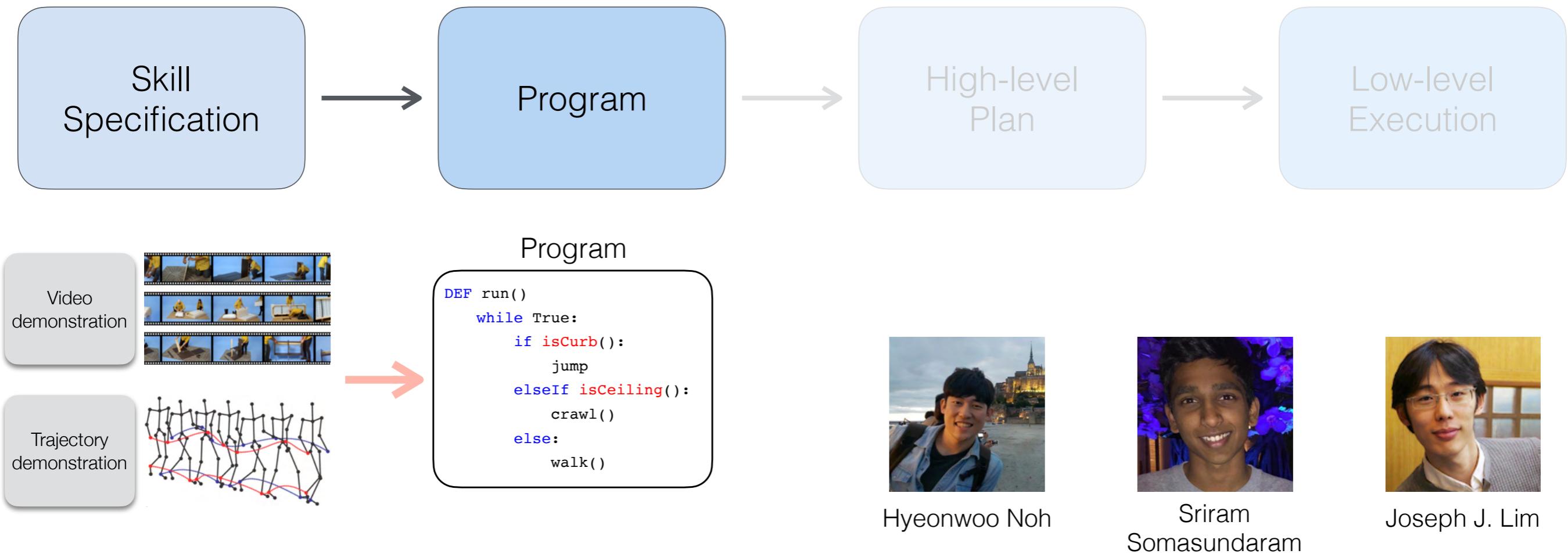


# Program Inference



# Neural Program Synthesis from Diverse Demonstration Videos

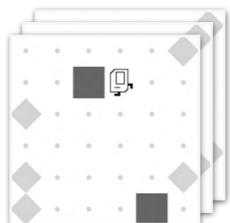
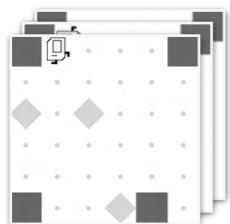
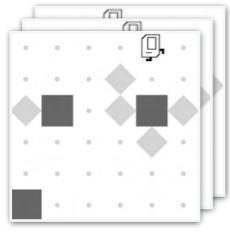
ICML 2018



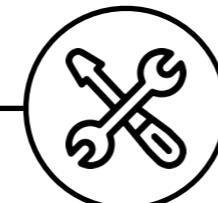
# Imitation Learning

---

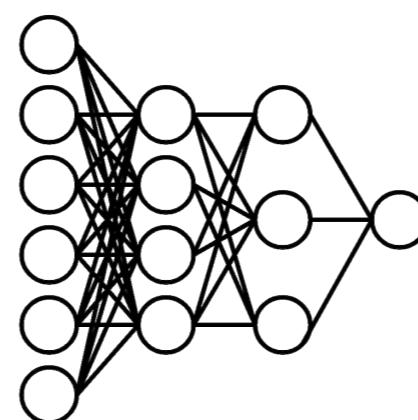
Demonstrations



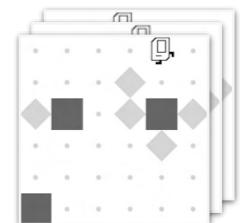
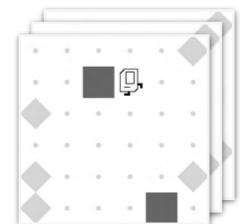
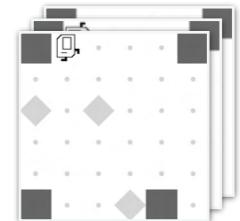
Imitate



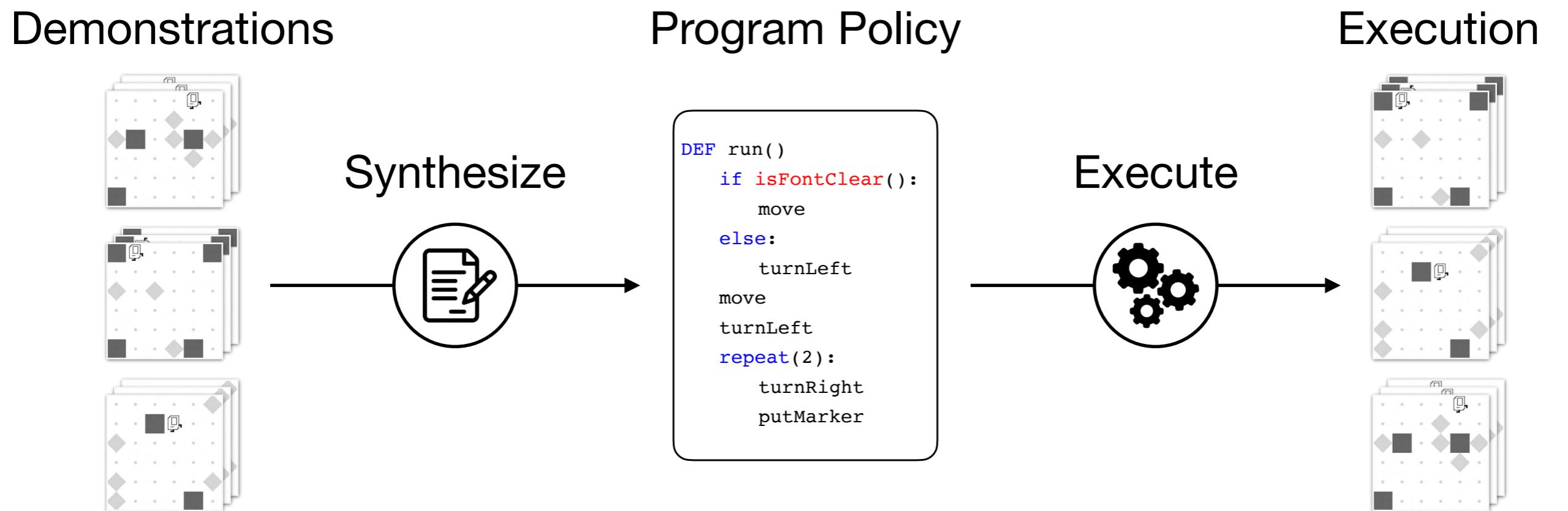
Neural Network  
Policy



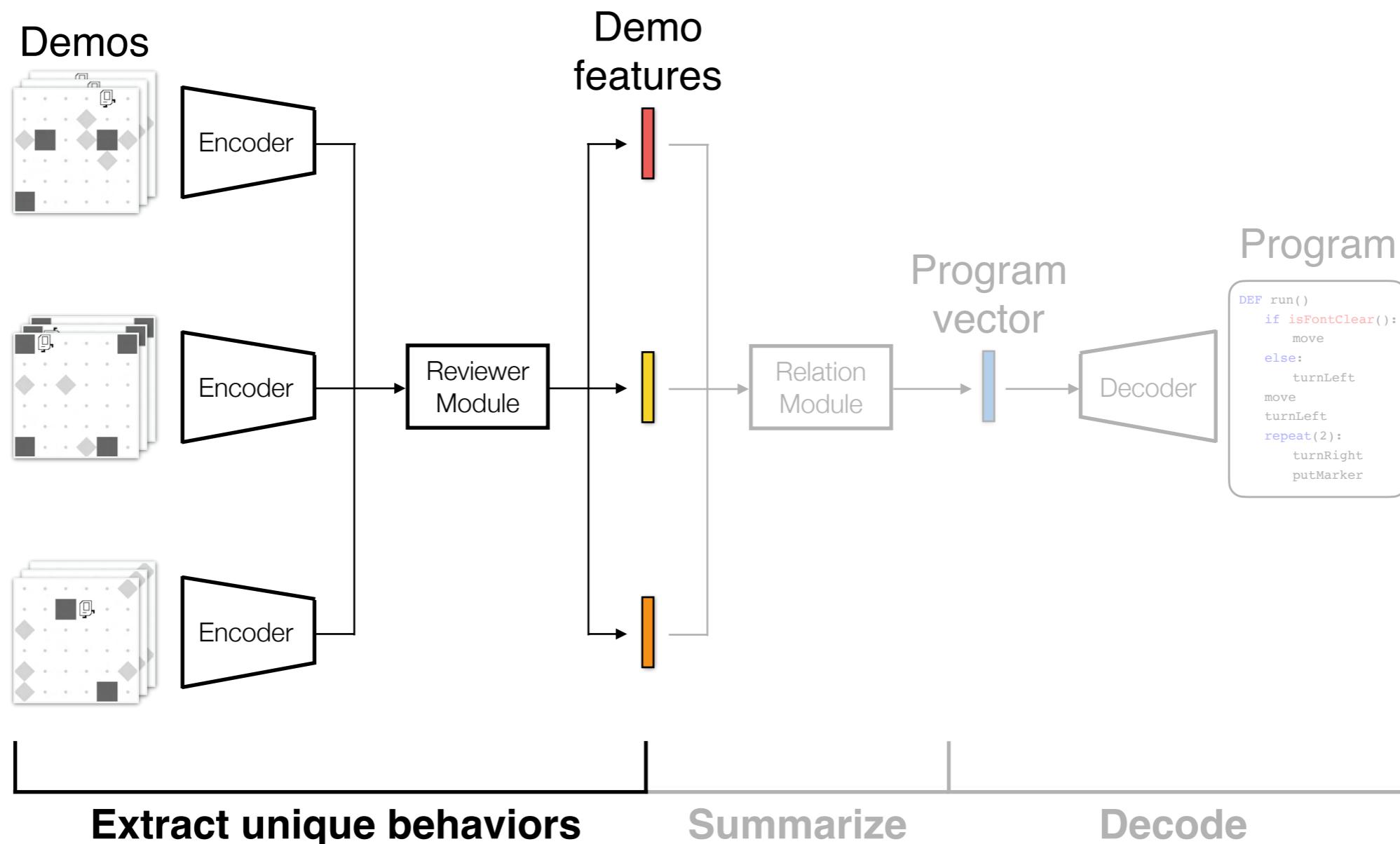
Execution



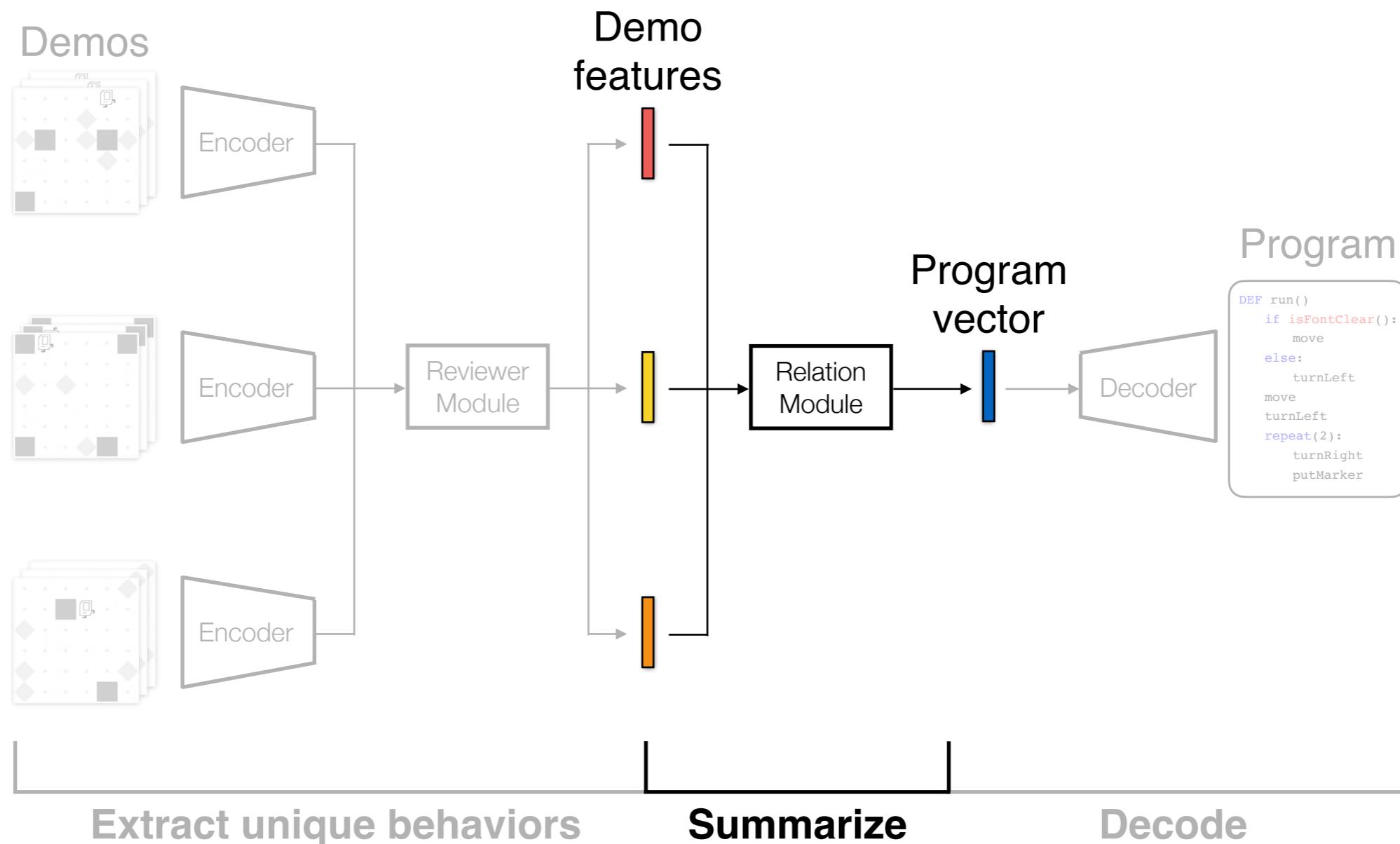
# Imitation Learning by Synthesizing Programs



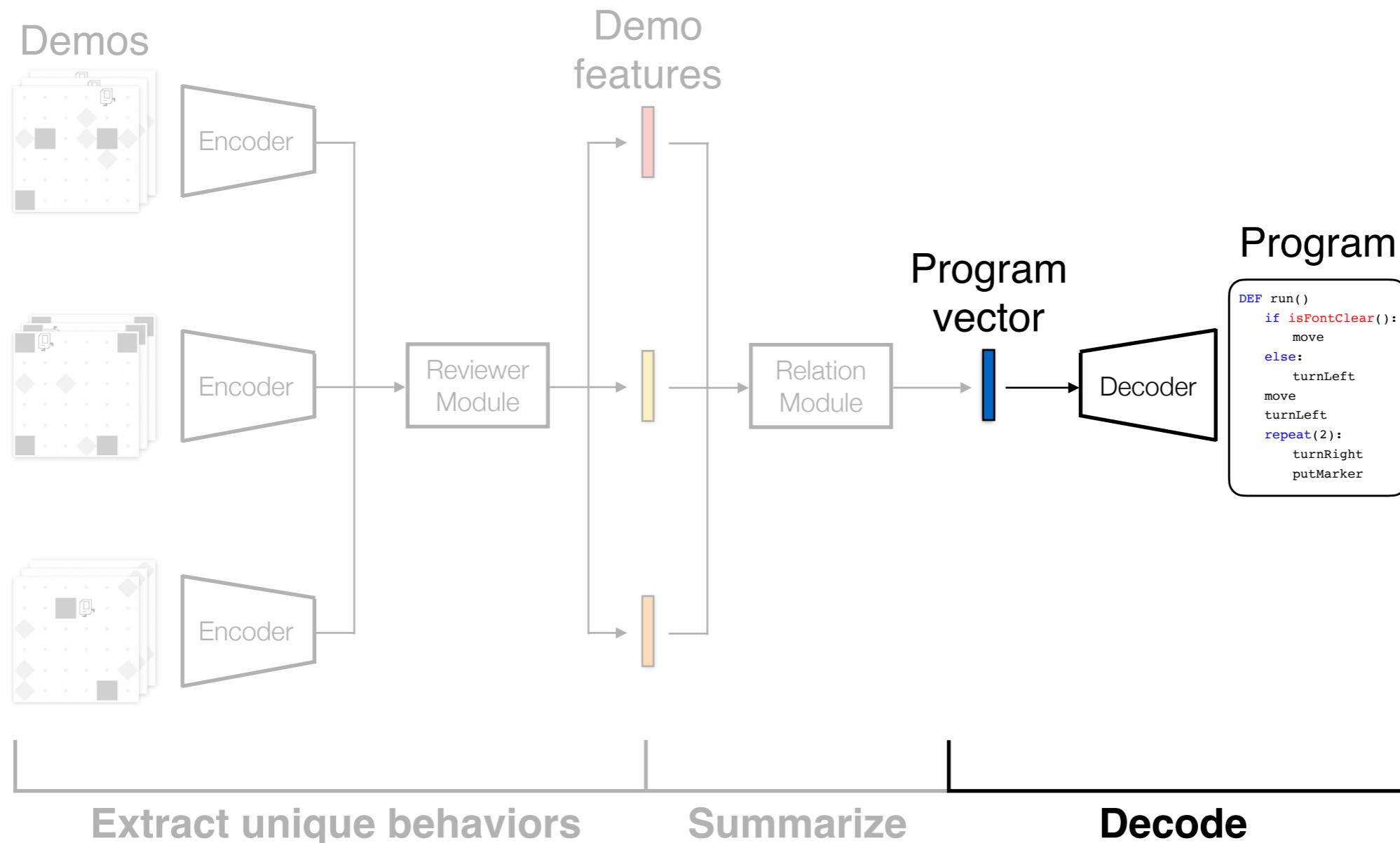
# Model Overview



# Model Overview



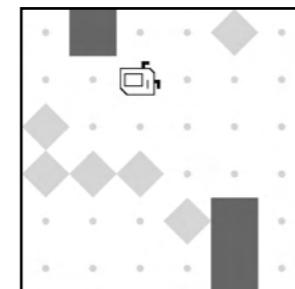
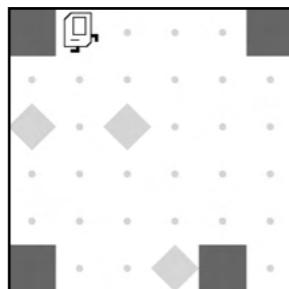
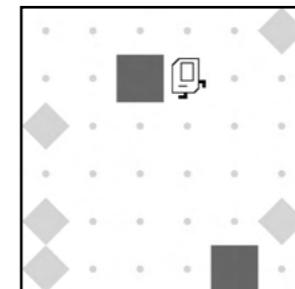
# Model Overview



# Environments

Karel

```
DEF run()
    if isFontClear():
        move
    else:
        turnLeft
    move
    turnLeft
    repeat(2):
        turnRight
        putMarker
```



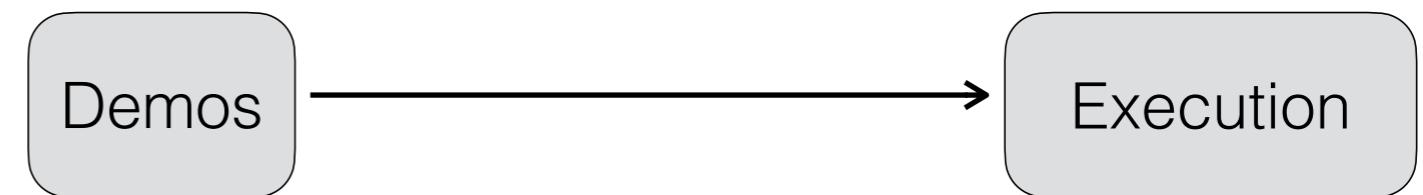
ViZDoom

```
DEF run()
    while isFontClear(HellKnight):
        attack
        moveForward
        if isThere(Demon):
            moveRight
        else:
            moveLeft
            moveBackward
```

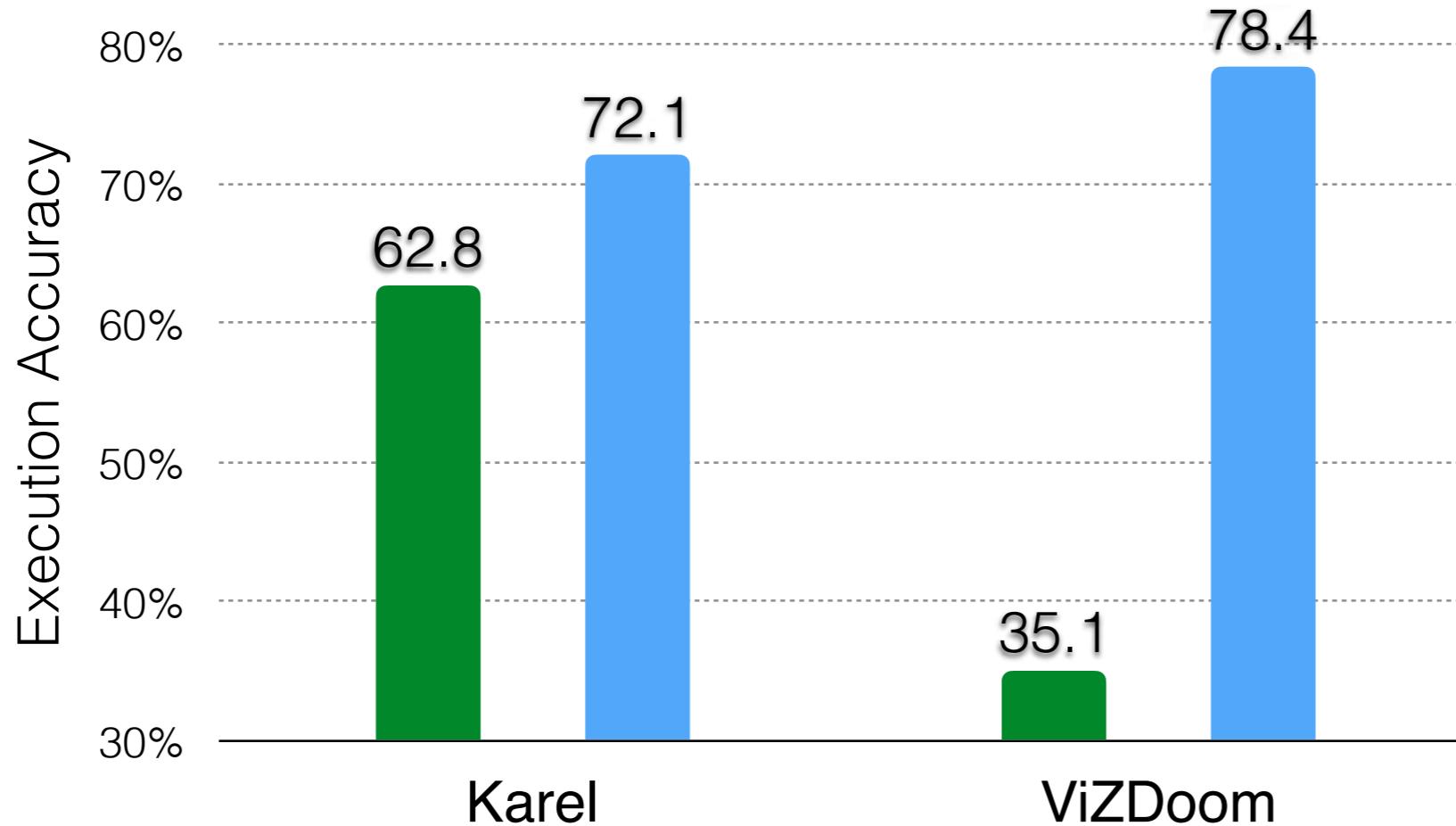
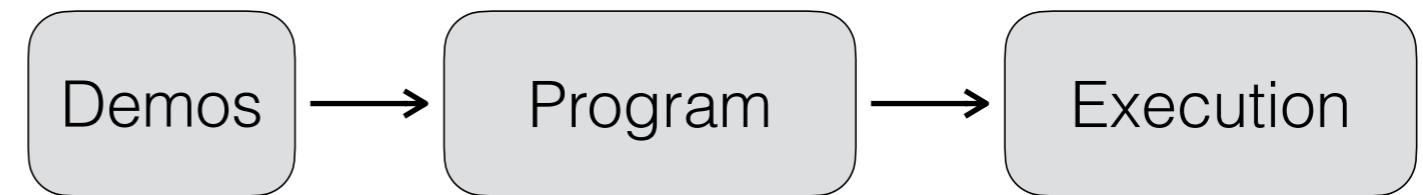


# Quantitative Results

Neural Network Policy



Program Policy



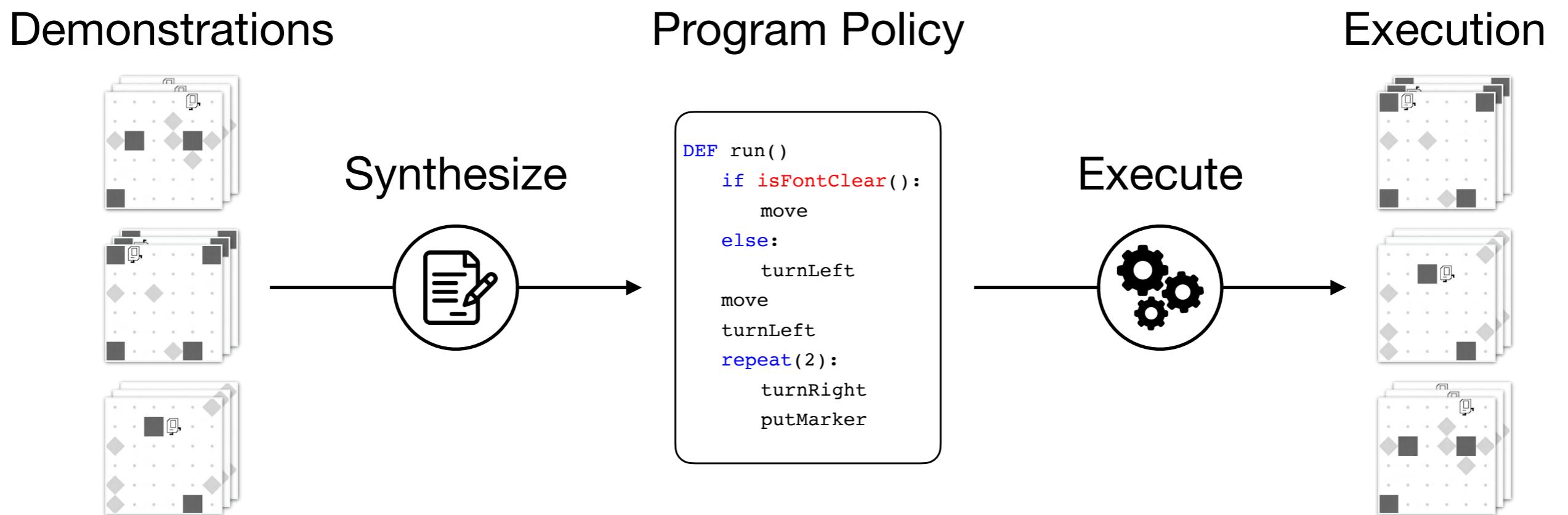
```
DEF run()
    if isFontClear():
        move
    else:
        turnLeft
    move
    turnLeft
    repeat(2):
        turnRight
        putMarker
```

9/10 demos  
Major branch  
Minor branch  
1/10 demos

# Takeaway

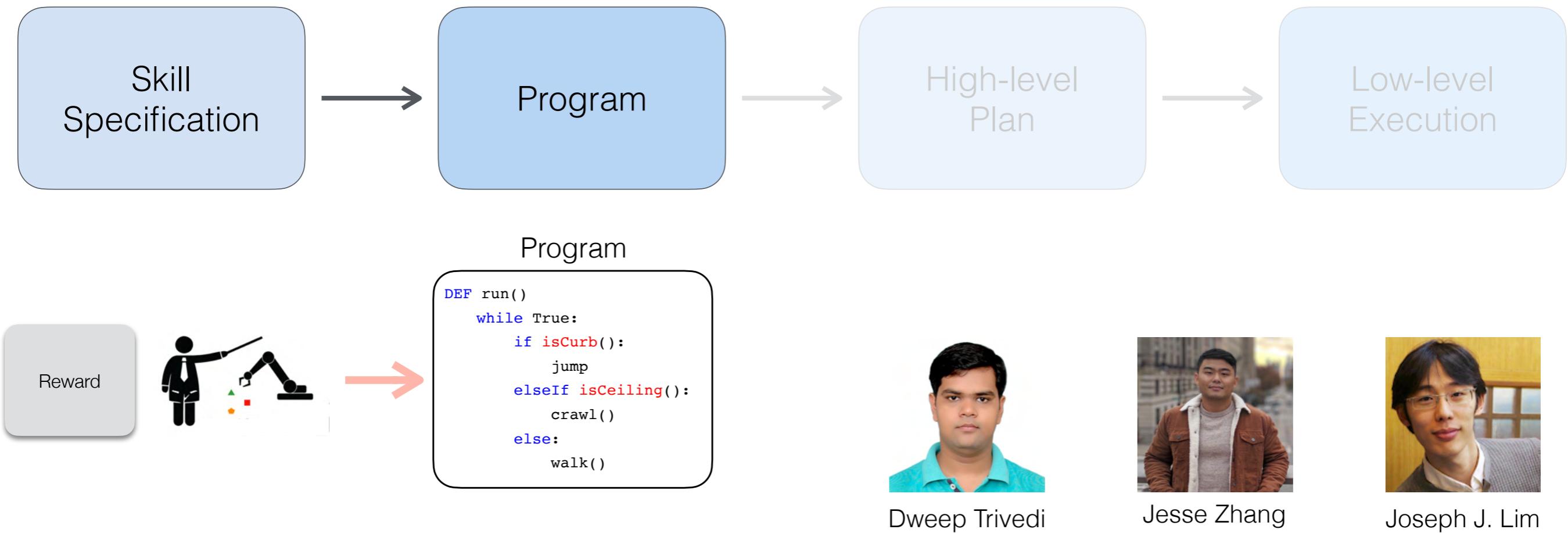
---

- Synthesize programs to imitate demonstrations

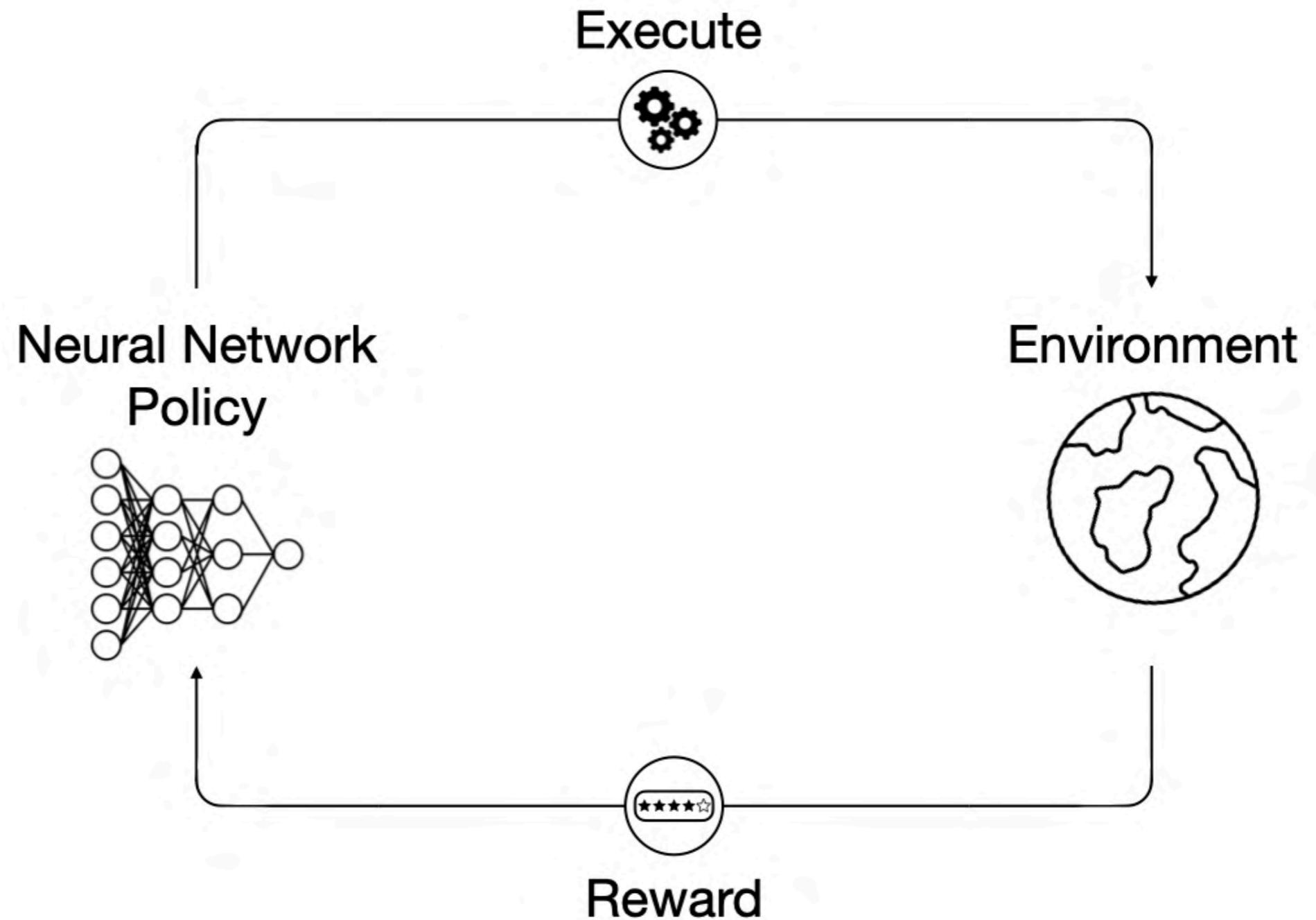


# Learning to Synthesize Programs as Interpretable and Generalizable Policies

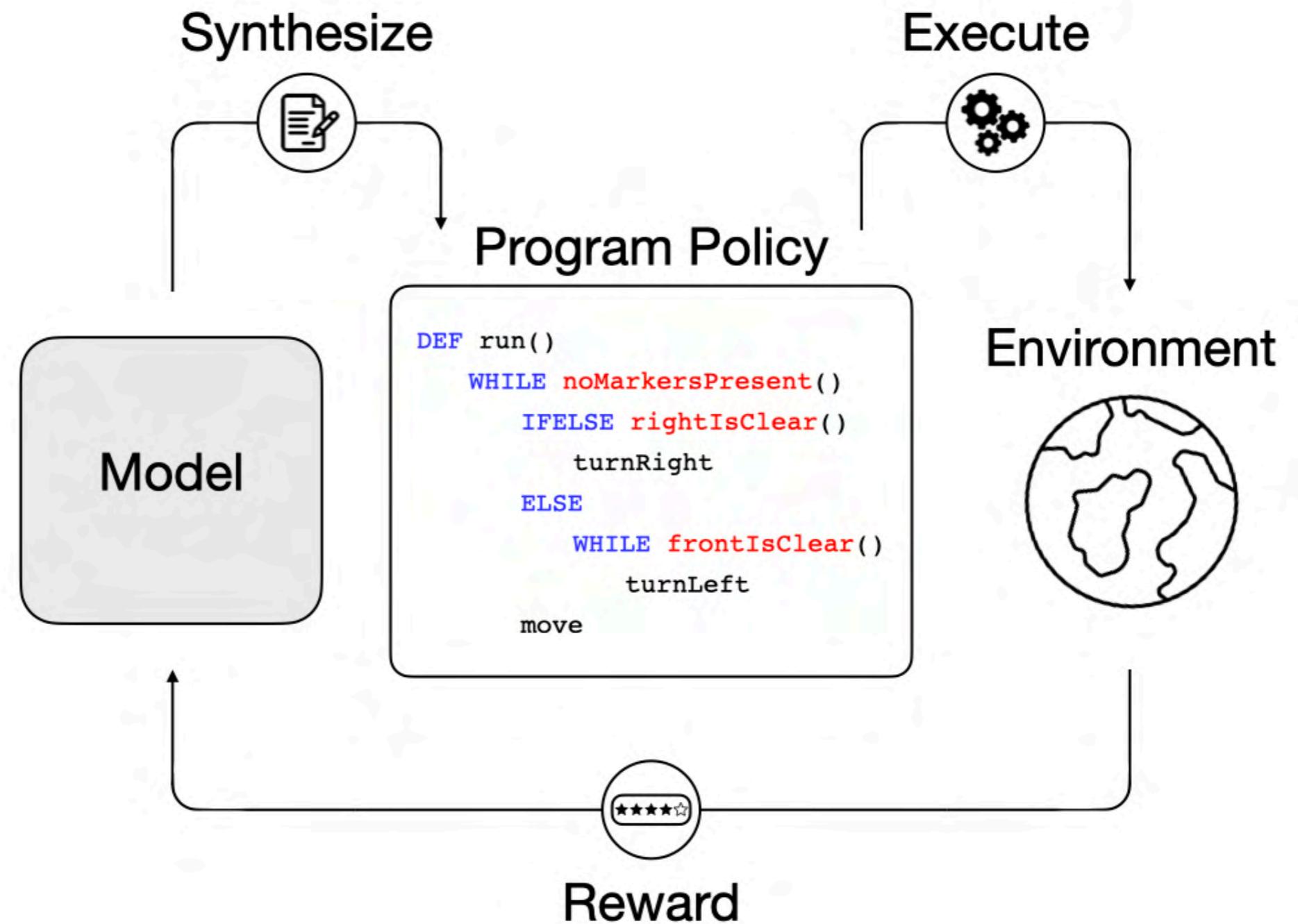
NeurIPS 2021



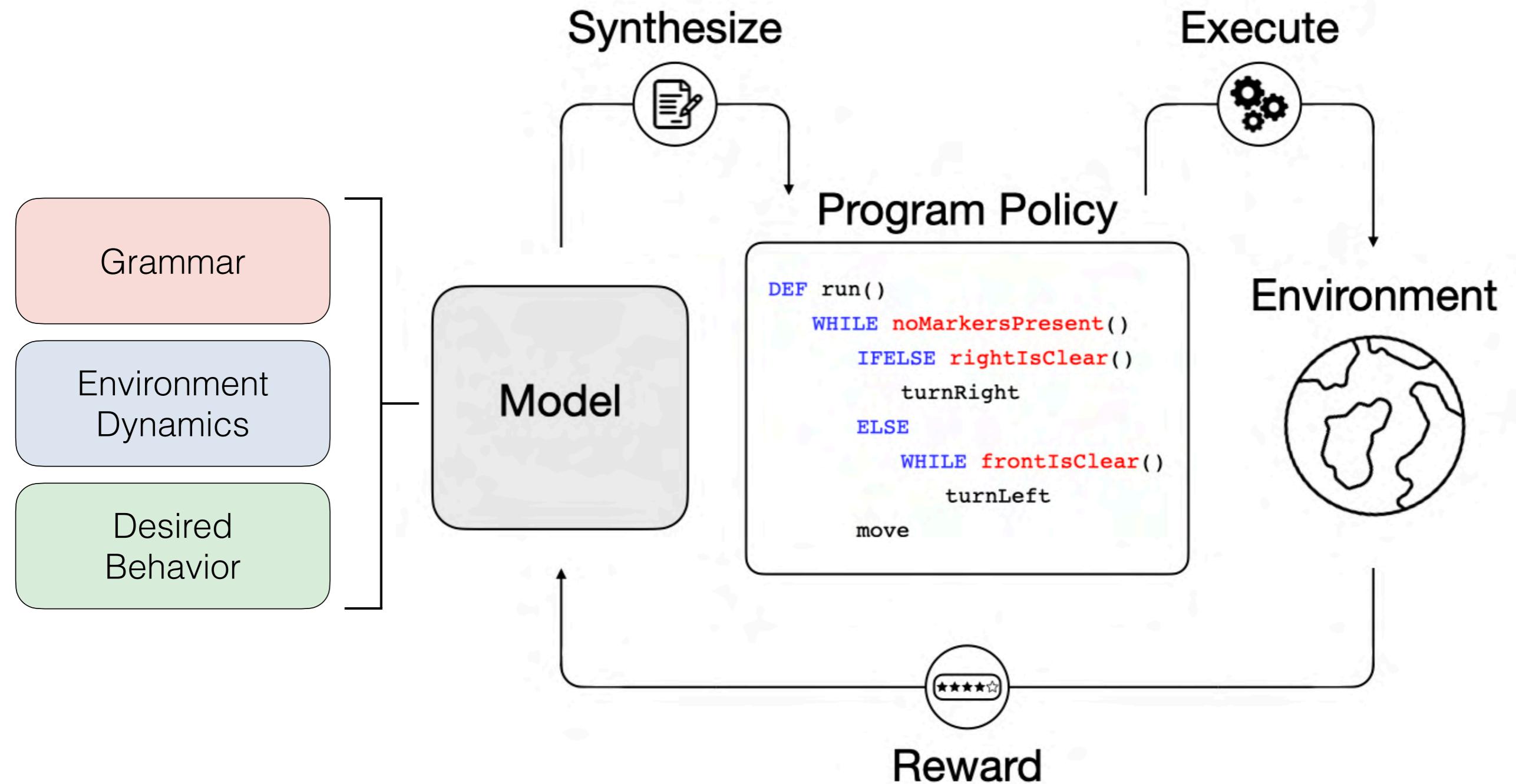
# Reinforcement Learning



# Reinforcement Learning by Synthesizing Programs



# Reinforcement Learning by Synthesizing Programs



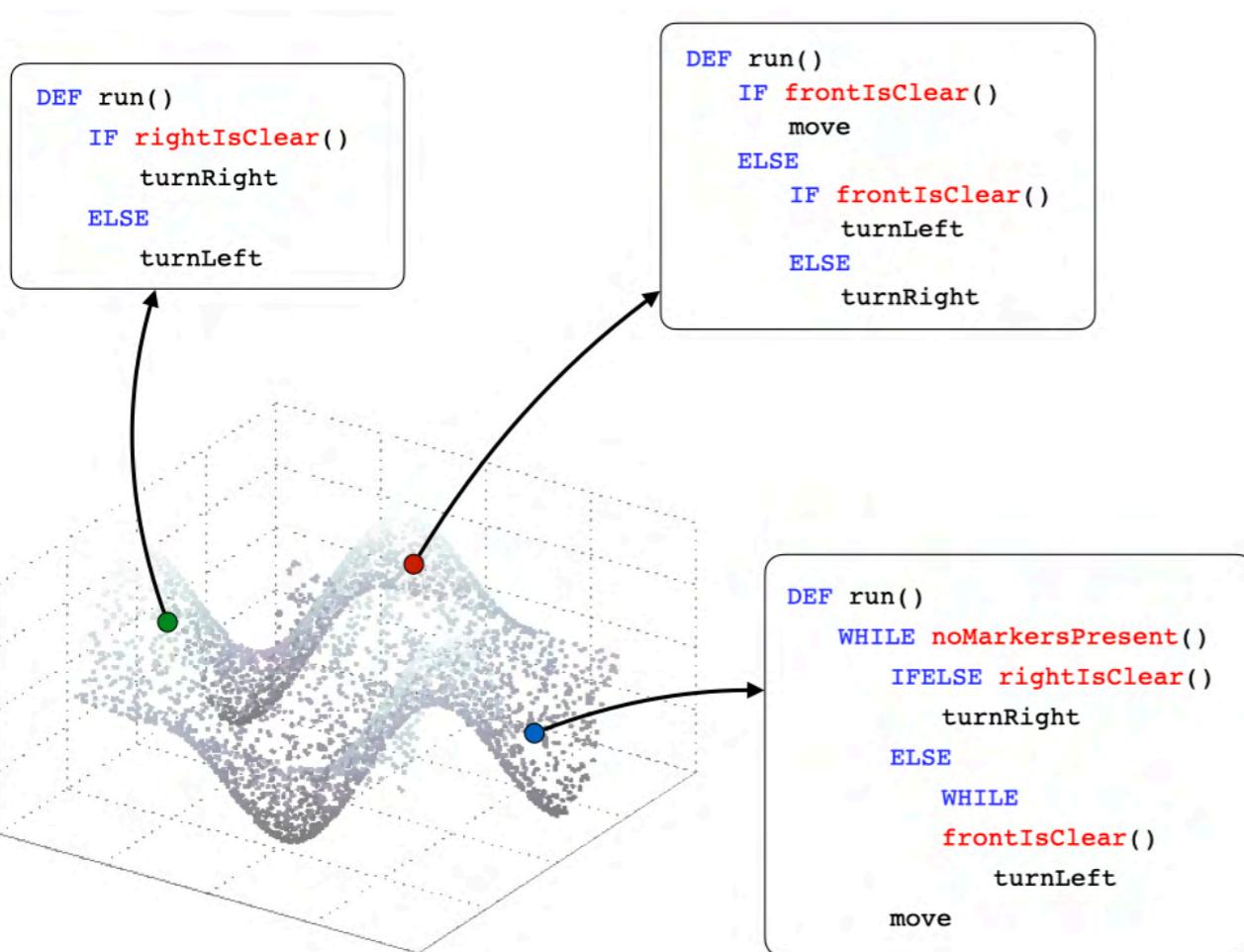
# Method Overview

## Stage 1

Learn a program embedding space from randomly generated programs

Grammar

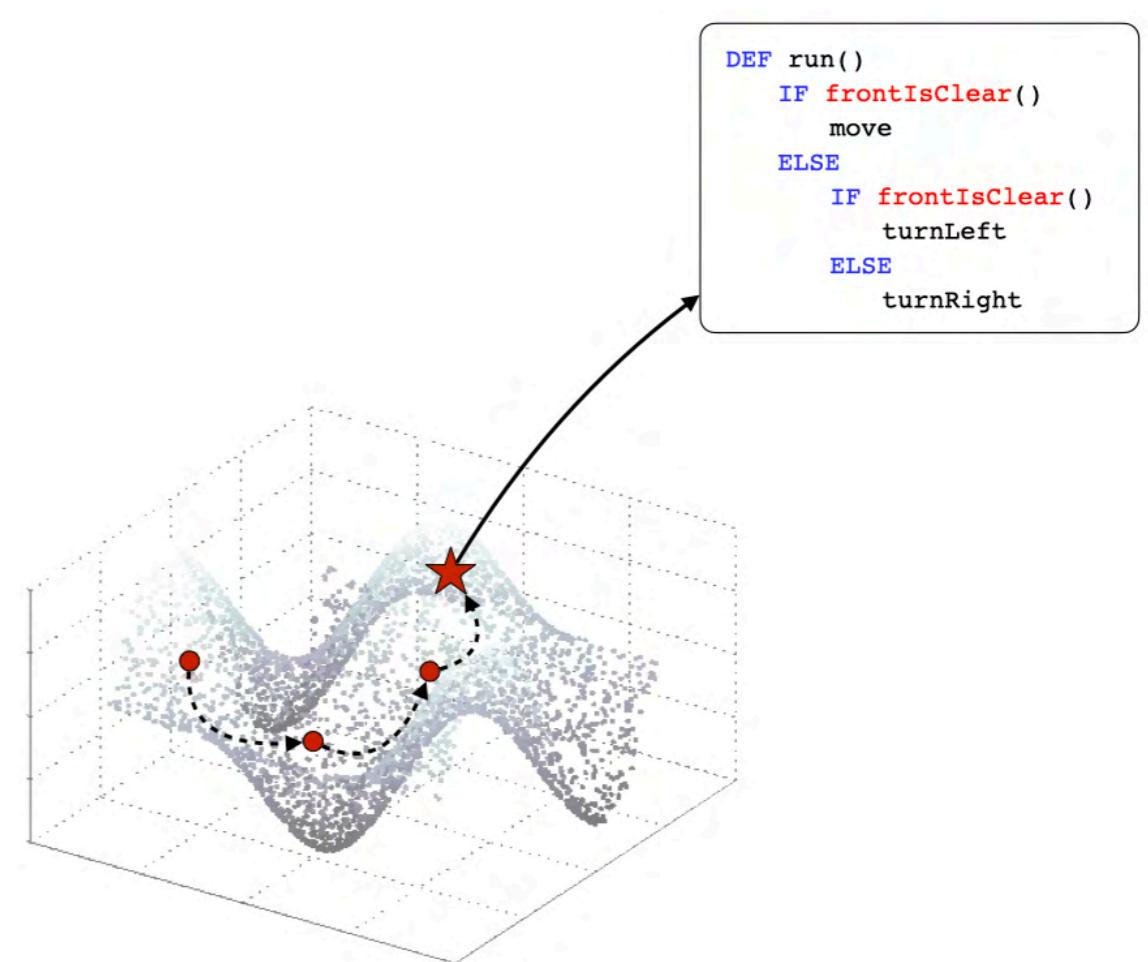
Environment  
Dynamics



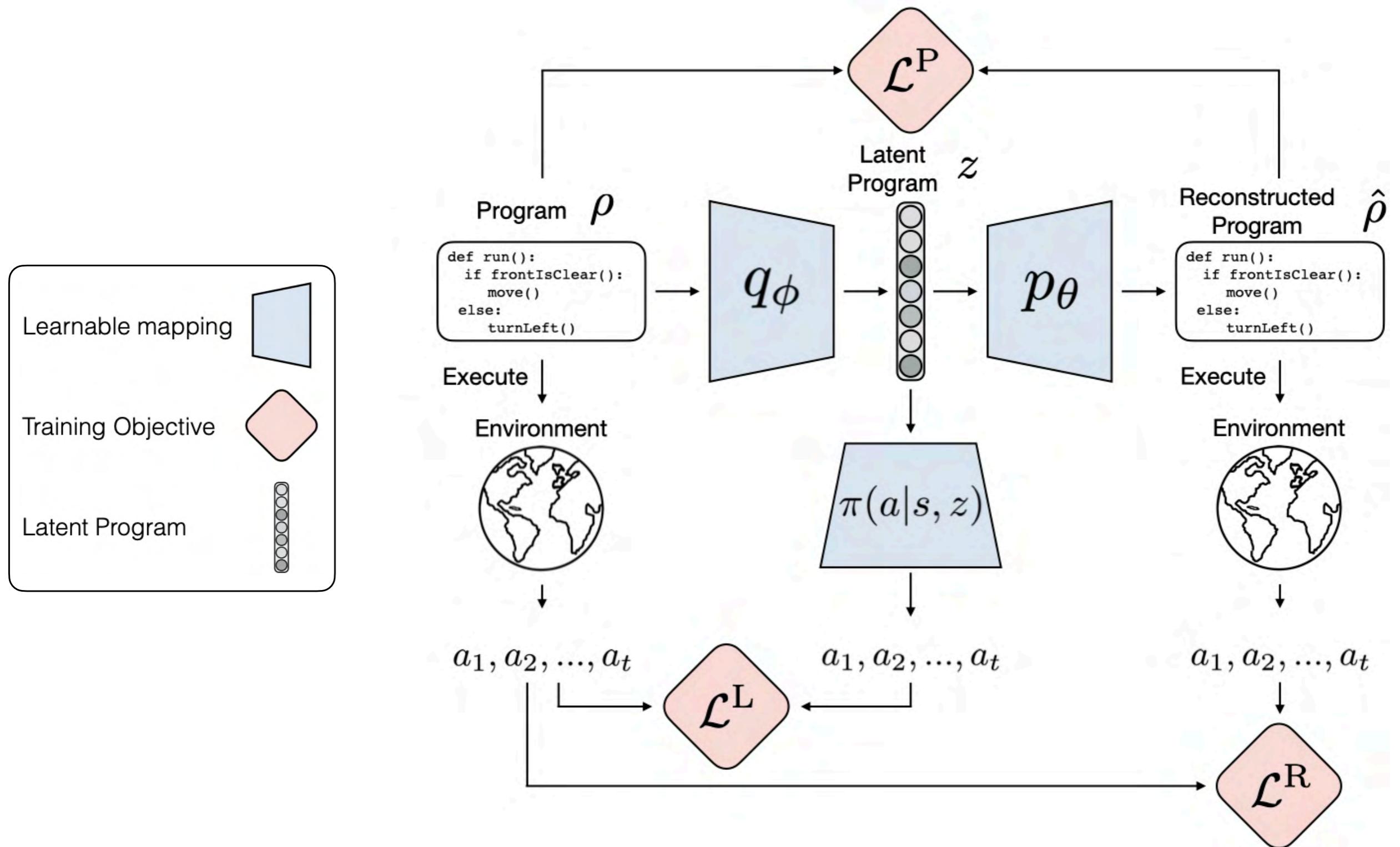
## Stage 2

Search for a task-solving program

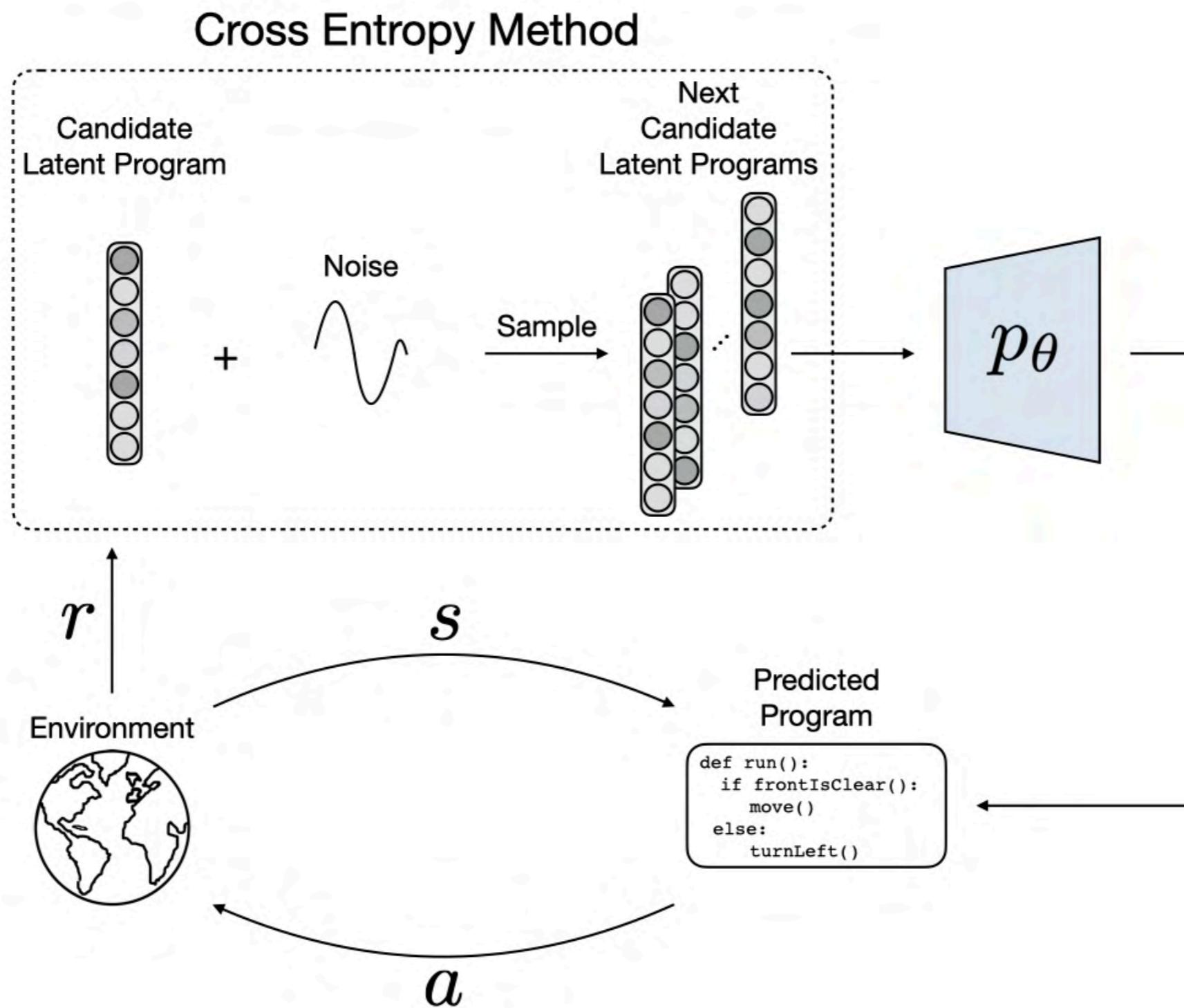
Desired Behavior



# Stage 1: Learning a Program Embedding Space

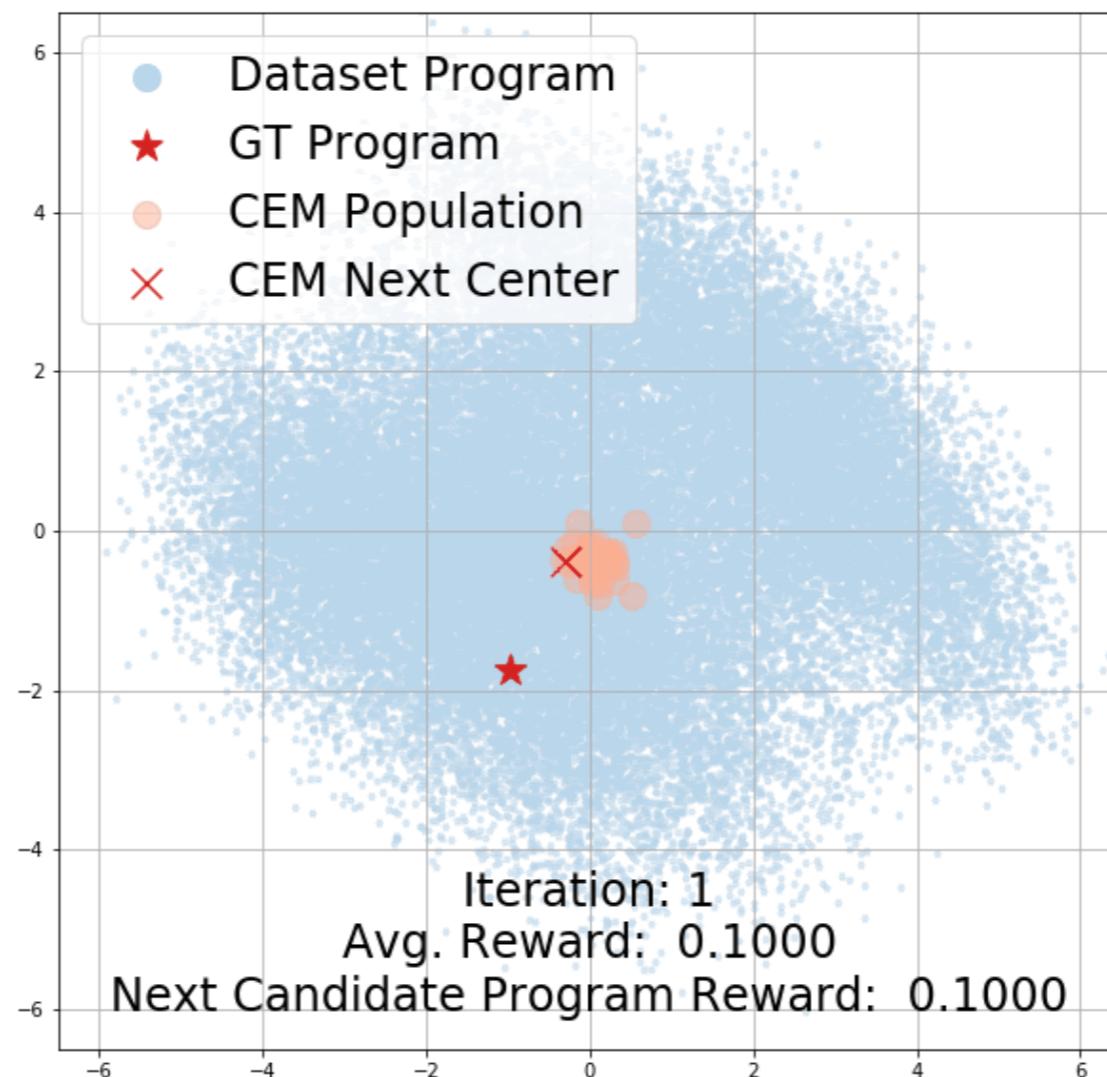


# Stage 2: Latent Program Search with Cross Entropy Method



# Cross Entropy Method Trajectory Visualization

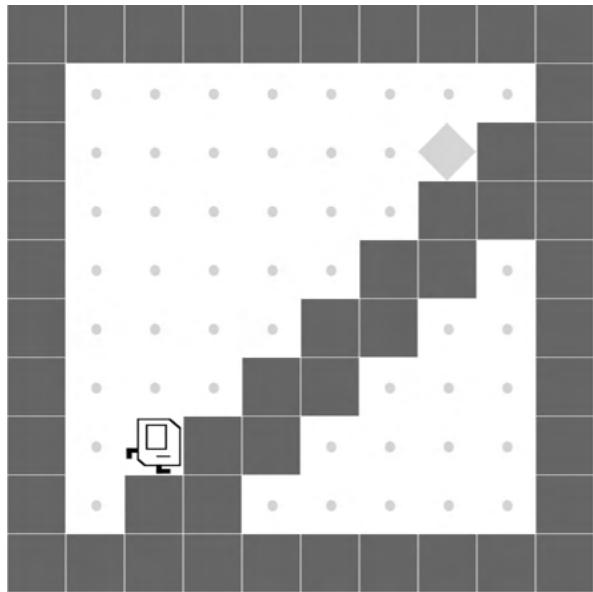
---



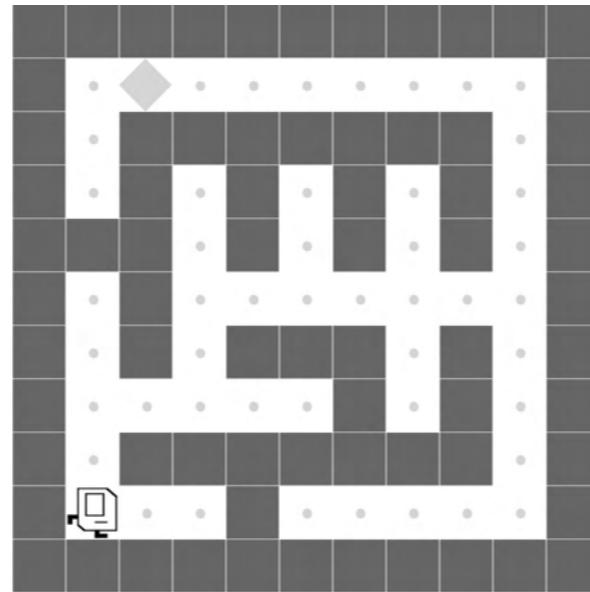
# Karel Tasks

---

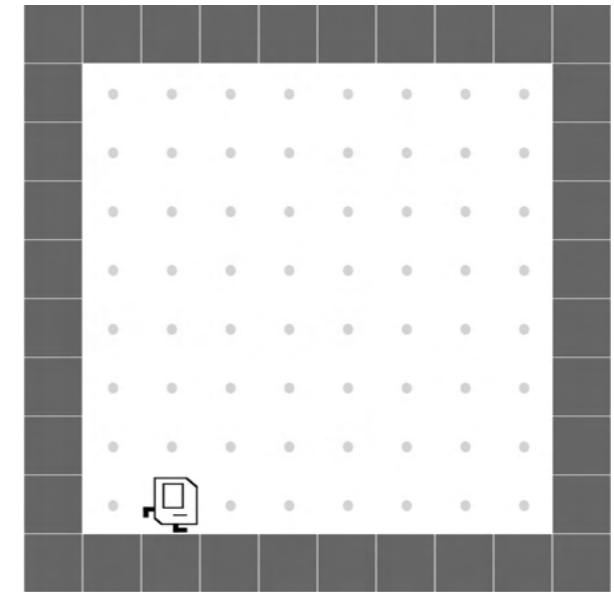
StairClimber



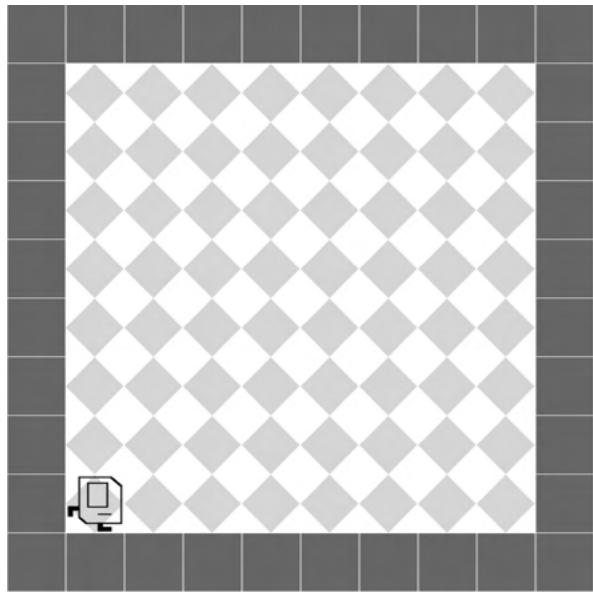
Maze



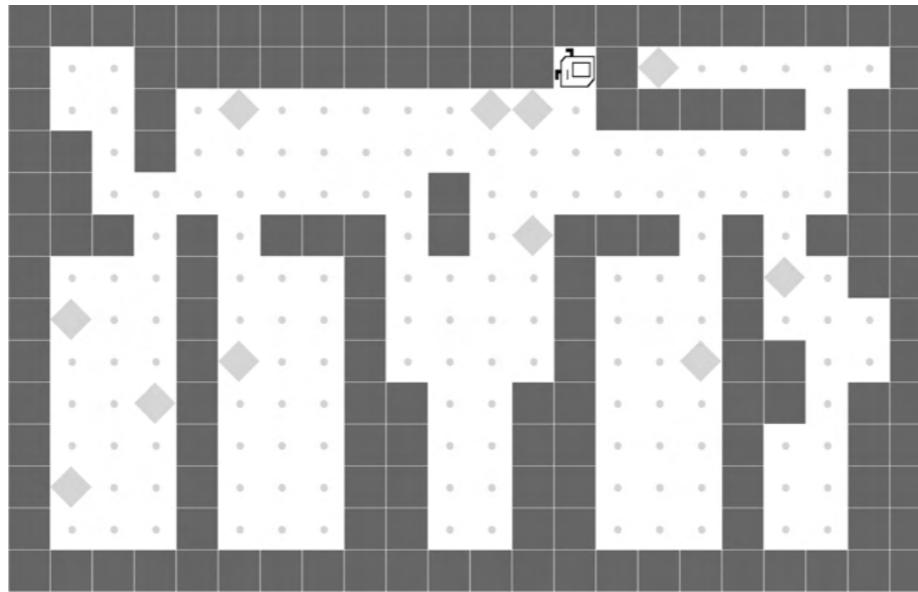
FourCorners



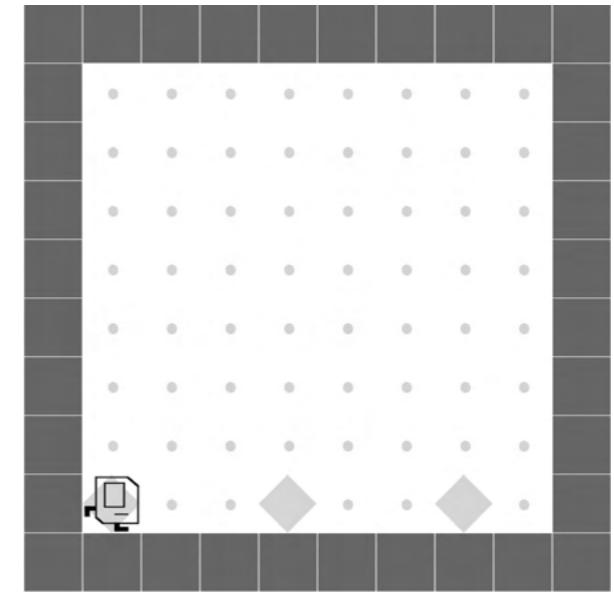
Harvester



CleanHouse



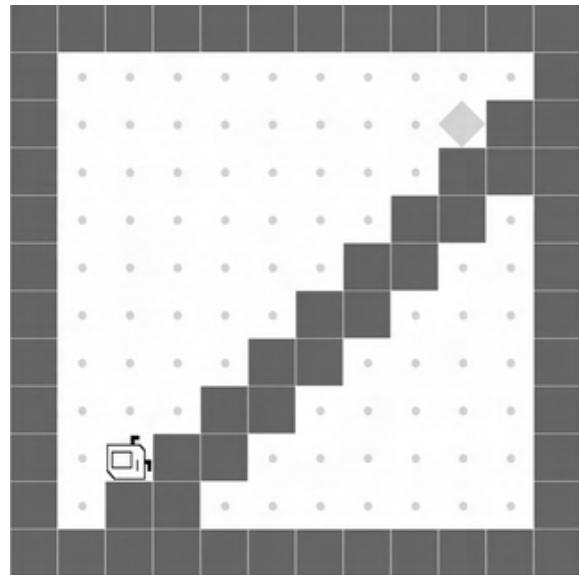
TopOff



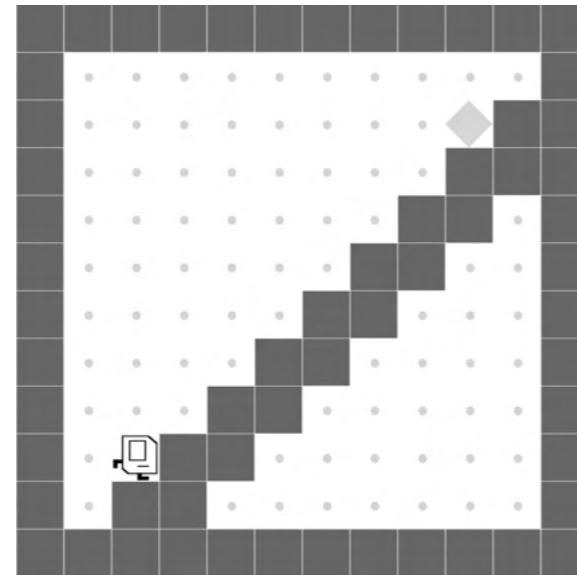
# Qualitative Results

---

StairClimber

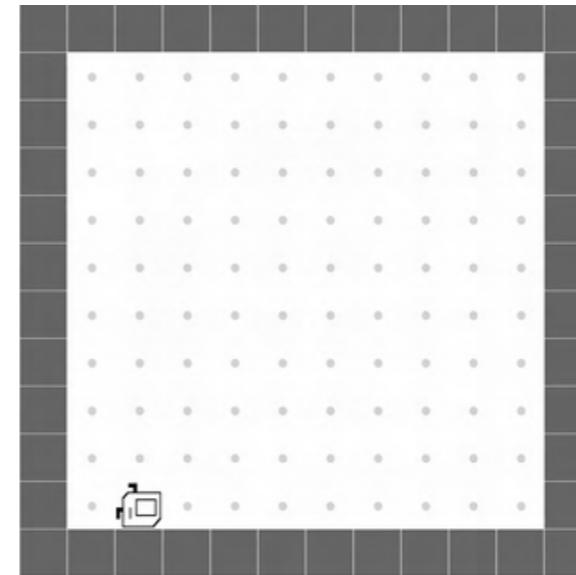


DRL

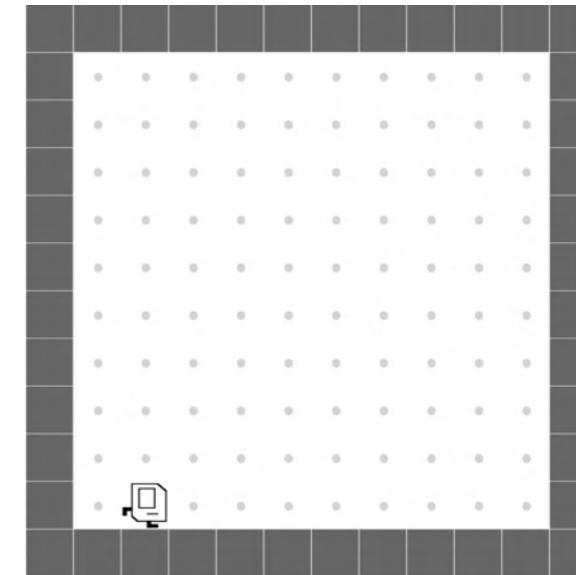


LEAPS

FourCorners

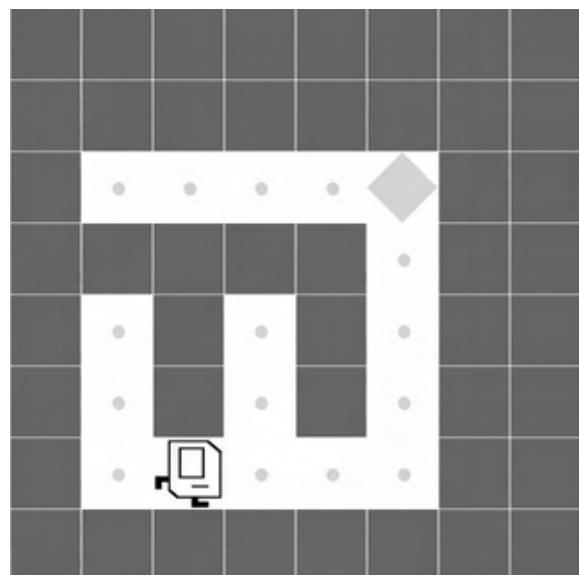


DRL

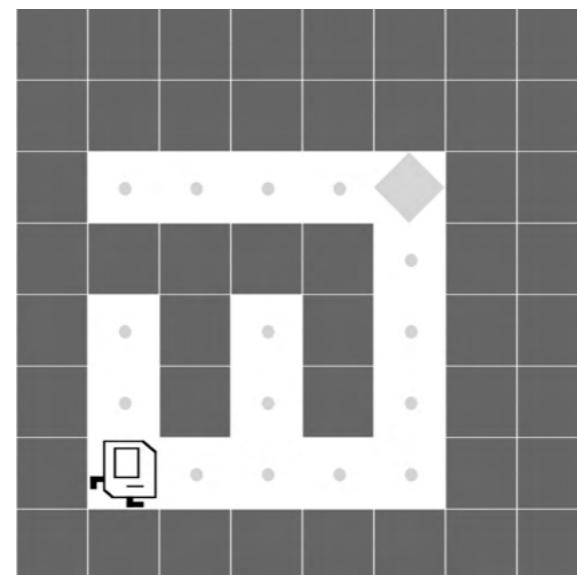


LEAPS

Maze

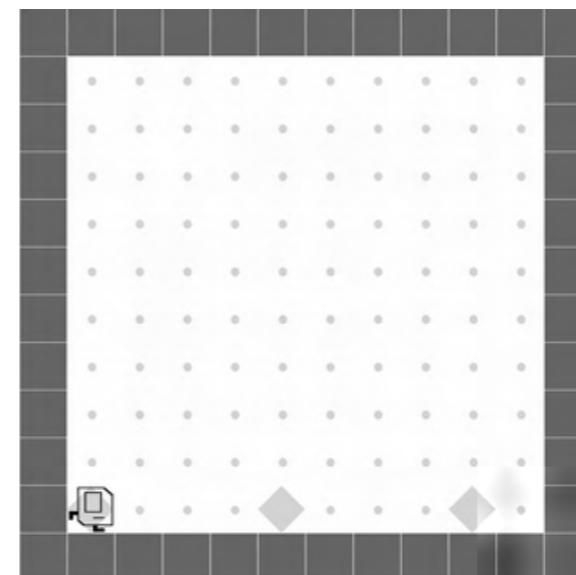


DRL

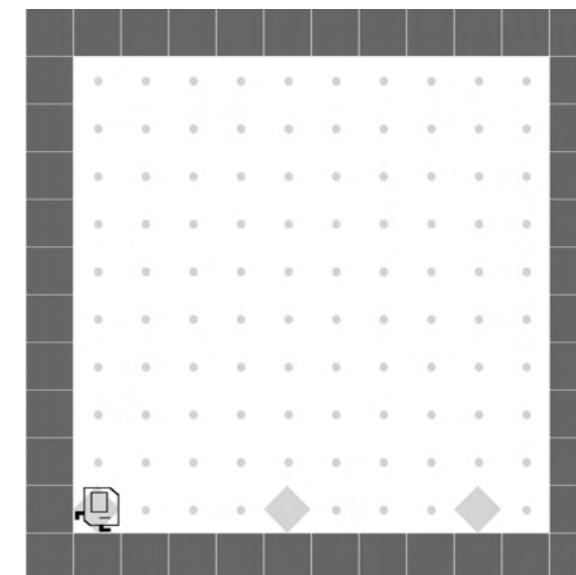


LEAPS

TopOff

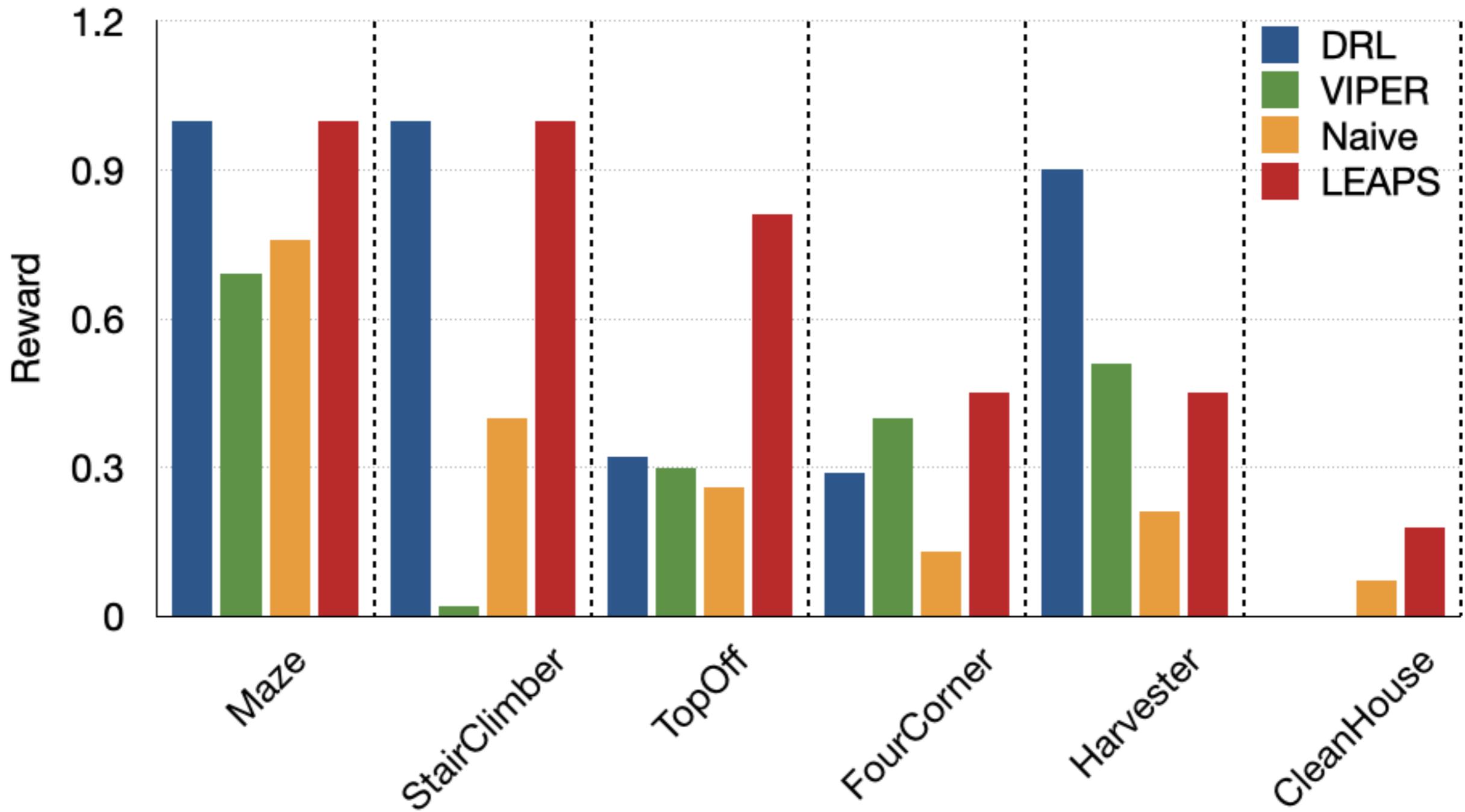


DRL



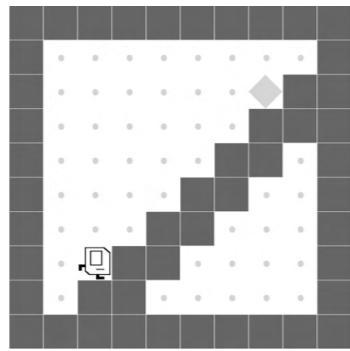
LEAPS

# Quantitative Results



# Zero-shot Generalization

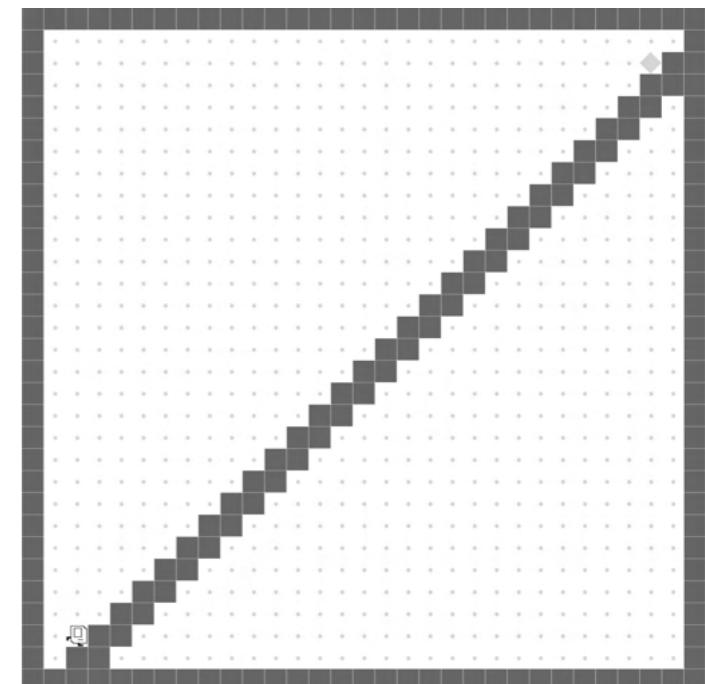
Learning on  
8x8 grids



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

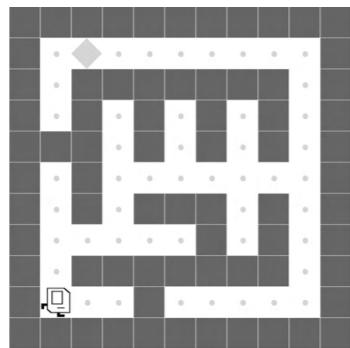


Evaluation on  
100x100 grids

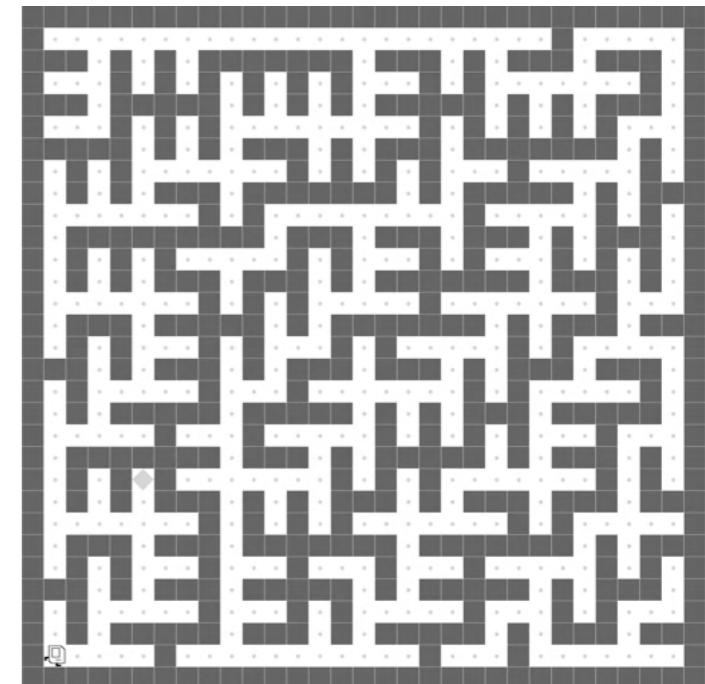


StairClimber

Maze

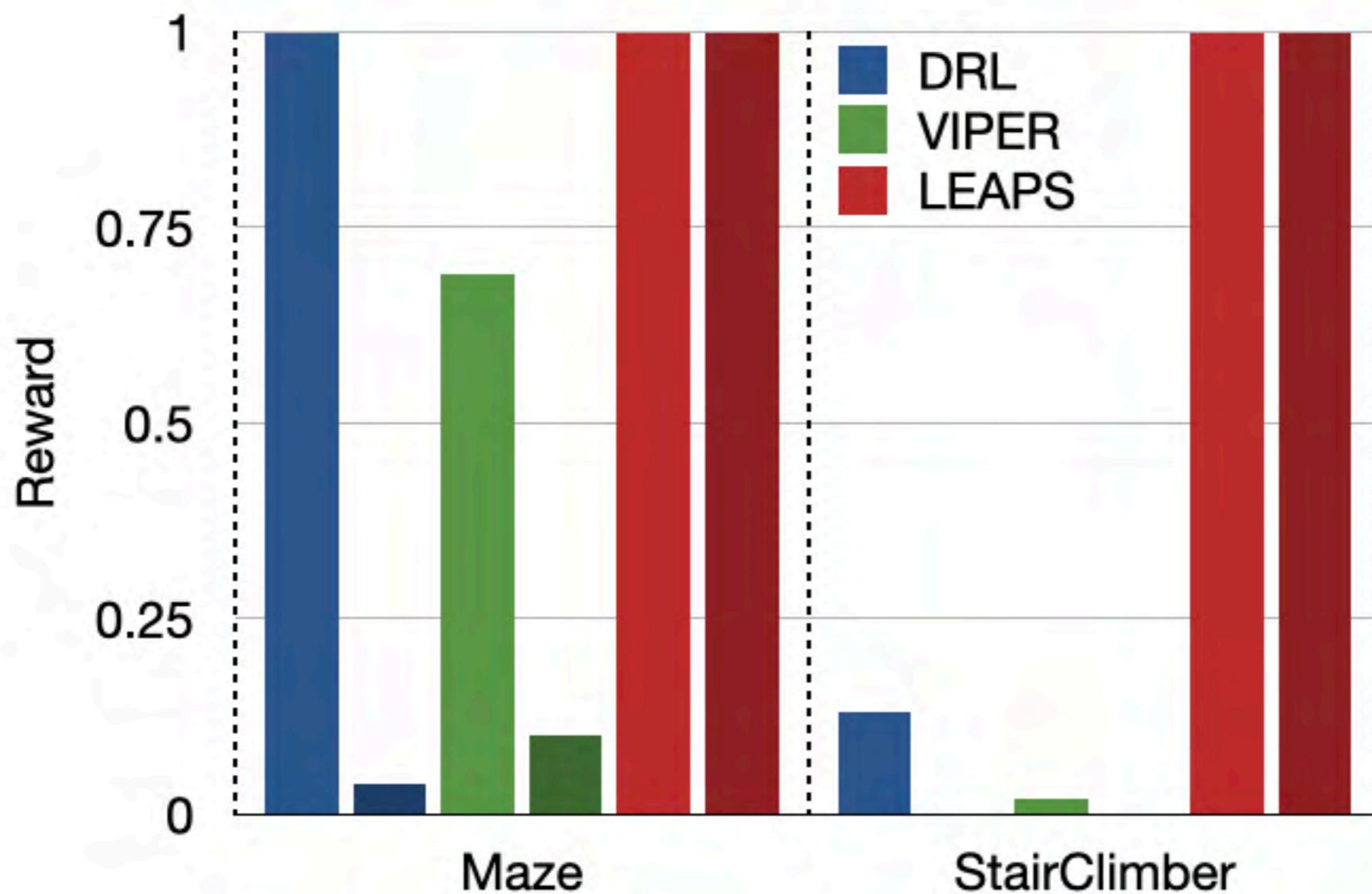


```
DEF run()
  IF frontIsClear()
    turnLeft
  WHILE noMarkersPresent()
    turnRight
    move
```



# Zero-shot Generalization

---

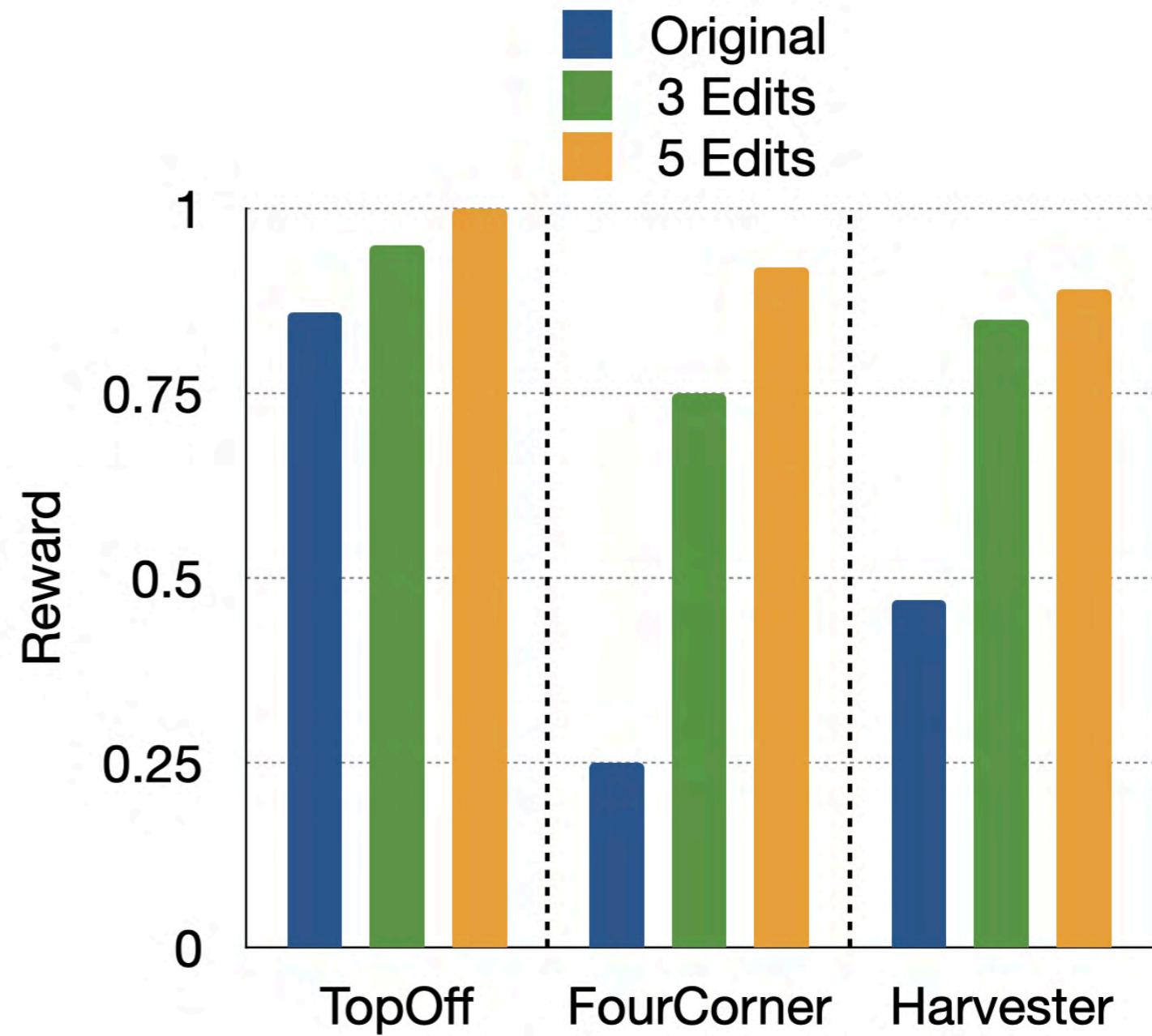


# Interpretability

## Human Debugging Interface



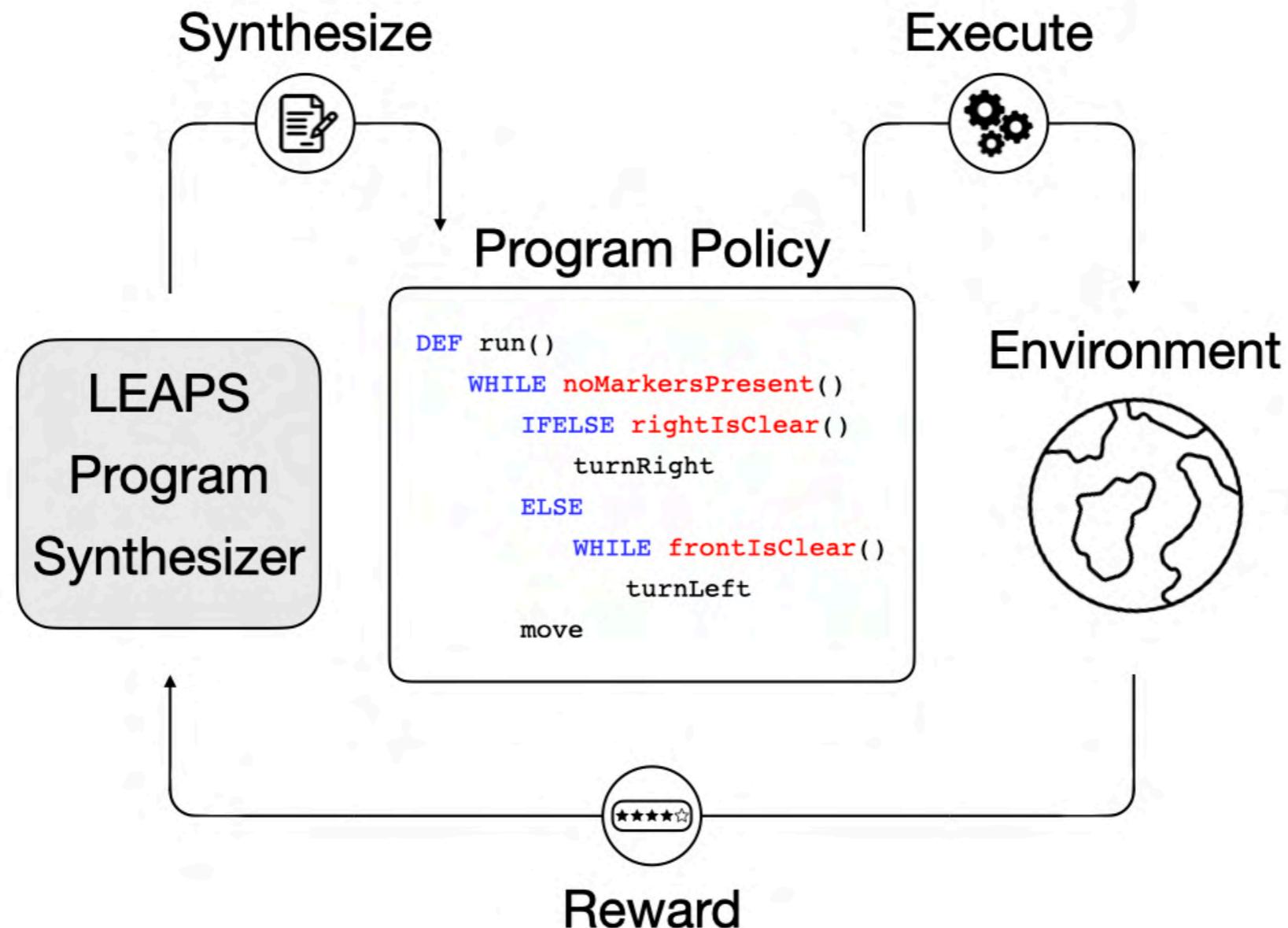
## Improved Performance



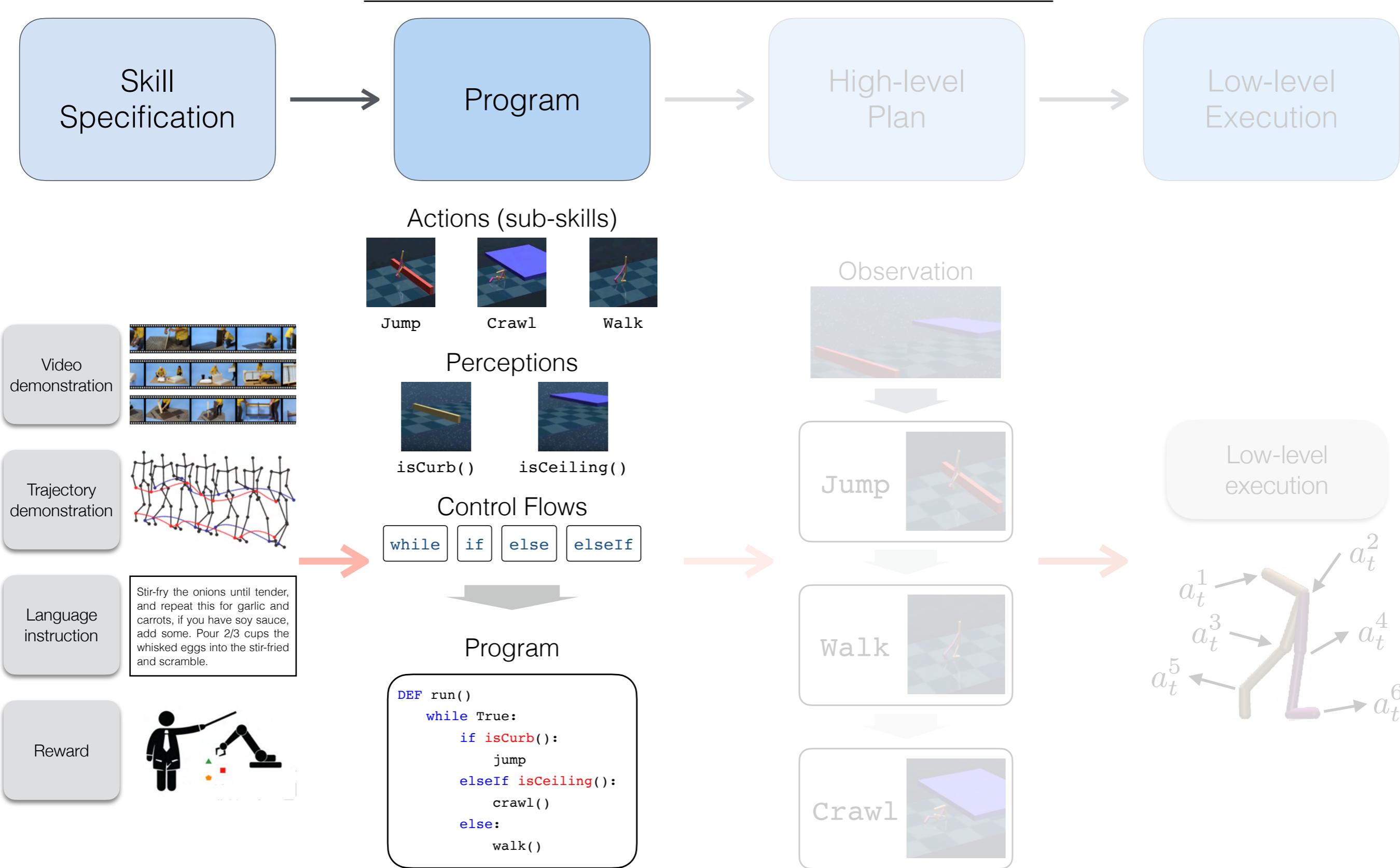
# Takeaway

---

- Synthesize generalizable and interpretable programs from rewards

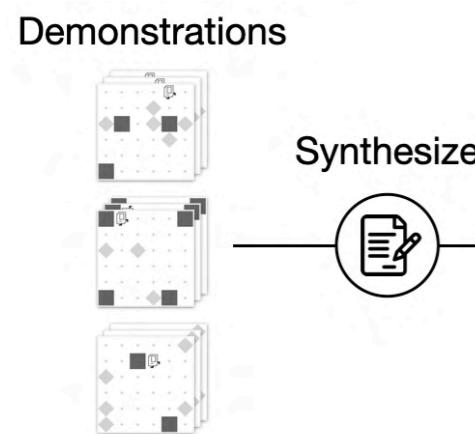


# Program Inference

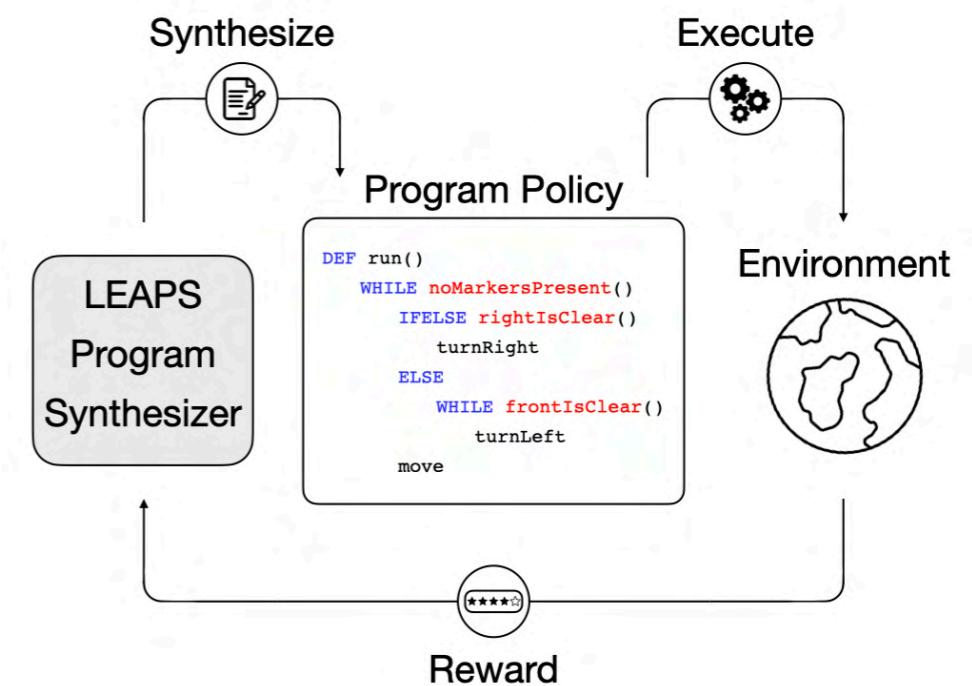


# Program Inference

Imitation learning from demonstrations



Reinforcement learning from rewards



# Program Inference

Imitation learning from demonstrations

Demonstrations



Synthesize



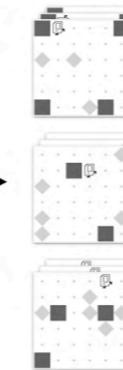
Program

```
DEF run()
  if isFrontClear():
    move
  else:
    turnLeft
    move
    turnLeft
  repeat(2):
    turnRight
    putMarker
```

Execute



Execution



Reinforcement learning from rewards

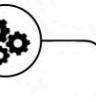
Synthesize



Program Policy

```
DEF run()
  WHILE noMarkersPresent()
    IFELSE rightIsClear()
      turnRight
    ELSE
      WHILE frontIsClear()
        turnLeft
      move
```

Execute



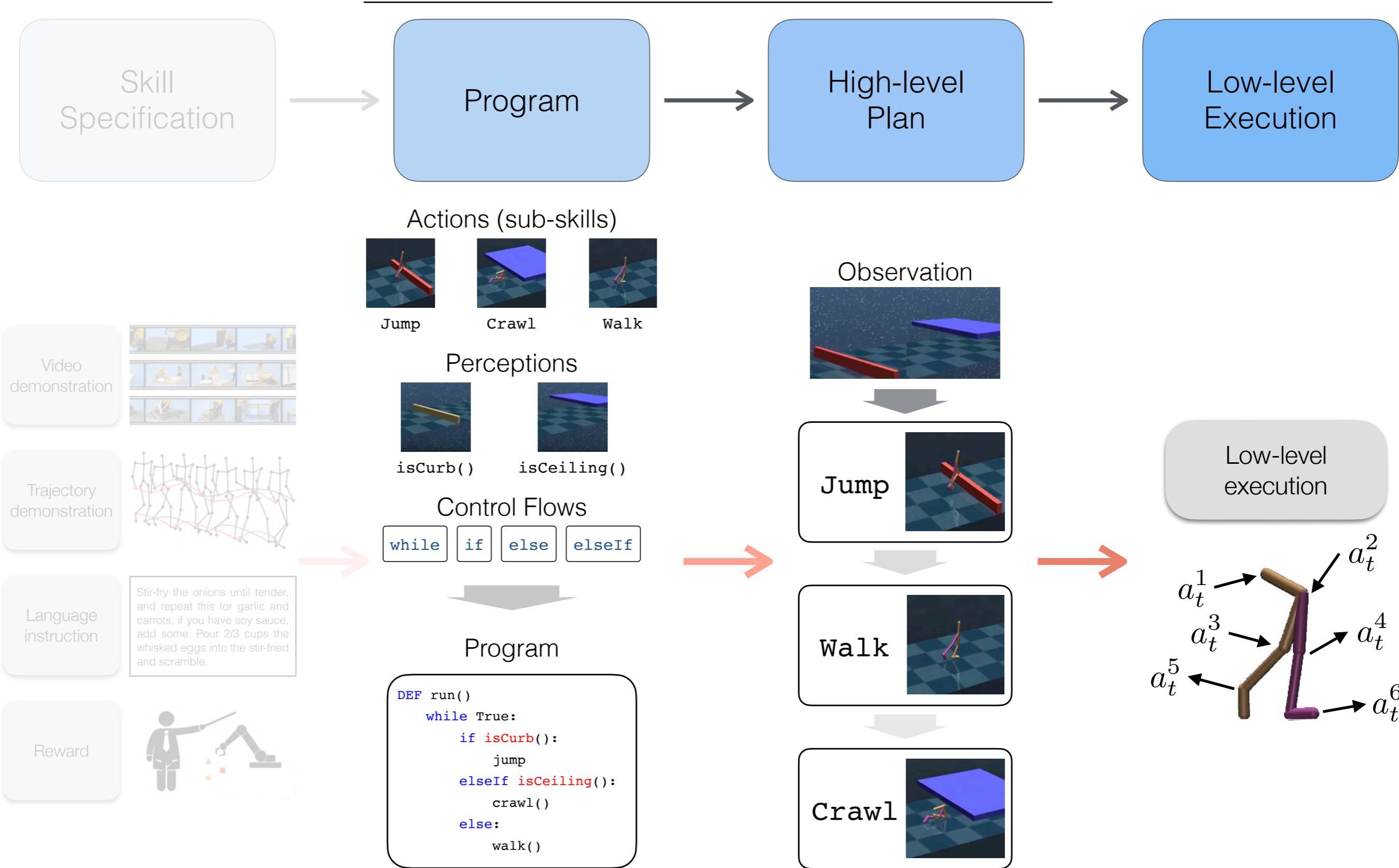
Environment



Reward

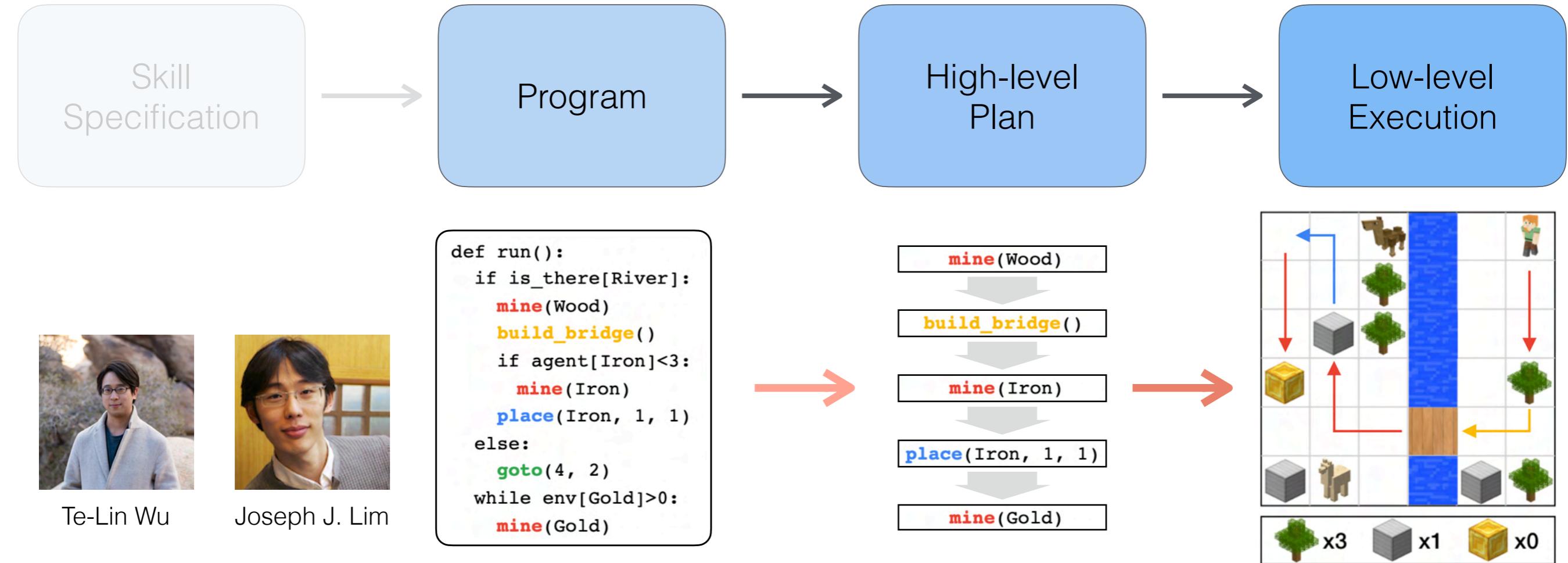
LEAPS  
Program  
Synthesizer

# Task Execution



# Program Guided Agent

ICLR 2020 (Spotlight)



Te-Lin Wu



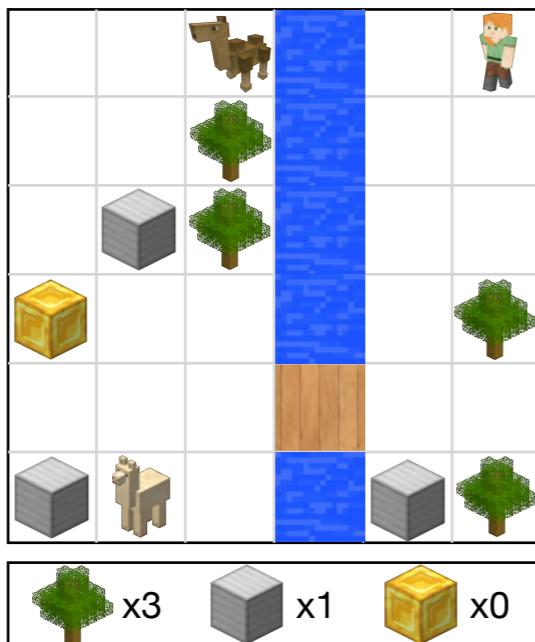
Joseph J. Lim

# Problem Formulation

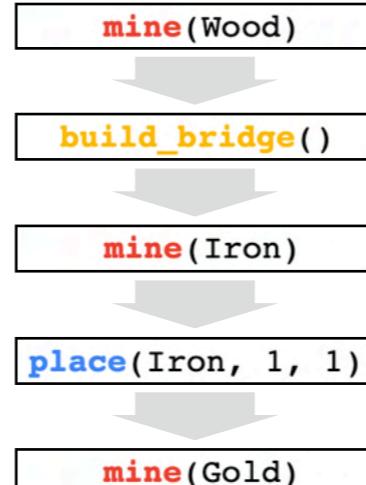
Program  
(task)

```
def run():
    if is_there[River]:
        mine(Wood)
        build_bridge()
    if agent[Iron]<3:
        mine(Iron)
        place(Iron, 1, 1)
    else:
        goto(4, 2)
    while env[Gold]>0:
        mine(Gold)
```

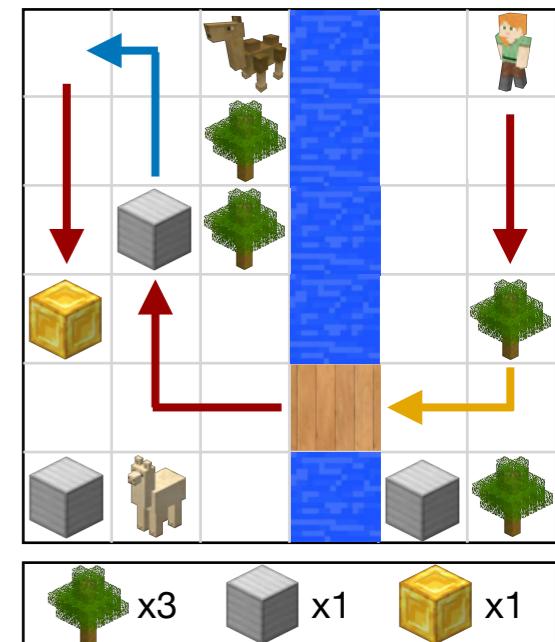
Observation



Plan  
(subtasks)



Execution



# Instructions

---

## Programs

```
def run()
    if is_there[River]:
        mine(Wood)
        build_bridge()
        if agent[Iron] < 3:
            mine(Iron)
            place(Iron, 2, 3)
        else:
            goto(4, 2)
    while env[Gold] > 0:
        mine(Gold)
```

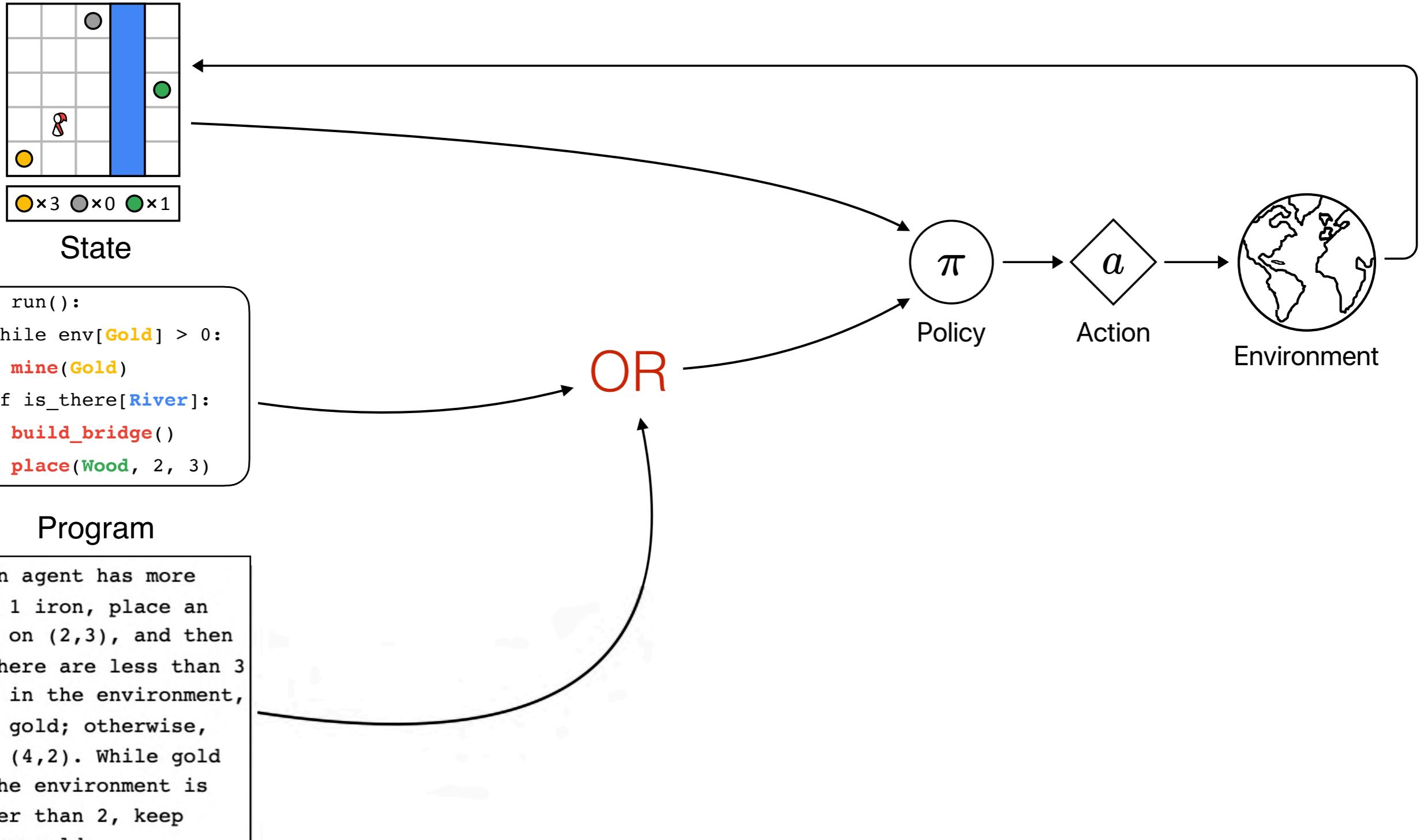
```
def run()
    while agent[Wood] <= 11:
        place(Wood, 2, 4)
        place(Iron, 1, 1)
        place(Iron, 8, 5)
        mine(Gold)
        mine(Gold)
        mine(Gold)
        repeat(4):
            sell(Gold)
            sell(Iron)
```

## Natural Language Descriptions

If a river is in the environment, mine a wood and then use it to build a bridge. And then if agent has less than there iron, place an iron at (2,3). Otherwise if no river, goto location (4,2). Finally, whenever there's still gold in the environment, mine a gold.

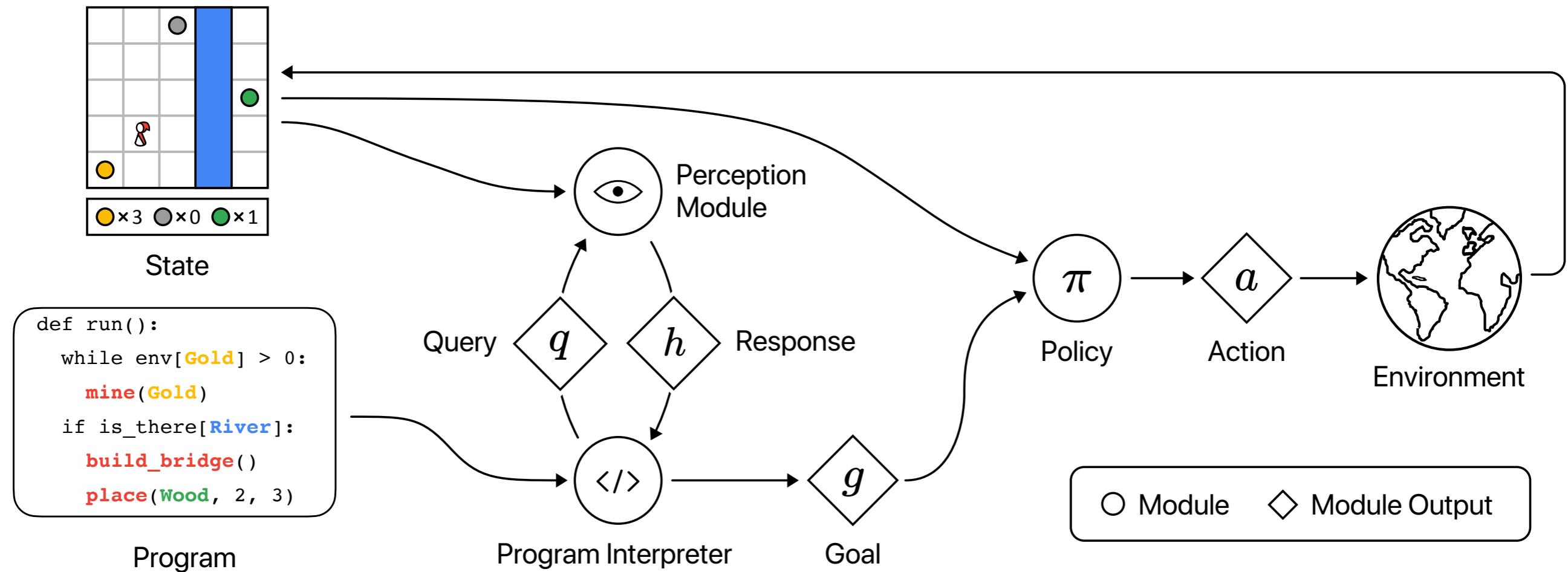
While agent has no more than 11 wood, place wood at (2,4) and iron at (1,1), then place iron at (8,5) and mine gold twice, then mine gold. After the preceding procedure, sell gold and sell iron 4 times.

# End-to-end Learning Baseline



Language description

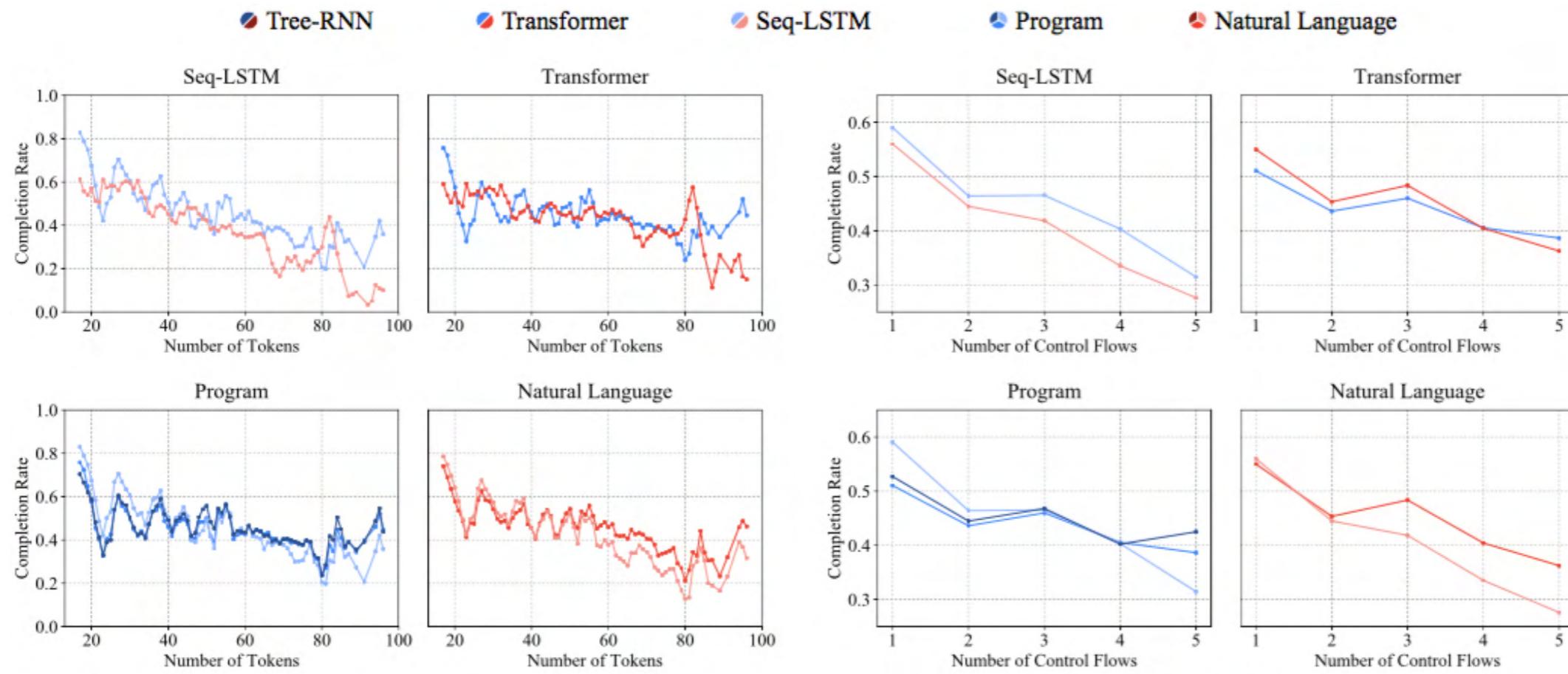
# Program Guided Agent



# Quantitative Results

## Generalization

Instruction Method		Natural language descriptions		Programs				
		Seq-LSTM	Transformer	Seq-LSTM	Tree-RNN	Transformer	Ours (concat)	Ours
<b>Dataset</b>	test	54.9±1.8%	52.5±2.6%	56.7±1.9%	50.1±1.2%	49.4±1.6%	88.6±0.8%	94.0±0.5%
	test-complex	32.4±4.9%	38.2±2.6%	38.8±1.2%	42.2±2.4%	40.9±1.5%	85.2±0.8%	91.8±0.2%
<b>Generalization gap</b>		40.9%	27.2%	31.6%	15.8%	17.2%	3.8%	2.3%



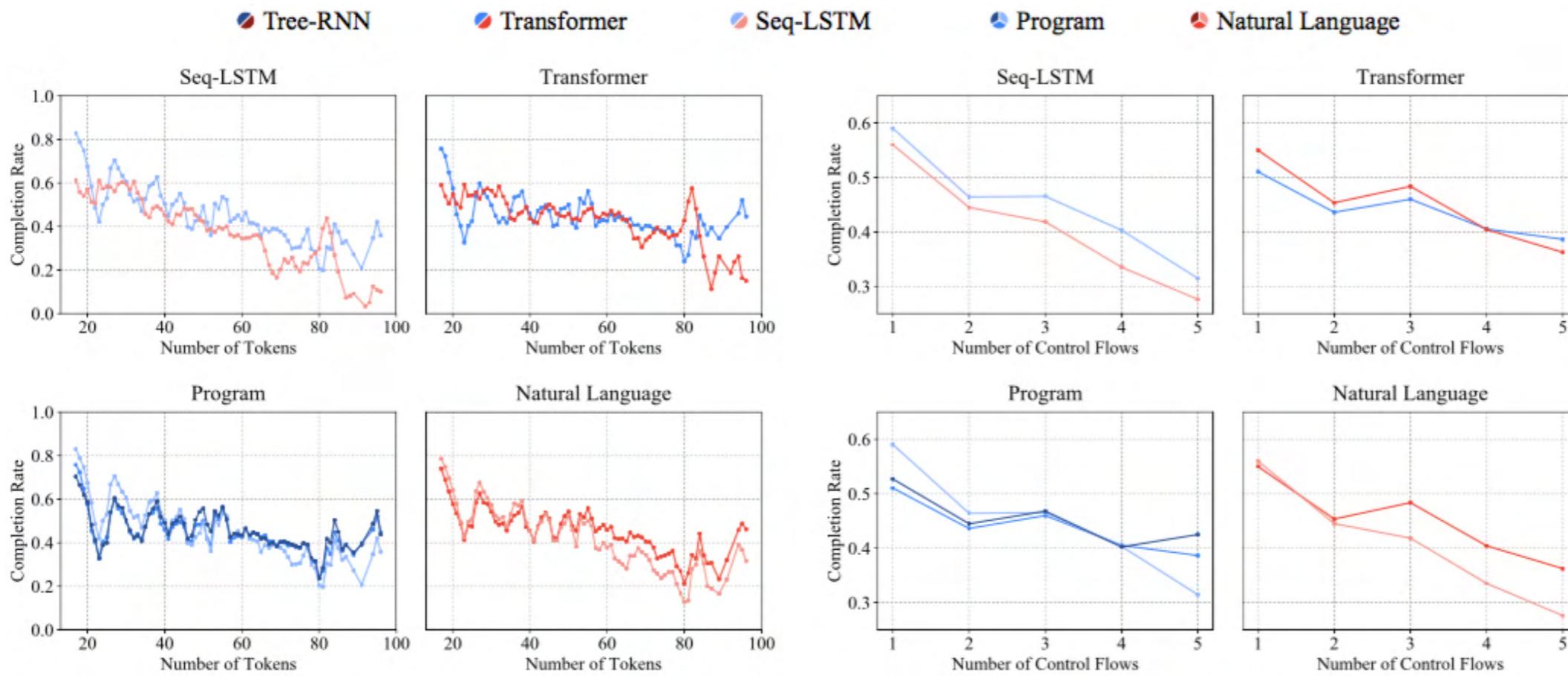
(a) Instruction Length

(b) Instruction Complexity

# Quantitative Results

## Natural Languages < Programs

Instruction Method		Natural language descriptions		Programs				
		Seq-LSTM	Transformer	Seq-LSTM	Tree-RNN	Transformer	Ours (concat)	Ours
Dataset	test	54.9±1.8%	52.5±2.6%	56.7±1.9%	50.1±1.2%	49.4±1.6%	88.6±0.8%	94.0±0.5%
	test-complex	32.4±4.9%	38.2±2.6%	38.8±1.2%	42.2±2.4%	40.9±1.5%	85.2±0.8%	91.8±0.2%
Generalization gap		40.9%	27.2%	31.6%	15.8%	17.2%	3.8%	2.3%



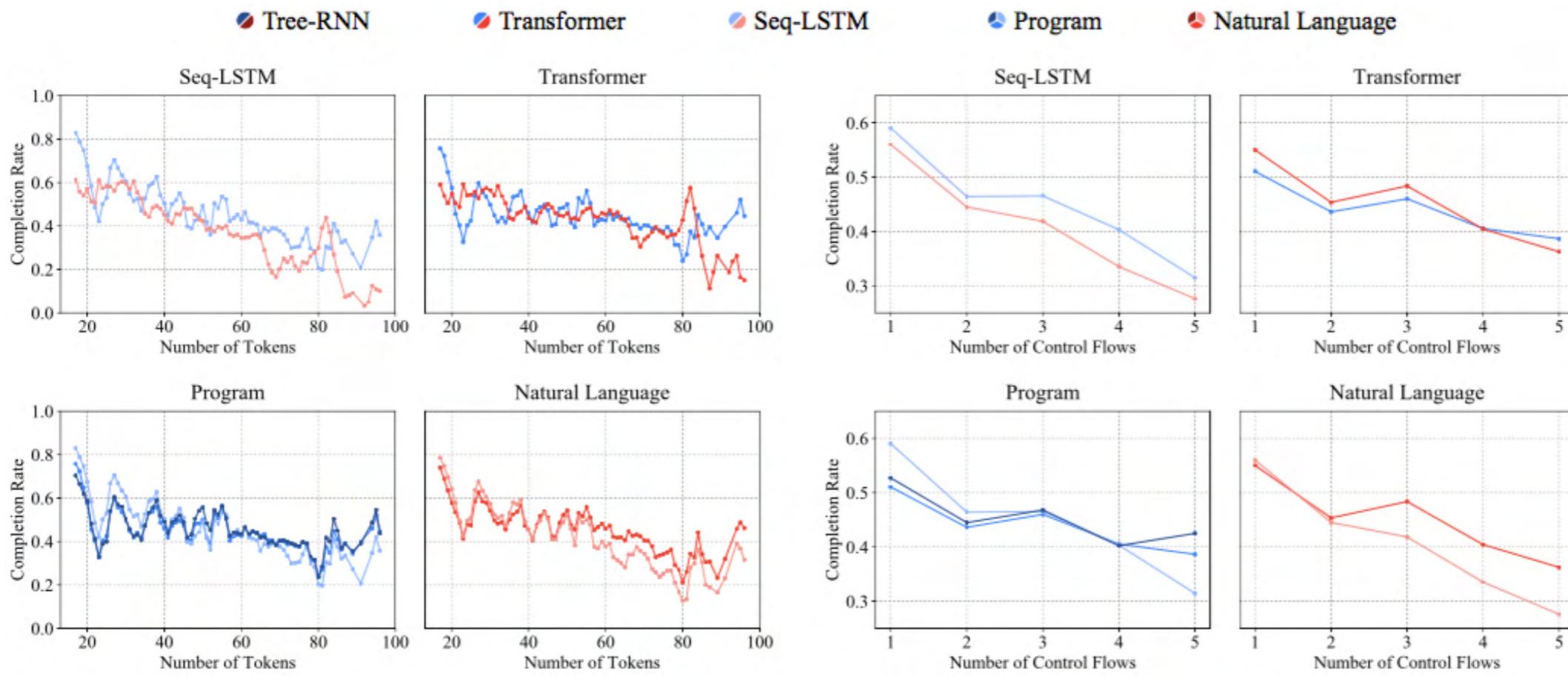
(a) Instruction Length

(b) Instruction Complexity

# Quantitative Results

End-to-end < Ours (modular)

Instruction Method		Natural language descriptions		Programs				
		Seq-LSTM	Transformer	Seq-LSTM	Tree-RNN	Transformer	Ours (concat)	Ours
<b>Dataset</b>	test	54.9±1.8%	52.5±2.6%	56.7±1.9%	50.1±1.2%	49.4±1.6%	88.6±0.8%	94.0±0.5%
	test-complex	32.4±4.9%	38.2±2.6%	38.8±1.2%	42.2±2.4%	40.9±1.5%	85.2±0.8%	91.8±0.2%
<b>Generalization gap</b>		40.9%	27.2%	31.6%	15.8%	17.2%	3.8%	2.3%



(a) Instruction Length

(b) Instruction Complexity

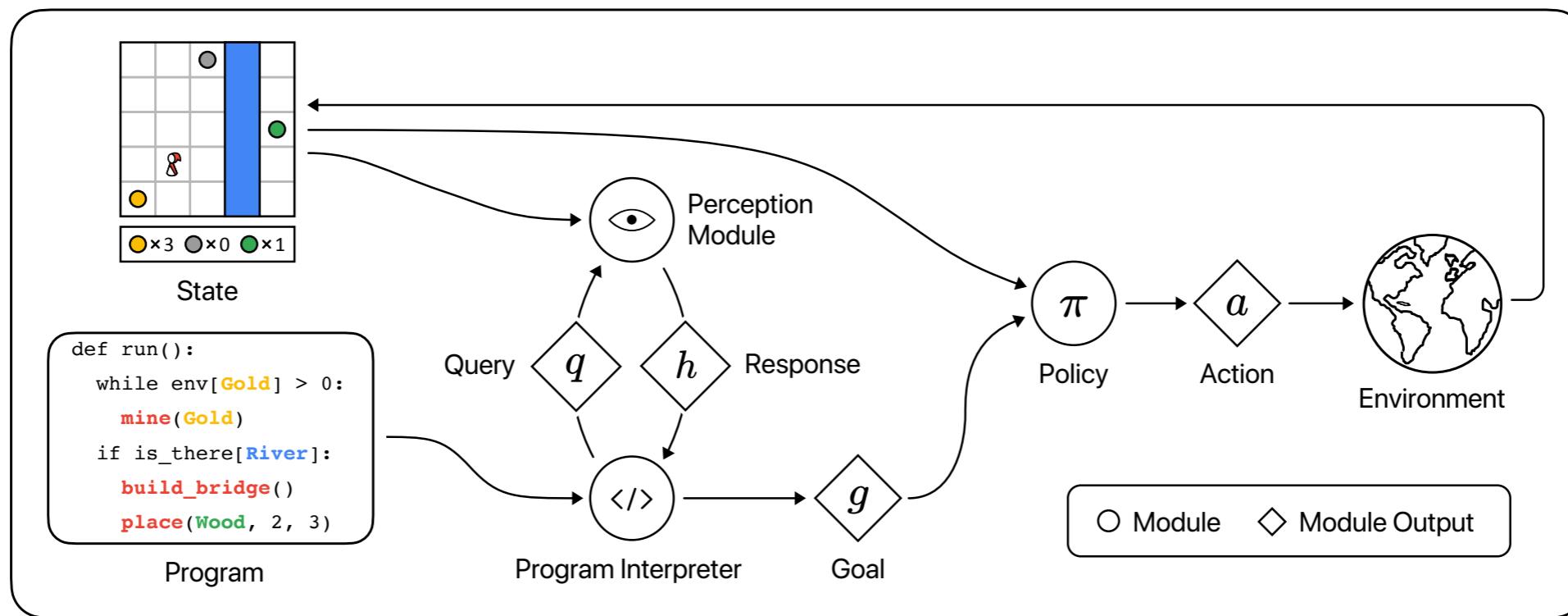
# Takeaway

- Specific tasks using programs

Program

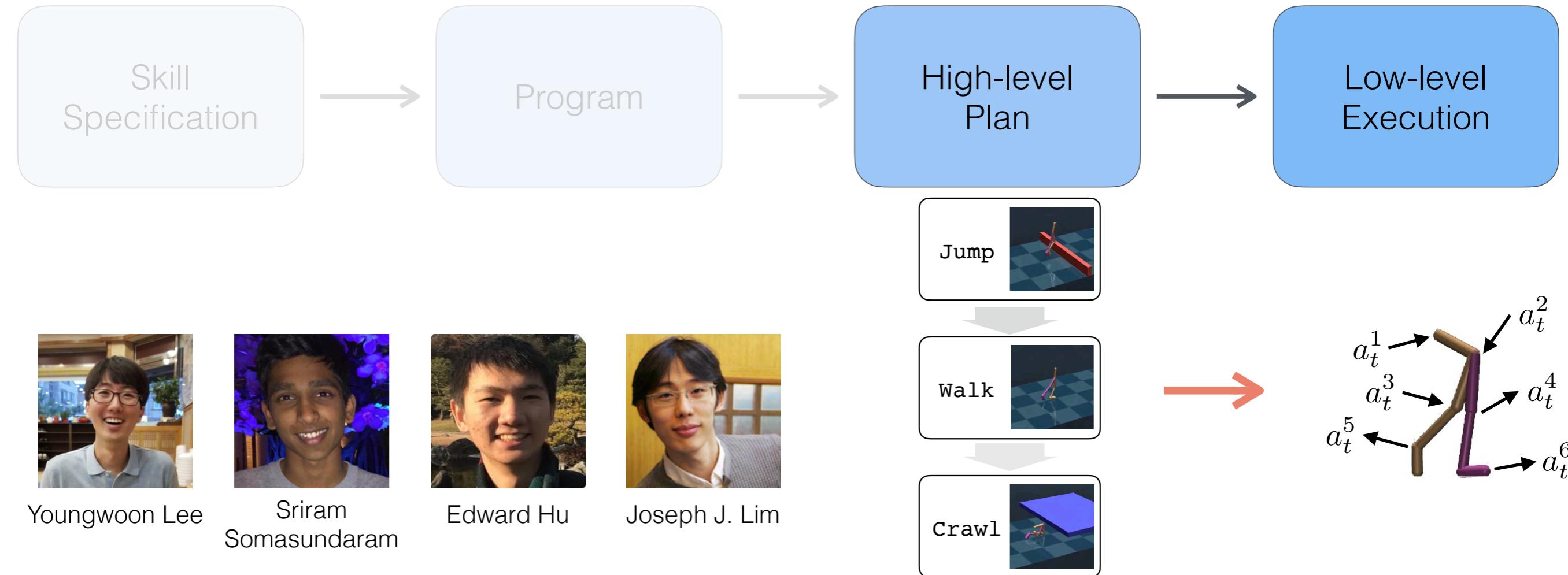
```
def run():
    if is_there[River]:
        mine(Wood)
        build_bridge()
        if agent[Iron]<3:
            mine(Iron)
            place(Iron, 1, 1)
        else:
            goto(4, 2)
    while env[Gold]>0:
        mine(Gold)
```

- Leverage the structure of programs with a modular framework



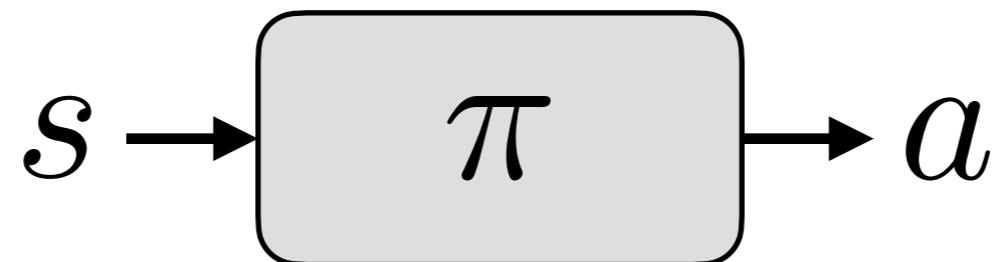
# Composing Complex Skills by Learning Transition Policies

ICLR 2019

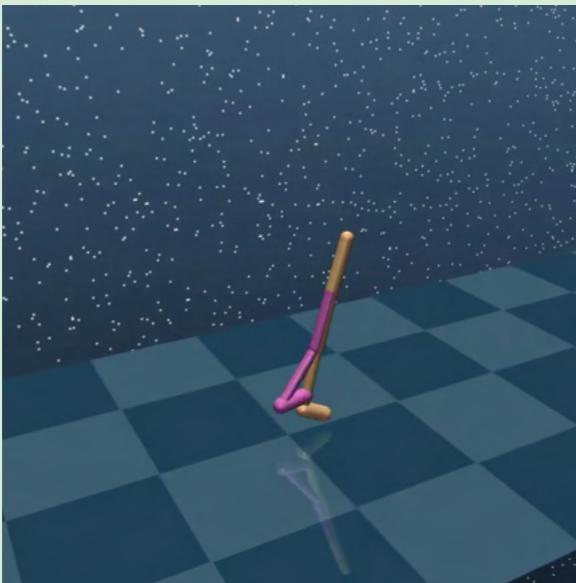


# Learned Skills

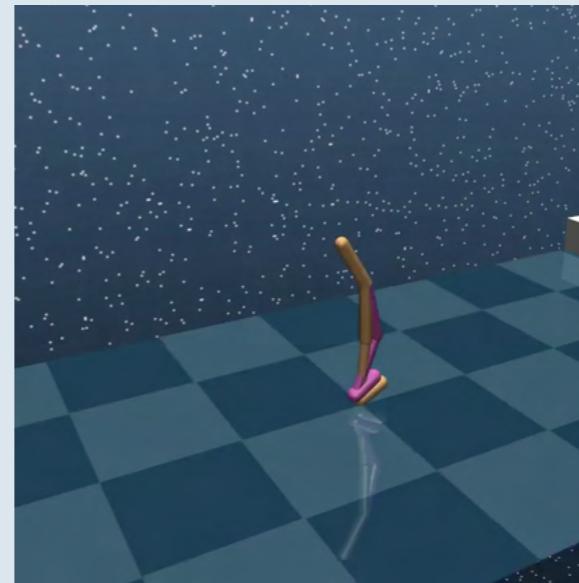
---



Walk



Jump



Crawl

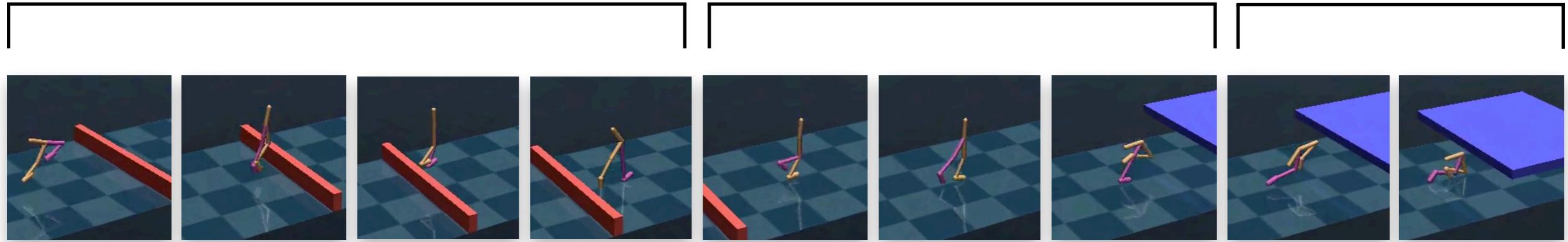


# Compose Complex Skills

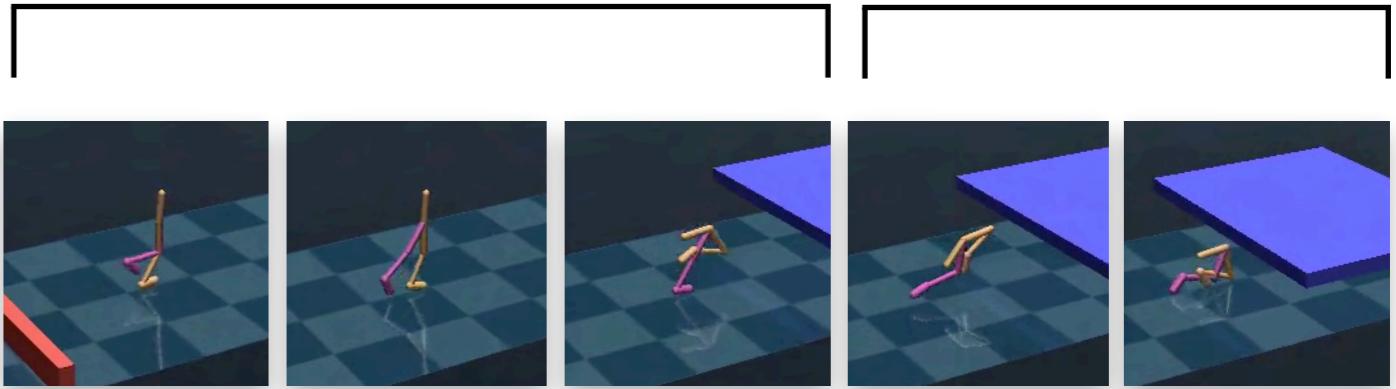
---

High-level plan

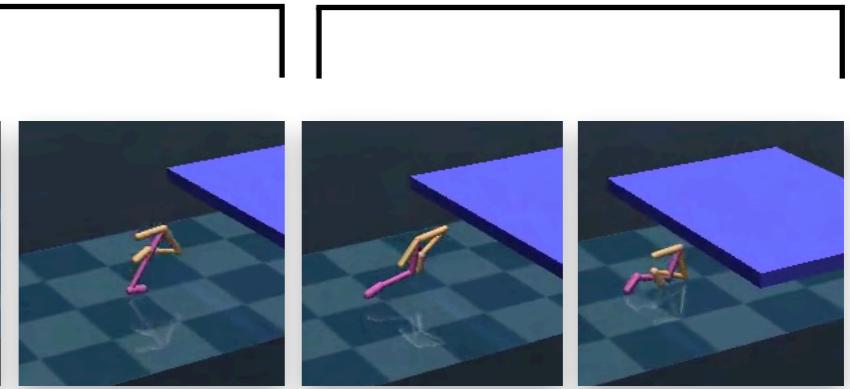
Jump



Walk



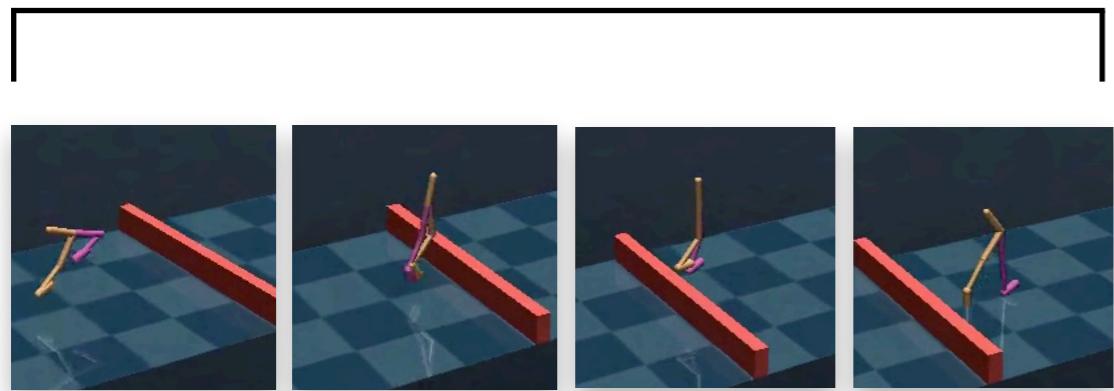
Crawl



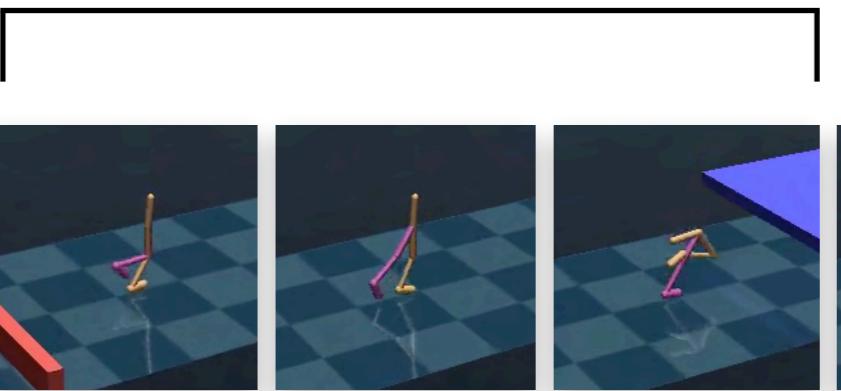
# Compose Complex Skills

High-level plan

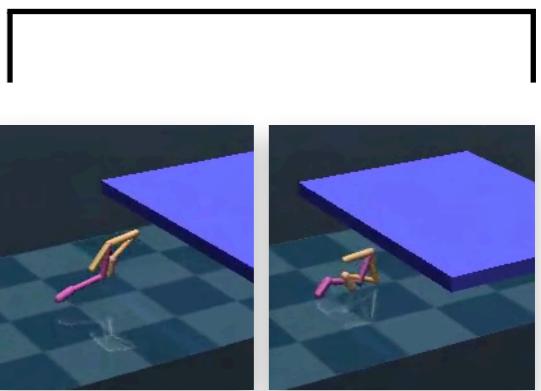
Jump



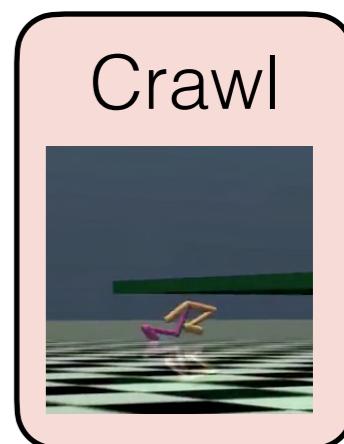
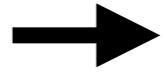
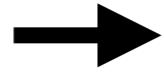
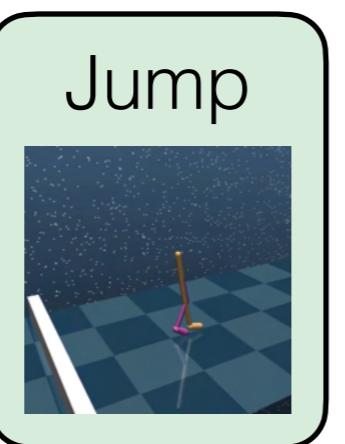
Walk



Crawl



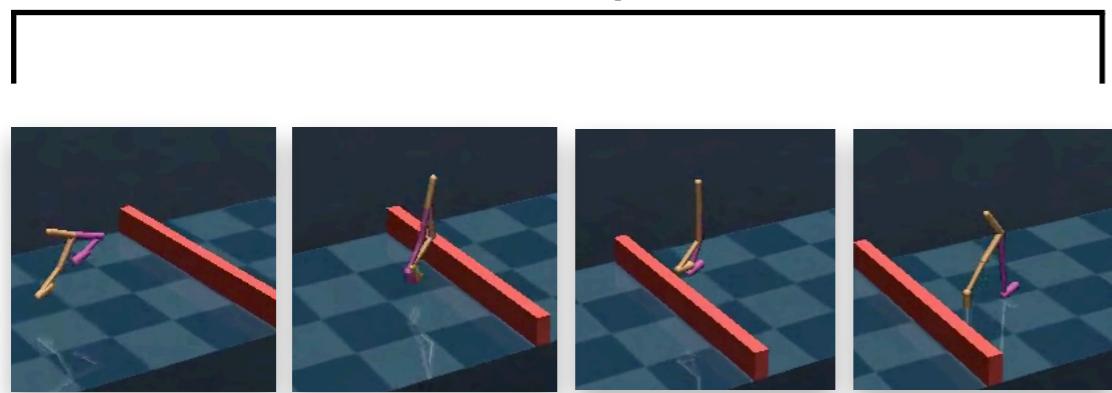
Sequentially execute corresponding policies



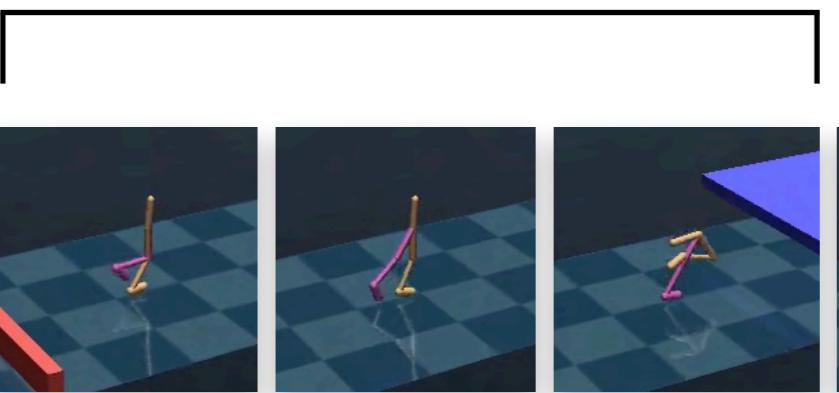
# Compose Complex Skills

High-level plan

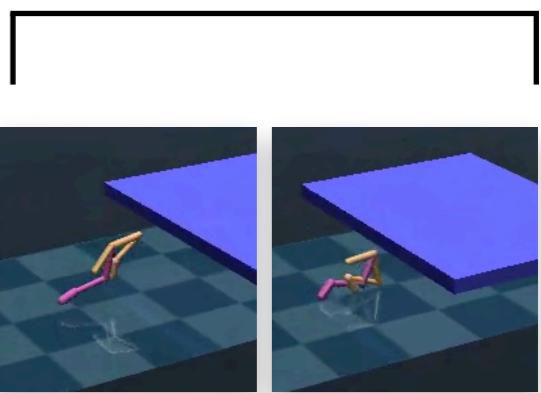
Jump



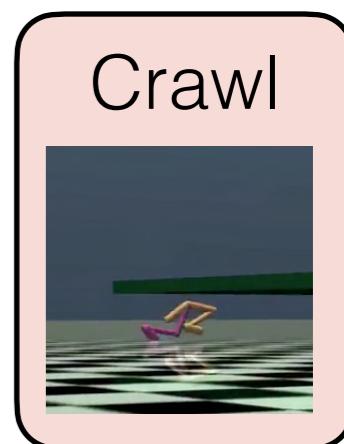
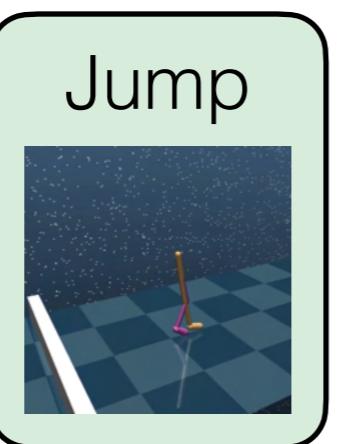
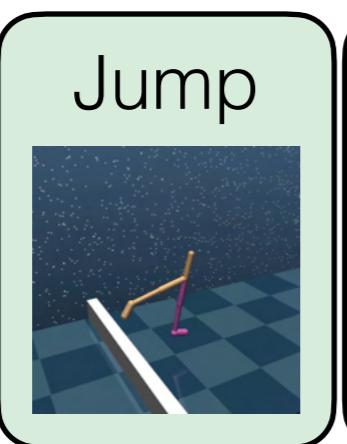
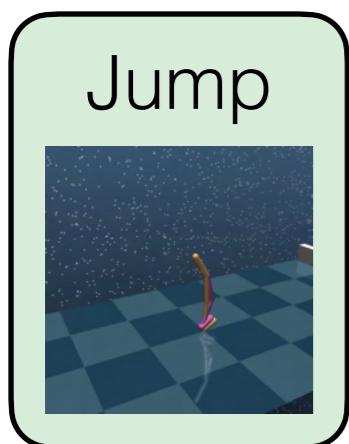
Walk



Crawl



Sequentially execute corresponding policies

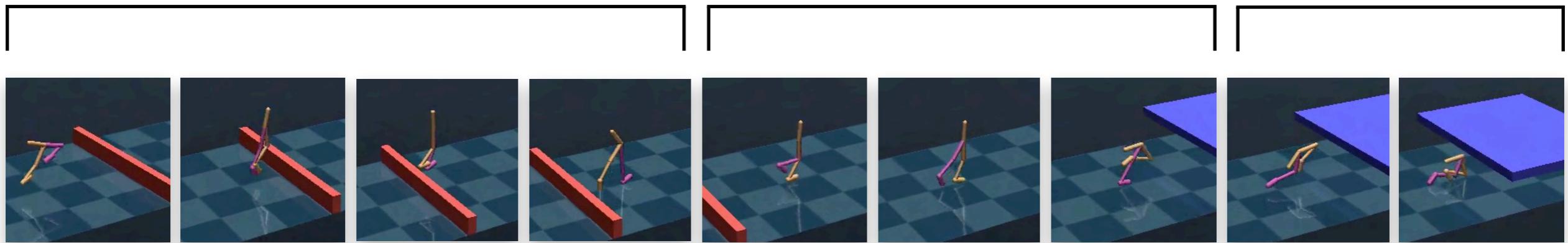


An end state of a previous policy might not be a good initial state of the following policy

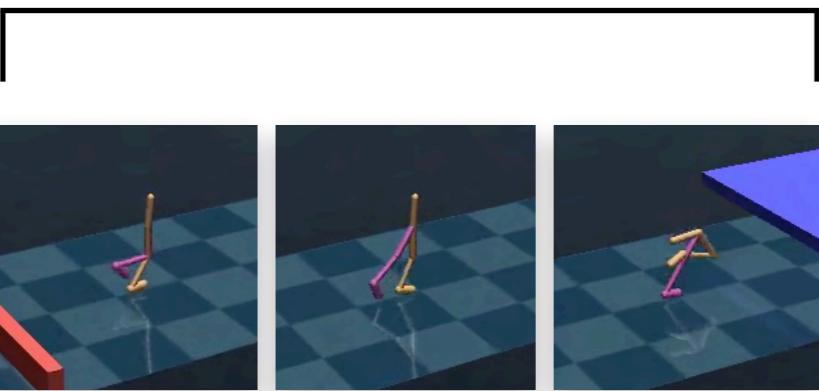
# Compose Complex Skills

High-level plan

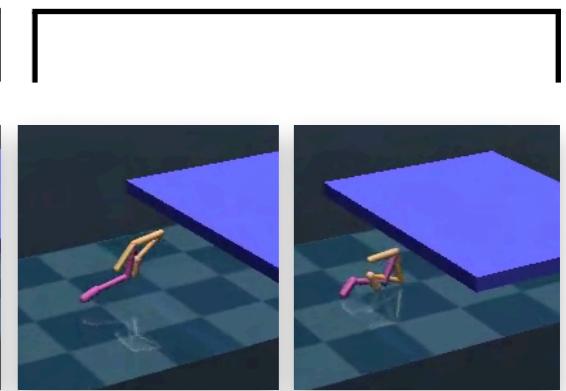
Jump



Walk

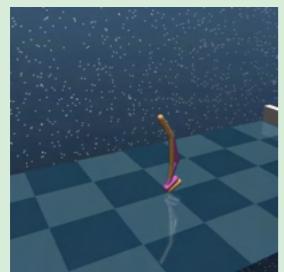


Crawl



Sequentially execute corresponding policies

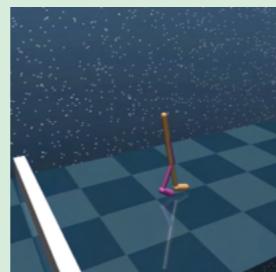
Jump



Jump

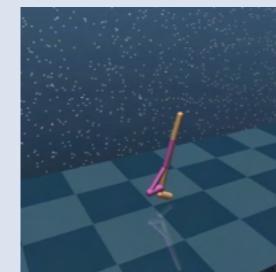


Jump



Trans  
 $\pi$

Walk



Walk



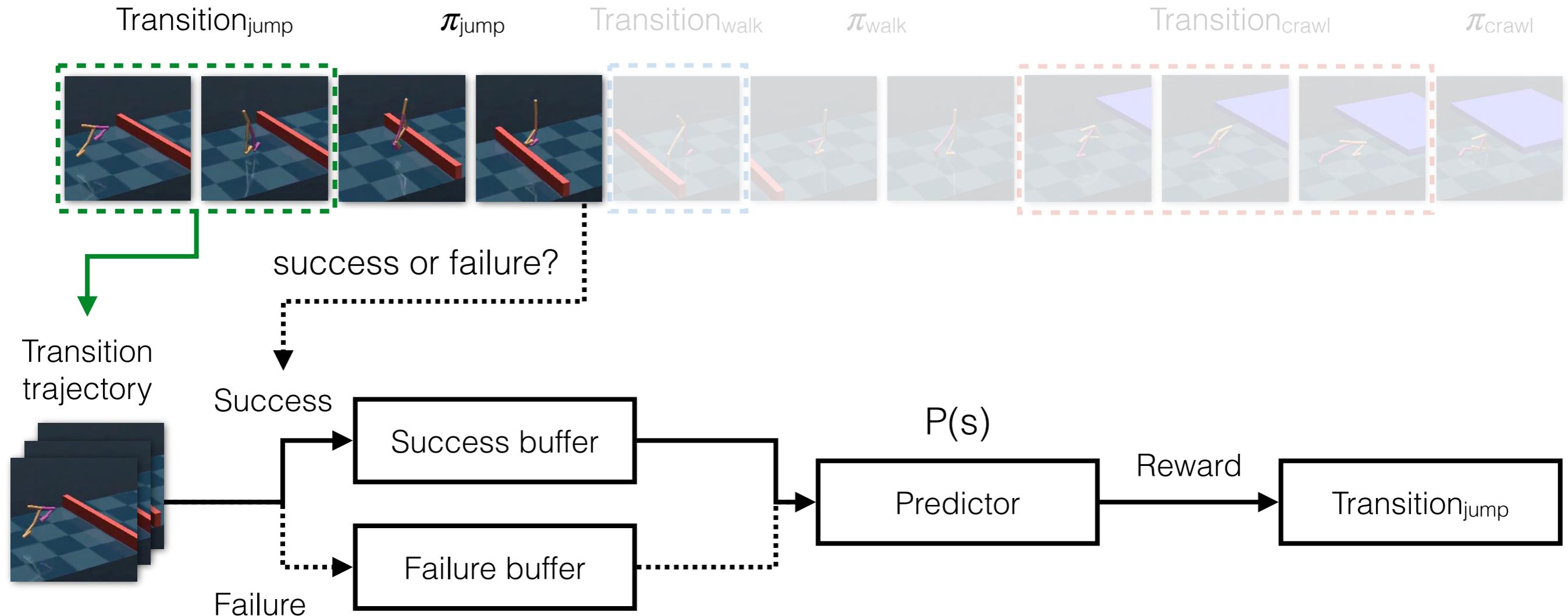
Trans  
 $\pi$

Crawl



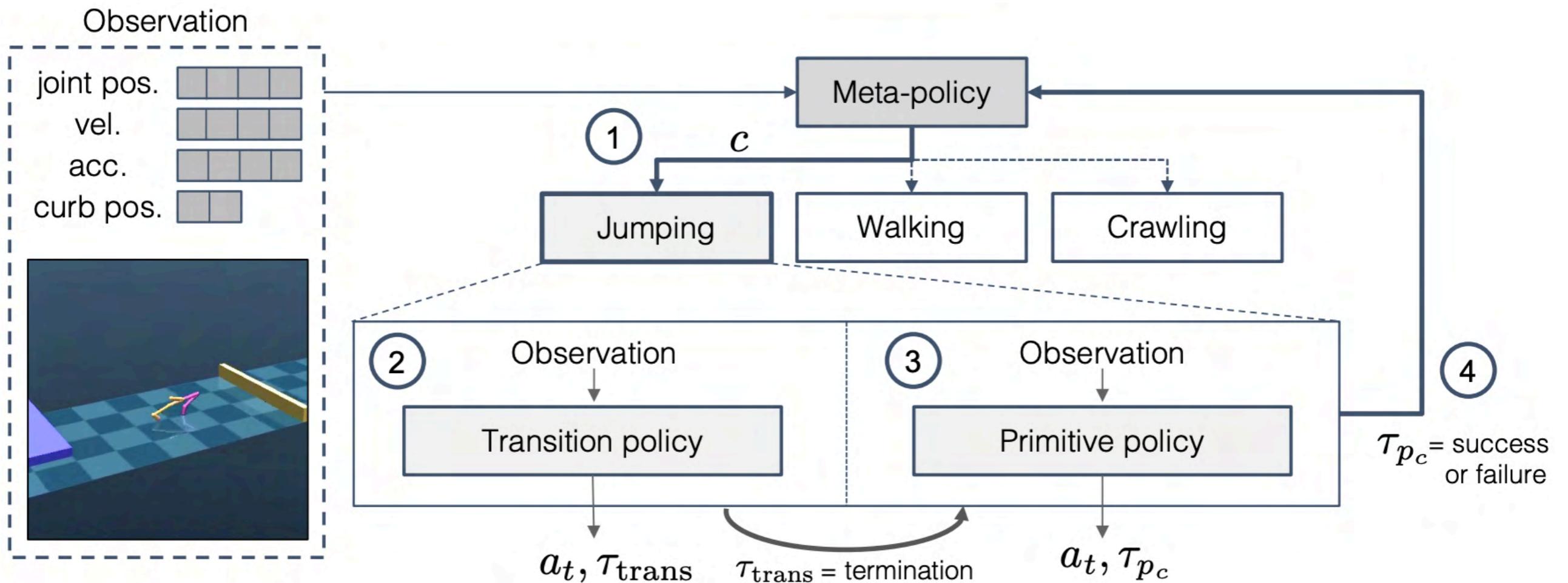
Transition policies

# Learning Transition Policies



- Predictor learns to judge if a state is good for executing the next policy
- Transition policy learns from the predicted rewards

# Modular Framework

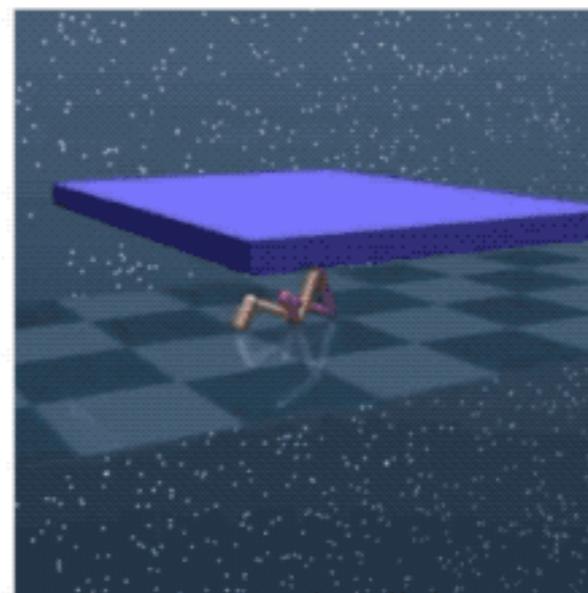


# Qualitative Results

---

Locomotion

Crawl



*Transition*

Walk

Walk Forward

*Transition*

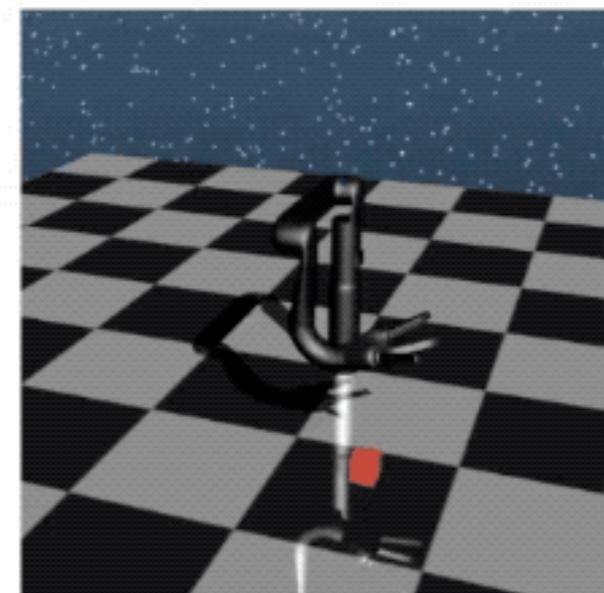
Walk Backward

Manipulation

Pick

*Transition*

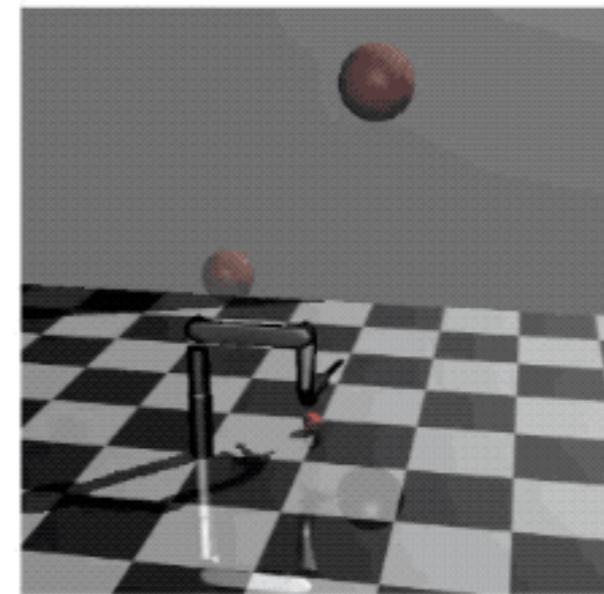
Pick



Toss

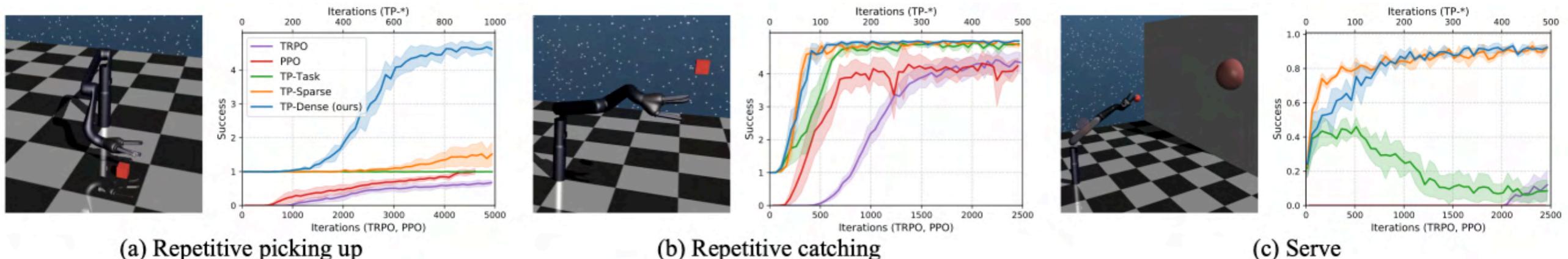
*Transition*

Hit

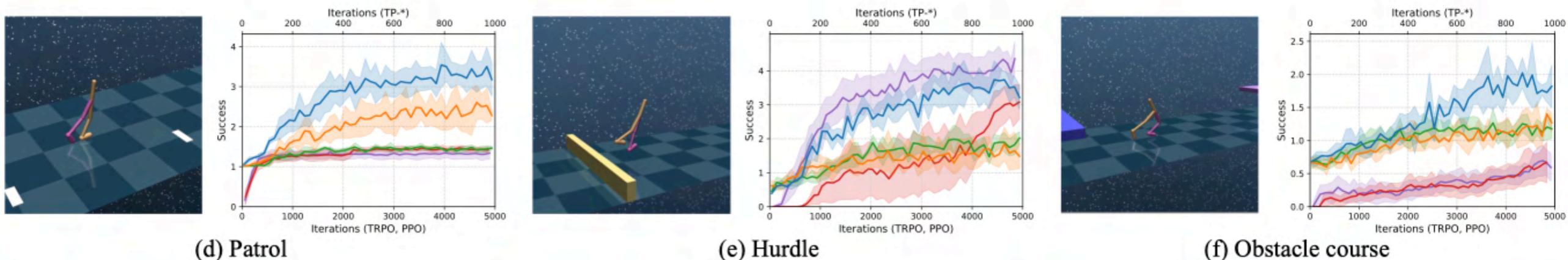


# Quantitative Results - Sample Efficiency

## Manipulation



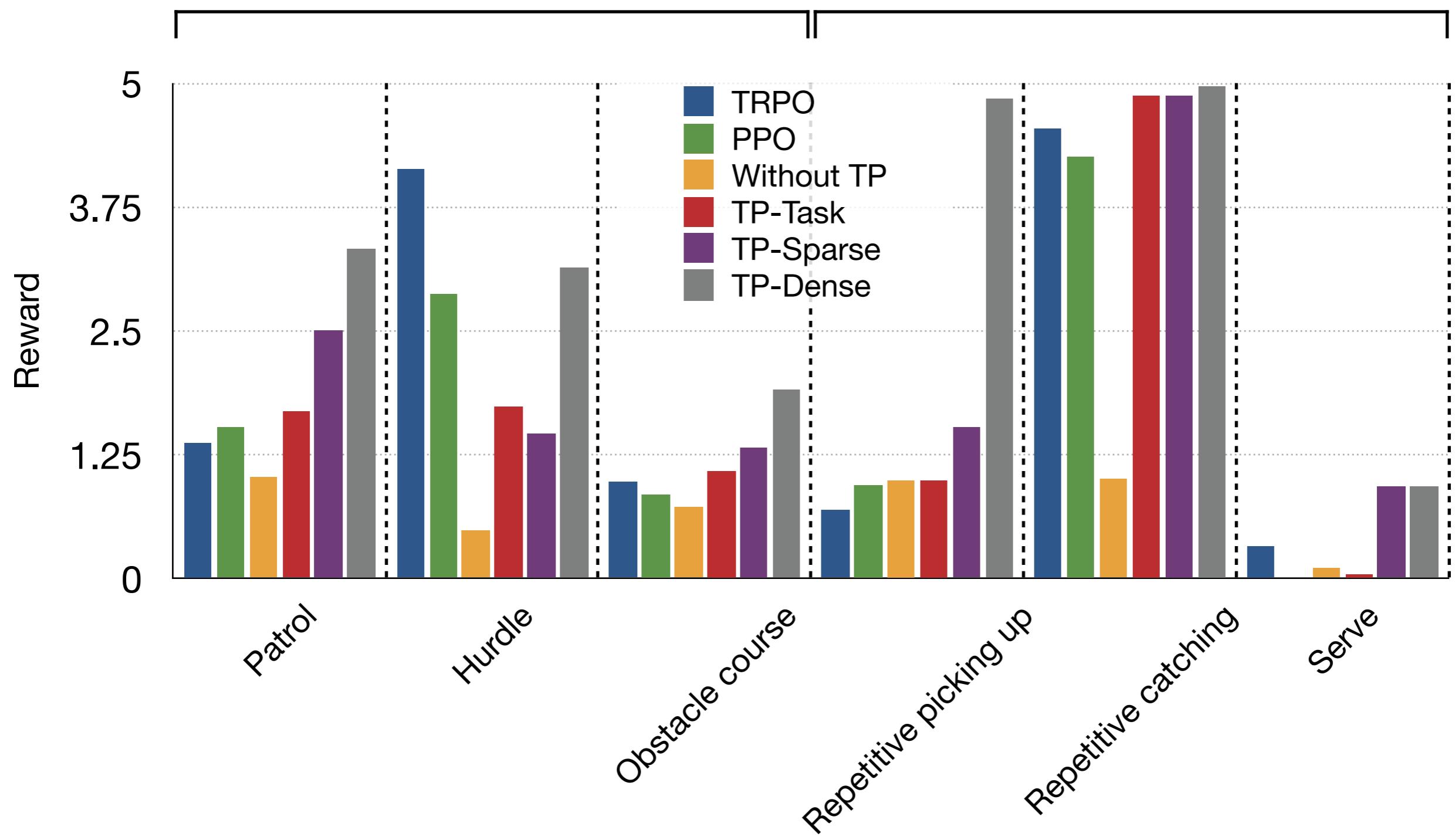
## Locomotion



# Quantitative Results - Reward

Locomotion

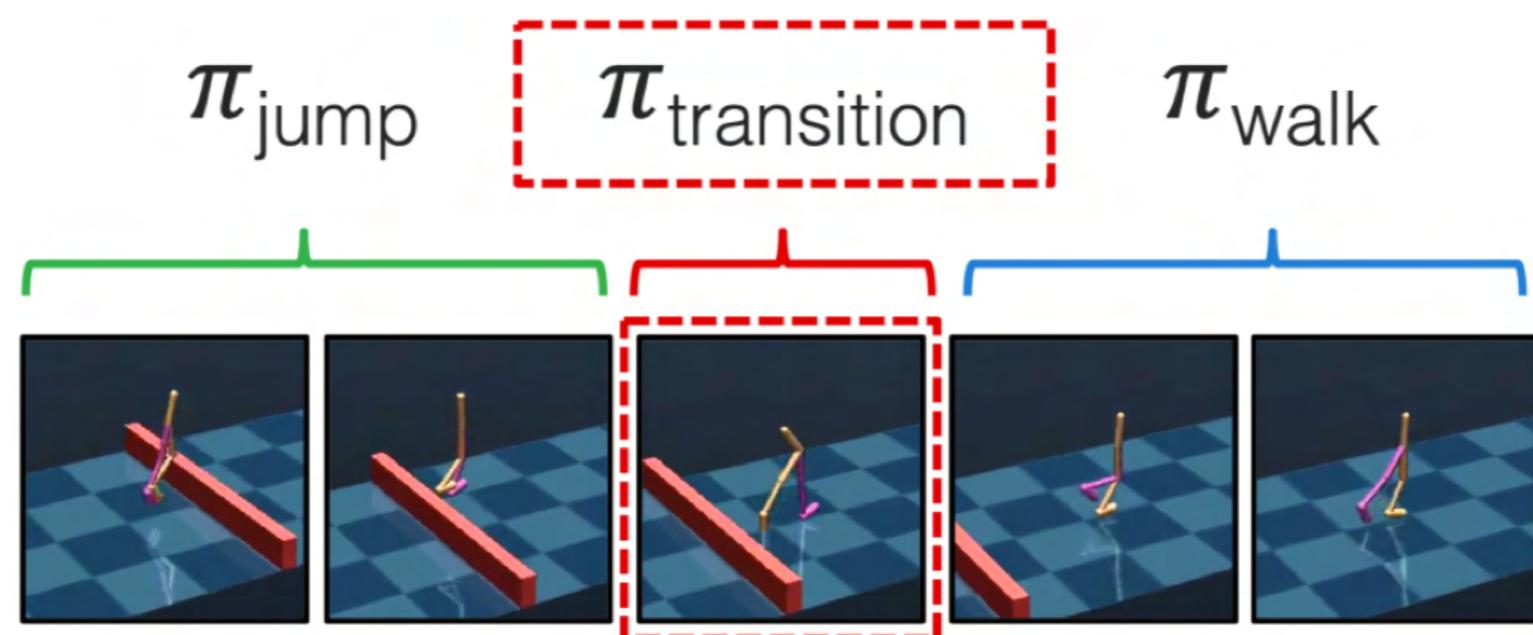
Manipulation



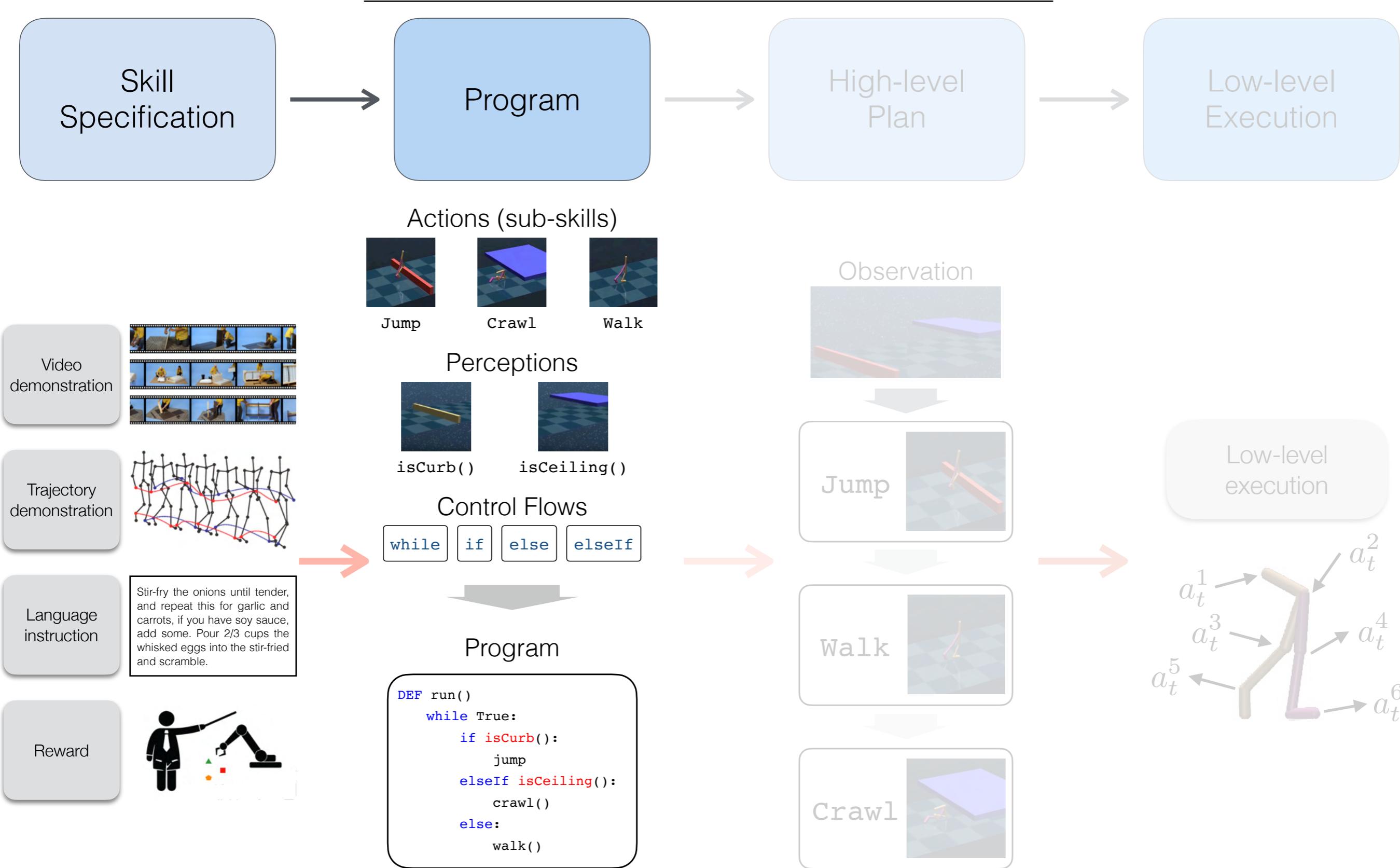
# Takeaway

---

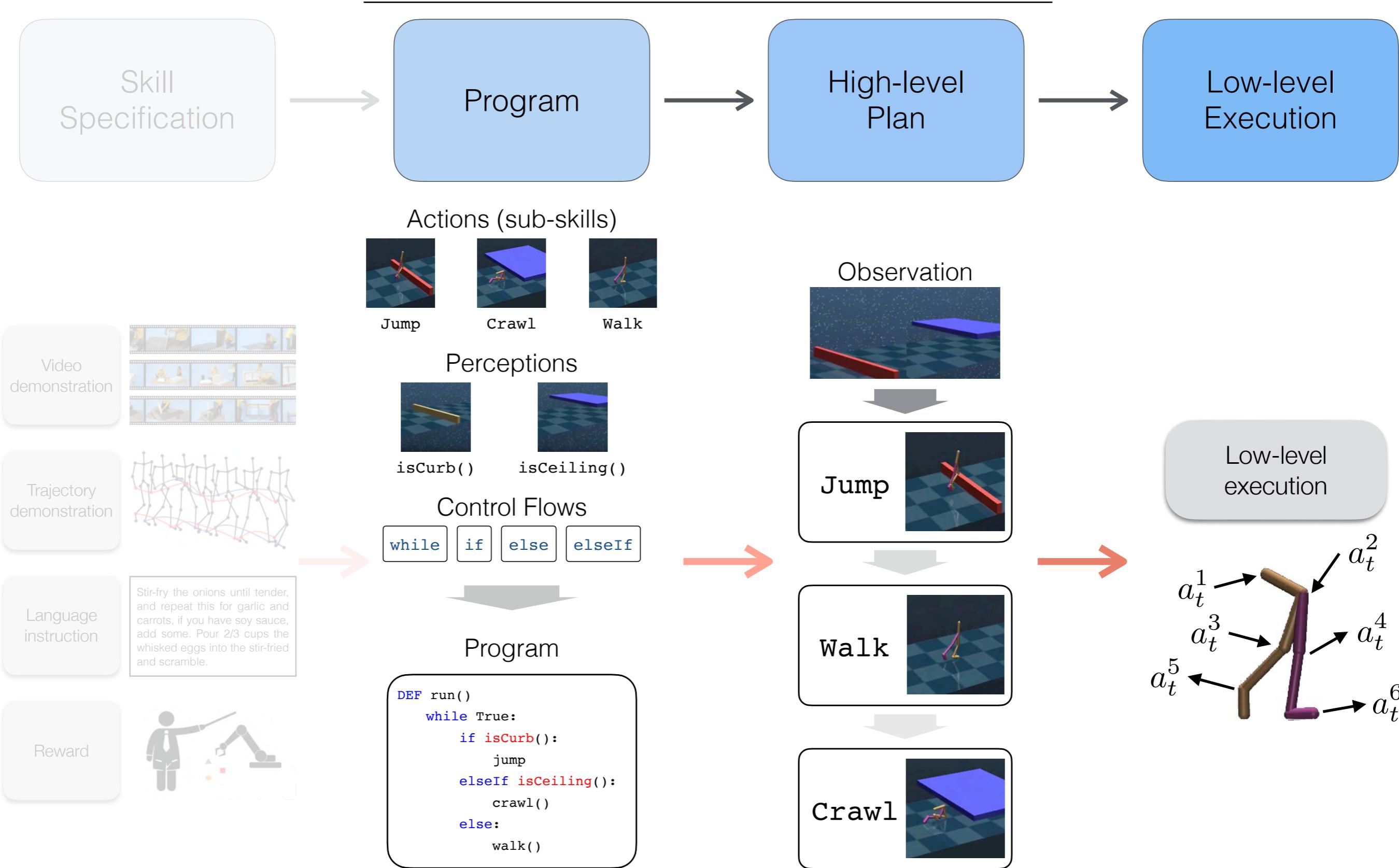
- Learning transition policies to smoothly compose learned skills



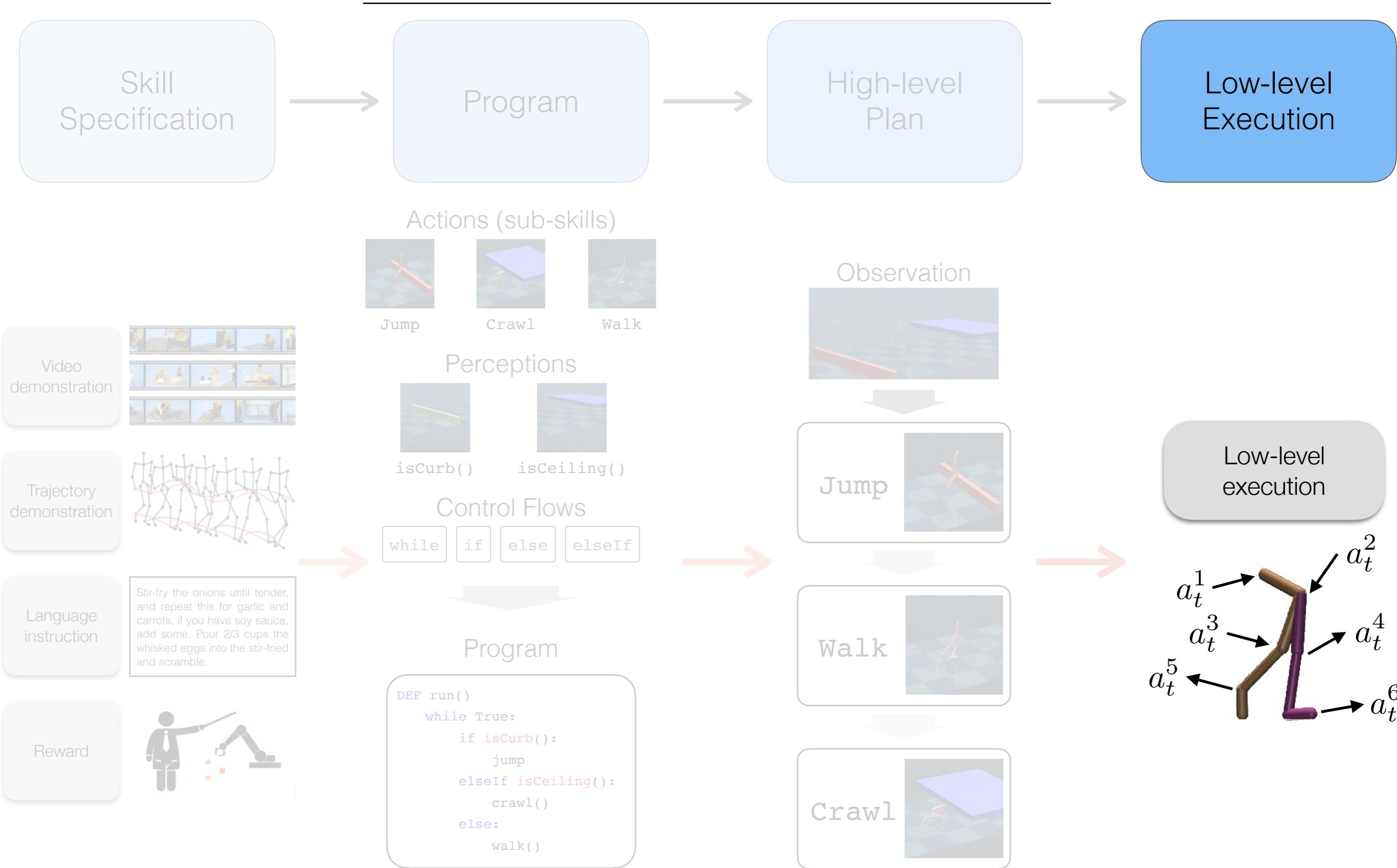
# Program Inference



# Task Execution



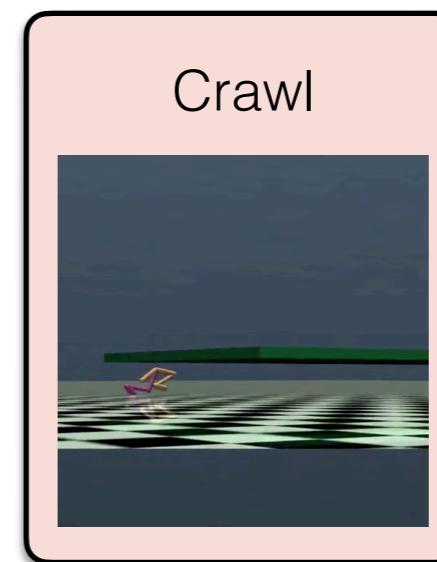
# Primitive Skill Acquisition



# Primitive Skill Acquisition

---

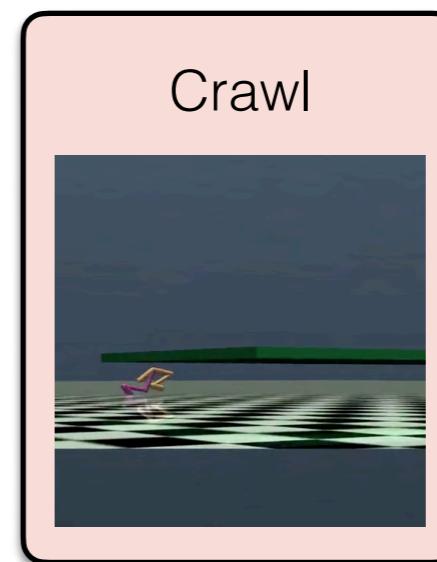
- Goal: acquire a diverse set of primitive skills efficiently



# Primitive Skill Acquisition

---

- Goal: acquire a diverse set of primitive skills efficiently



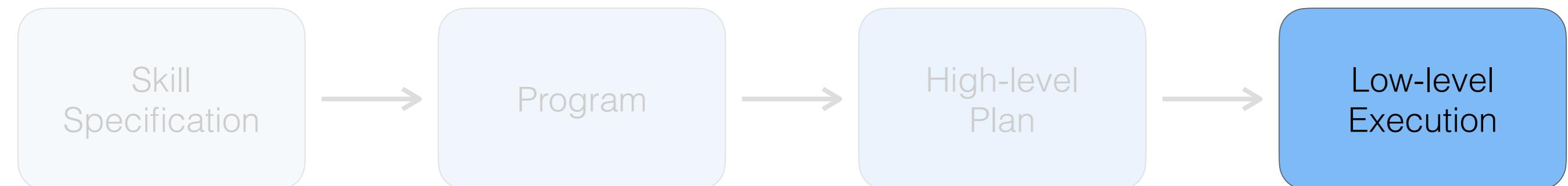
- Key directions

Meta-learning  
Meta-RL

Learning from  
experts

# Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation

NeurIPS 2019 (Spotlight)



Risto Vuorio

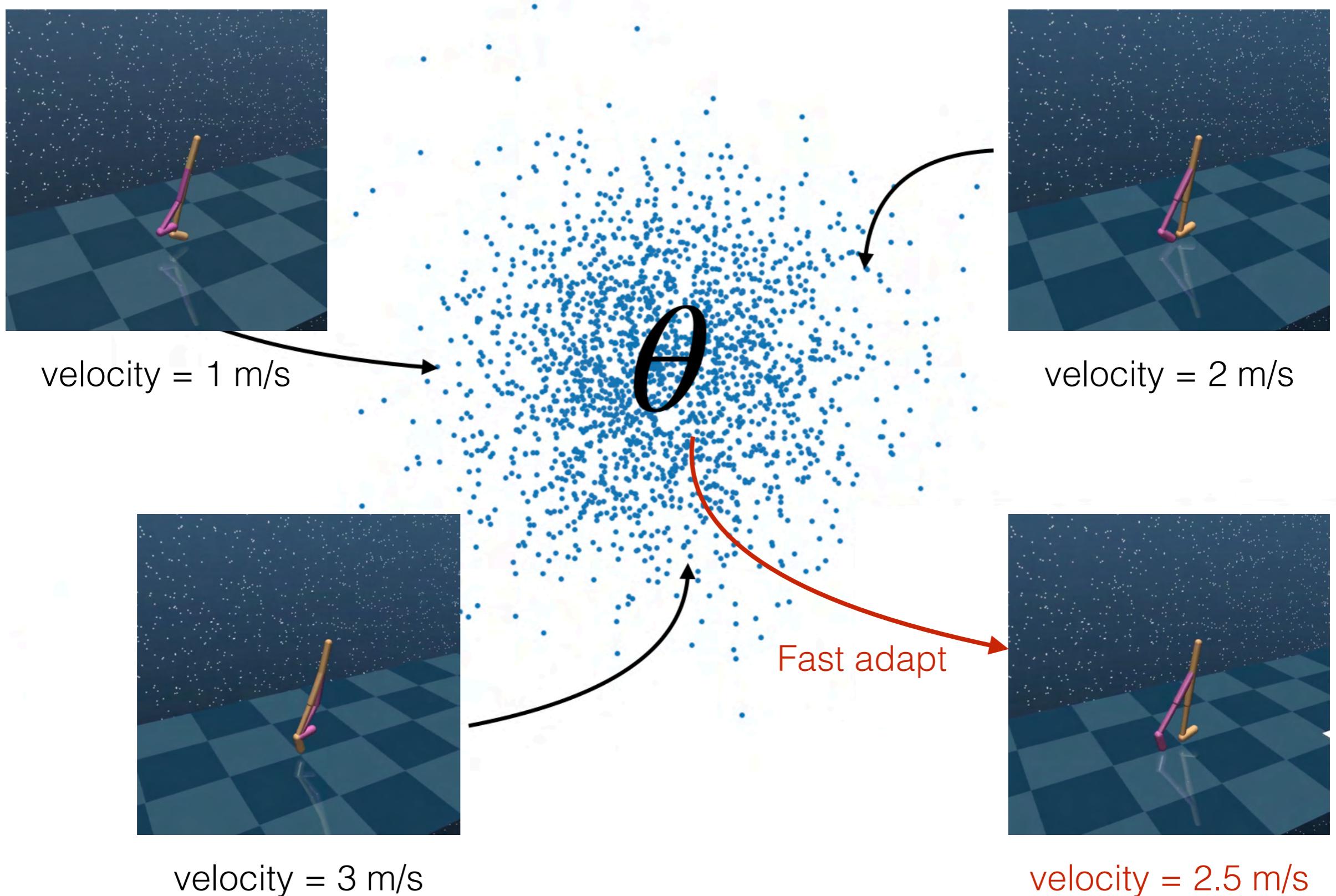


Hexiang Hu

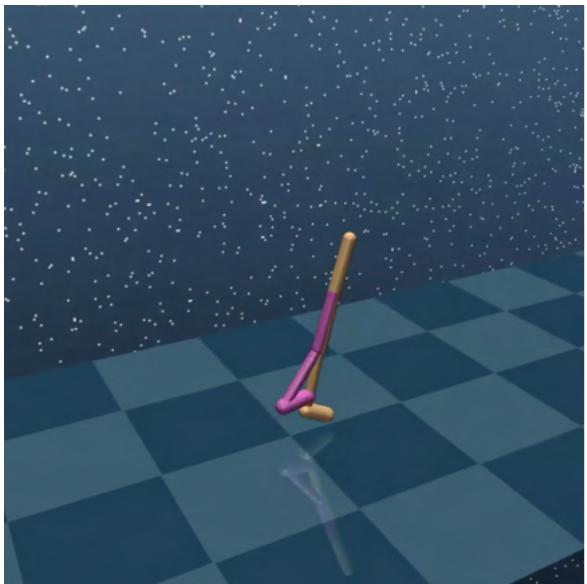


Joseph J. Lim

# Model-Agnostic Meta-Learning (MAML)

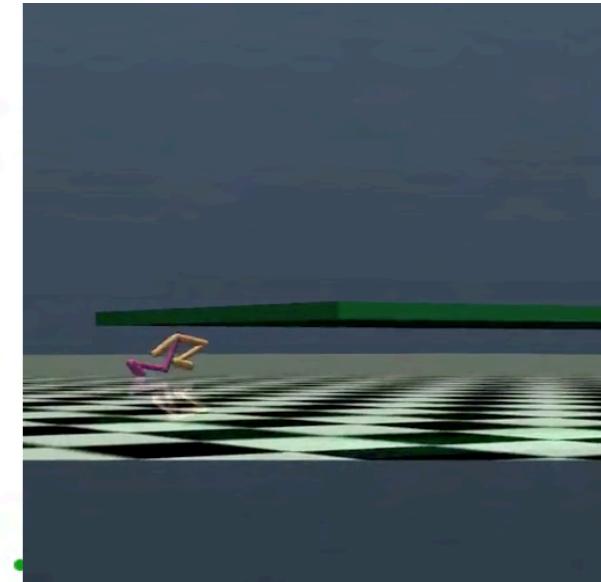


Walk

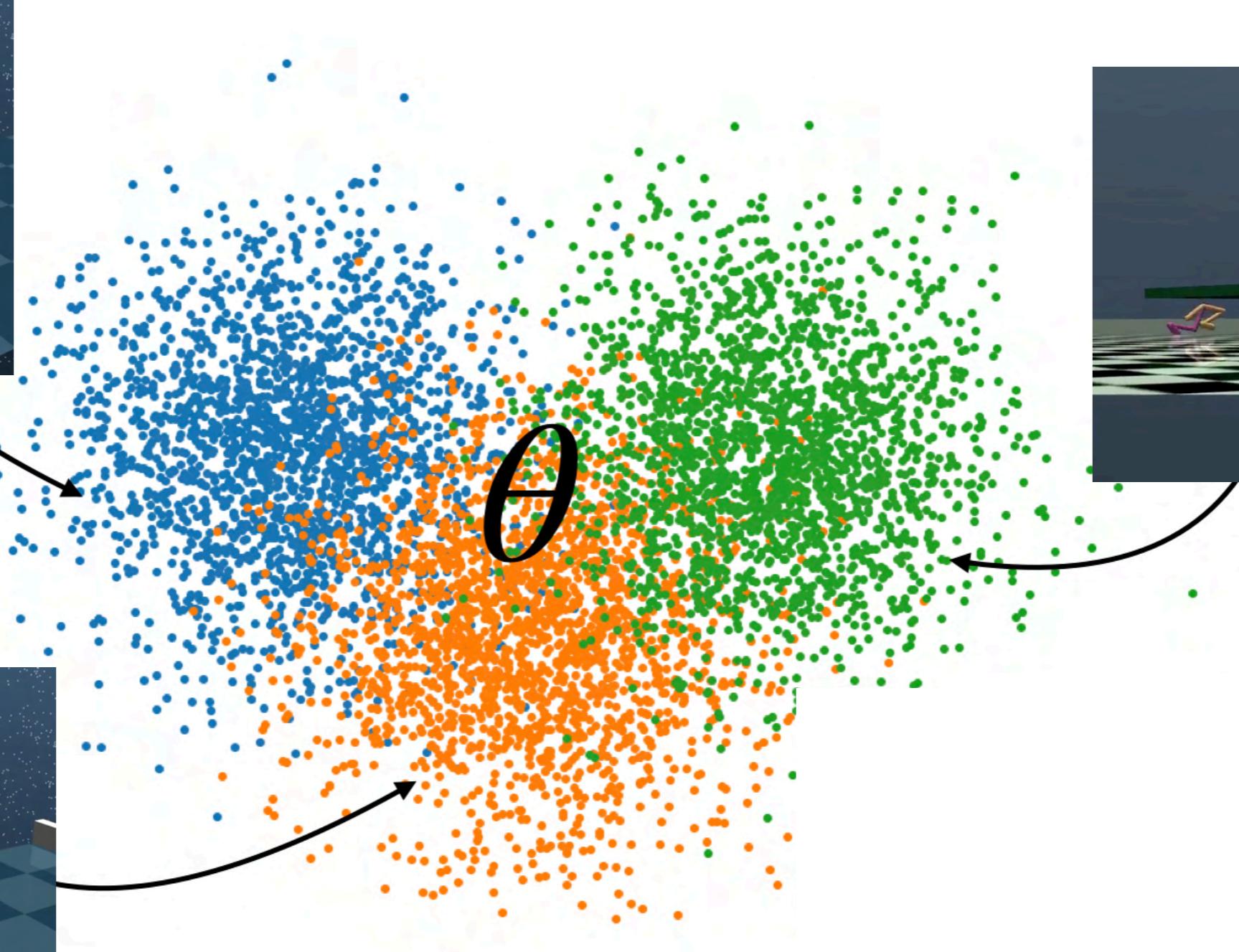
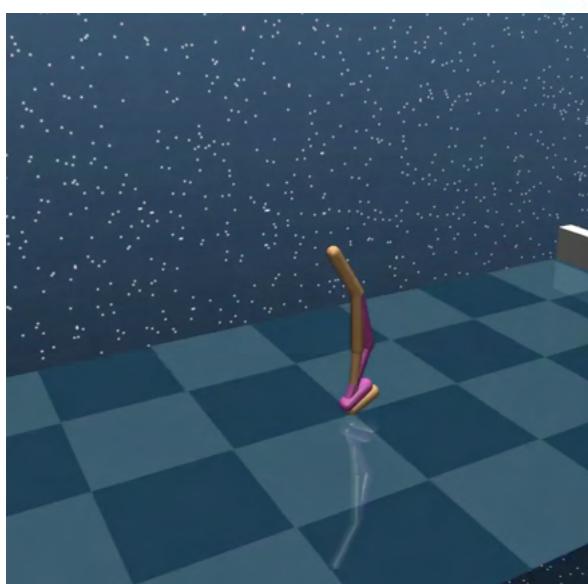


# Multimodal Task Distribution

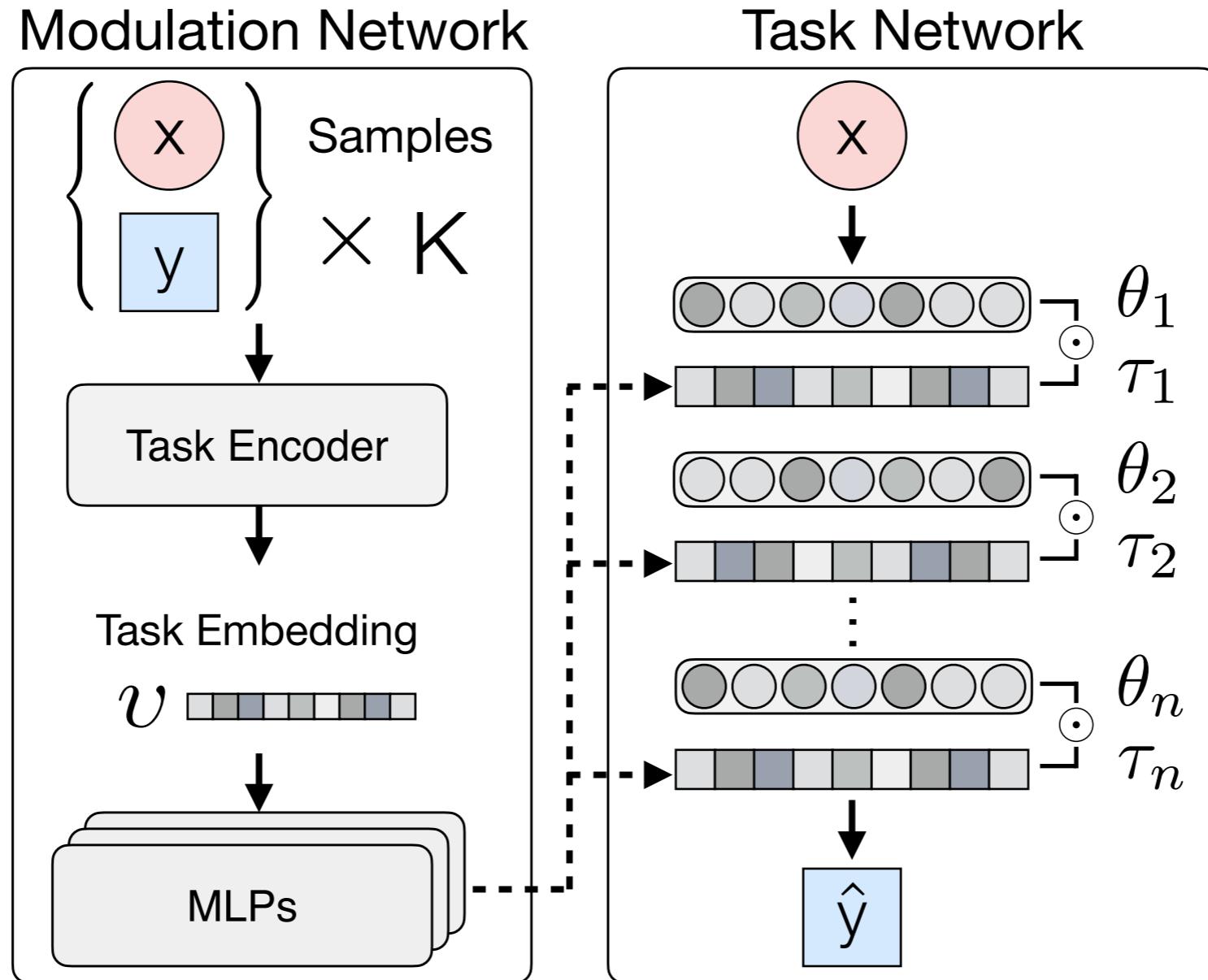
Crawl



Jump



# Multimodal Model-Agnostic Meta-Learning (MMAML)



# Training Algorithm

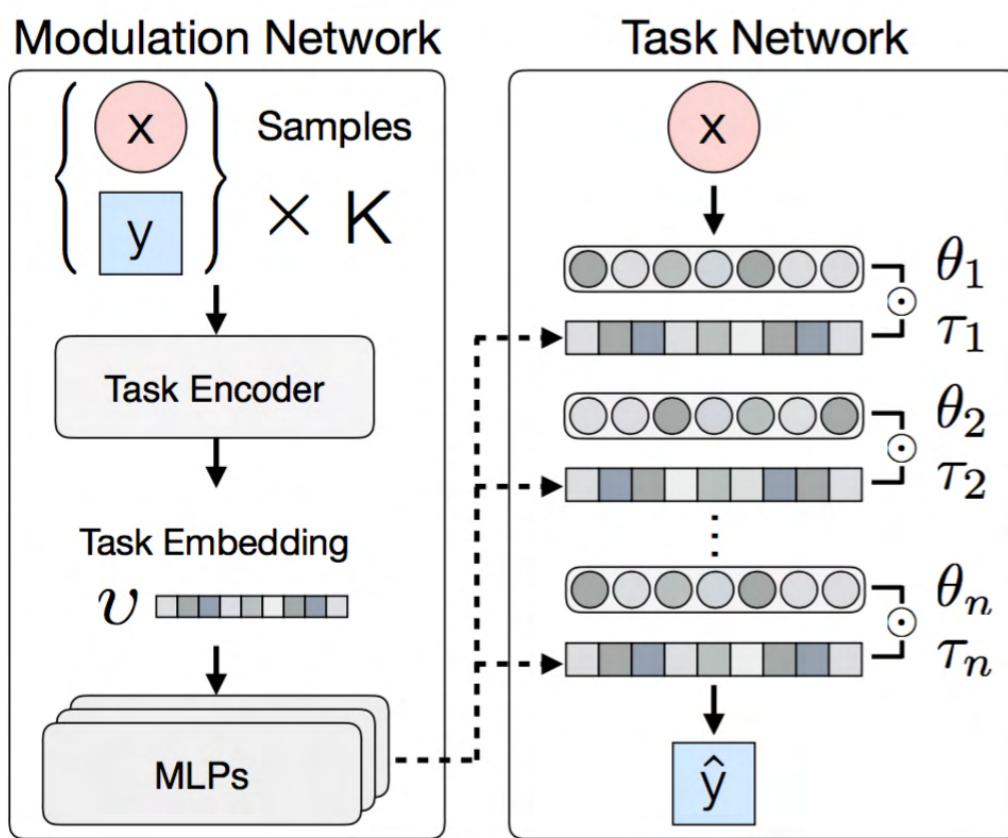
---

## Outer loop

- Task Encoder: produce the task embedding  $\omega_g$
- MLPs: modulate the task network blocks  $\omega_h$

## Inner loop

- Task network: fast adapt through gradient updates  $\theta$



**Algorithm 1** MMAML META-TRAINING PROCEDURE.

```

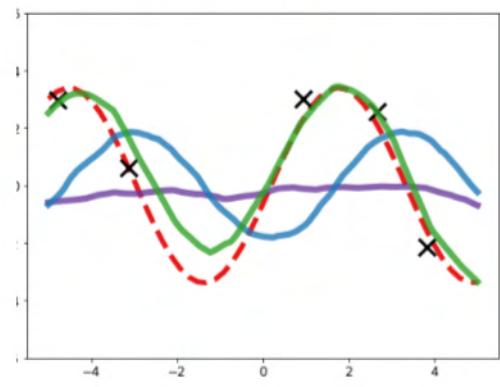
1: Input: Task distribution  $P(\mathcal{T})$ , Hyper-parameters  $\alpha$  and  $\beta$ 
2: Randomly initialize  $\theta$  and  $\omega$ .
3: while not DONE do
4:   Sample batches of tasks  $\mathcal{T}_j \sim P(\mathcal{T})$ 
5:   for all j do
6:     Infer  $v = h(\{x, y\}_K; \omega_h)$  with K samples from  $\mathcal{D}_{\mathcal{T}_j}^{\text{train}}$ .
7:     Generate parameters  $\tau = \{g_i(v; \omega_g) \mid i = 1, \dots, N\}$  to modulate each block of the task network  $f$ .
8:     Evaluate  $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_j}(f(x; \theta, \tau); \mathcal{D}_{\mathcal{T}_j}^{\text{train}})$  w.r.t the K samples
9:     Compute adapted parameter with gradient descent:
10:     $\theta'_{\mathcal{T}_j} = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_j}(f(x; \theta, \tau); \mathcal{D}_{\mathcal{T}_j}^{\text{train}})$ 
11:   end for
12:   Update  $\theta$  with  $\beta \nabla_{\theta} \sum_{T_j \sim P(\mathcal{T})} \mathcal{L}_{\mathcal{T}_j}(f(x; \theta', \tau); \mathcal{D}_{\mathcal{T}_j}^{\text{val}})$ 
13:   Update  $\omega_g$  with  $\beta \nabla_{\omega_g} \sum_{T_j \sim P(\mathcal{T})} \mathcal{L}_{\mathcal{T}_j}(f(x; \theta', \tau); \mathcal{D}_{\mathcal{T}_j}^{\text{val}})$ 
14: end while

```

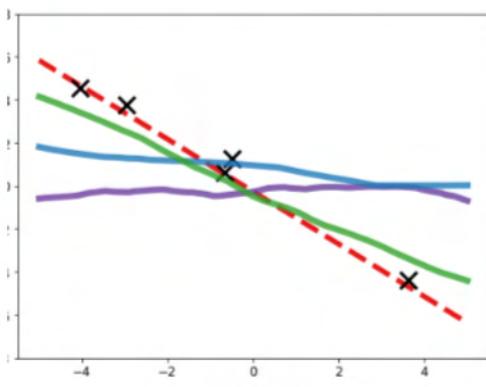
# Regression



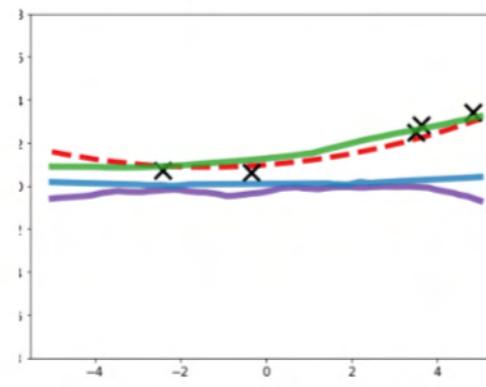
**Sinusoidal**



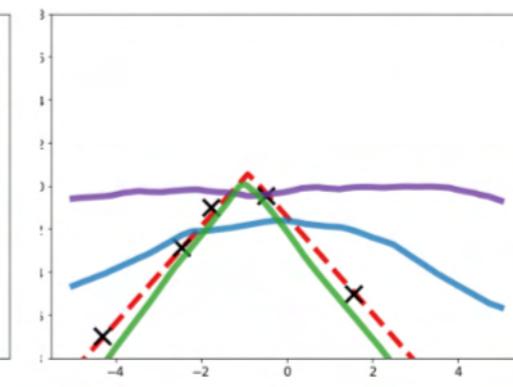
**Linear**



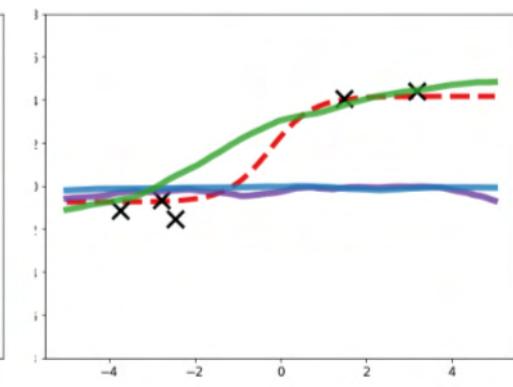
**Quadratic**



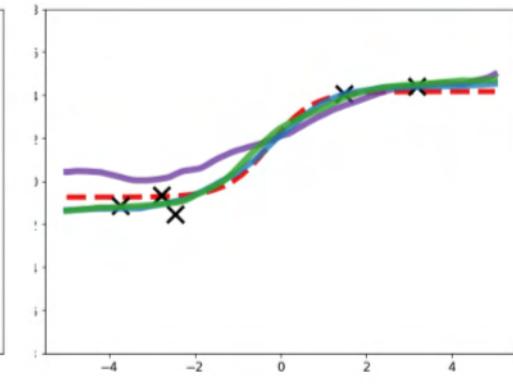
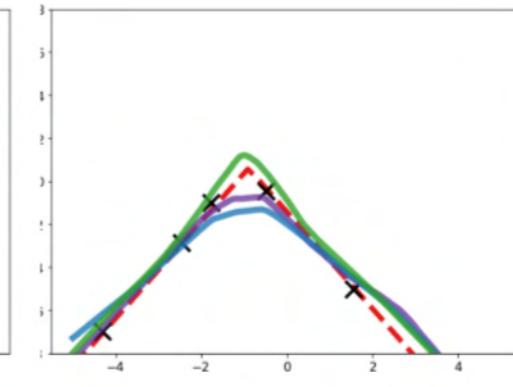
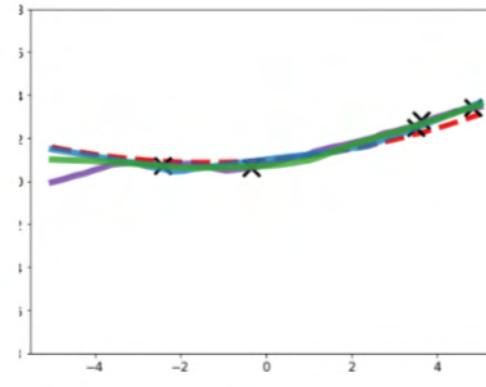
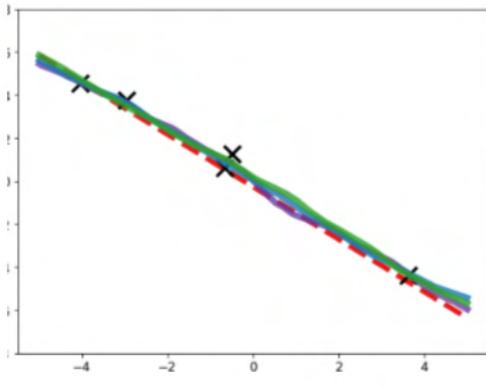
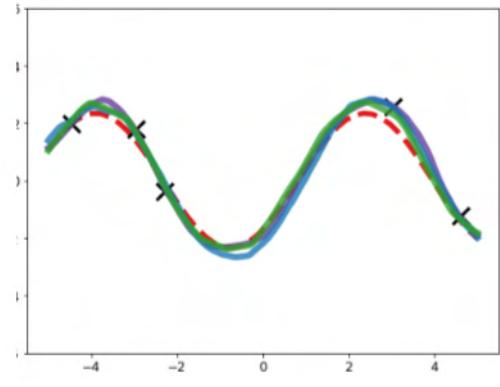
**Transformed  $\ell_1$  Norm**



**Tanh**



**(a) MMAML post modulation vs. other prior models**

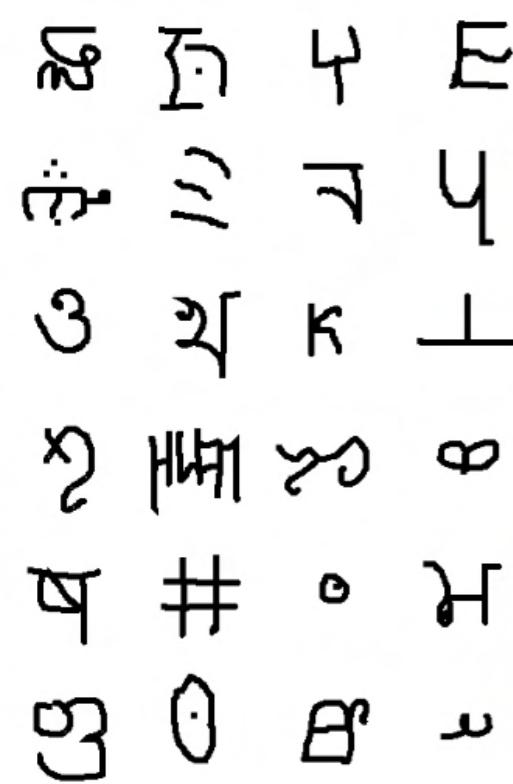


**(b) MMAML post adaptation vs. other posterior models**

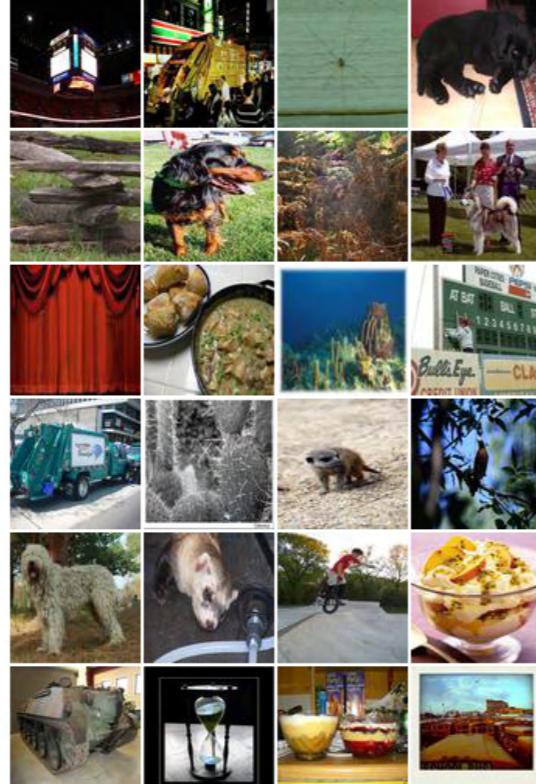
Method	2 Modes		3 Modes		5 Modes	
	Post Modulation	Post Adaptation	Post Modulation	Post Adaptation	Post Modulation	Post Adaptation
MAML [1]	-	1.085	-	1.231	-	1.668
Multi-MAML	-	0.433	-	0.713	-	1.082
LSTM Learner	0.362	-	0.548	-	0.898	-
<b>Ours: MMAML (Softmax)</b>	1.548	0.361	2.213	<b>0.444</b>	2.421	0.939
<b>Ours: MMAML (FiLM)</b>	2.421	<b>0.336</b>	1.923	<b>0.444</b>	2.166	<b>0.868</b>

# Image Classification

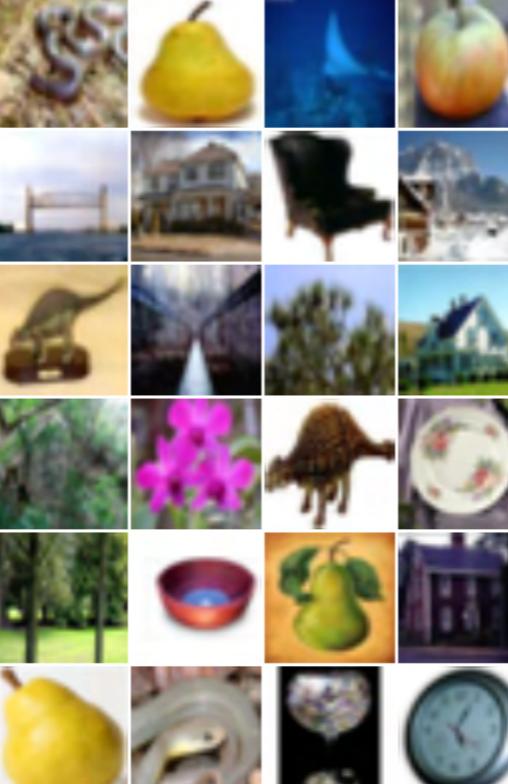
(a) Omniglot



(b) Mini-ImageNet



(c) FC100



(d) CUB



(e) Aircraft



**Method & Setup**

**2 Modes**

**3 Modes**

**5 Modes**

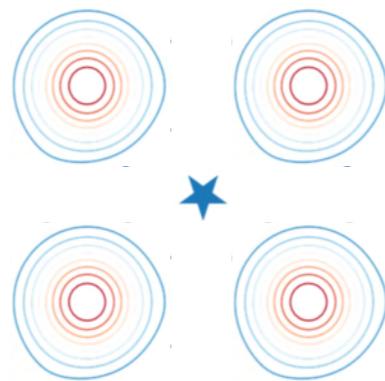
Way Shot	5-way 1-shot		20-way 1-shot		5-way 1-shot		20-way 1-shot		5-way 1-shot		20-way 3-shot	
MAML [1]	66.80%	77.79%	44.69%	54.55%	67.97%	28.22%	66.00%	70.87%	44.69%	56.57%		
Multi-MAML	66.85%	73.07%	<b>53.15%</b>	55.90%	62.20%	<b>39.77%</b>	71.94%	76.94%	58.66%	61.11%		
MMAML (ours)	<b>69.93%</b>	<b>78.73%</b>	47.80%	<b>57.47%</b>	<b>70.15%</b>	36.27%	<b>72.02%</b>	<b>78.13%</b>	<b>60.13%</b>	<b>63.52%</b>		

# Reinforcement Learning

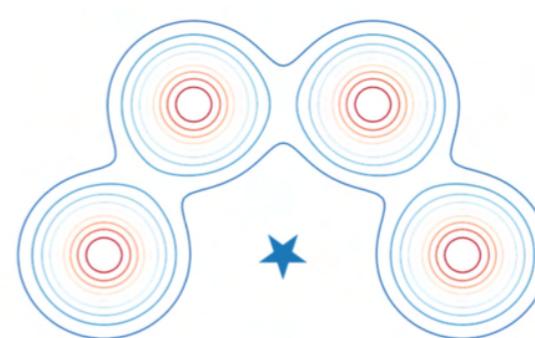
---

Goal modes

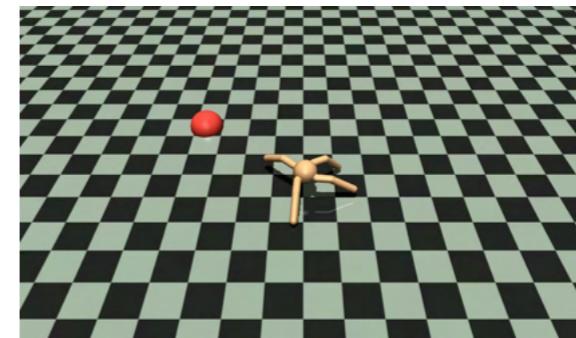
Reacher



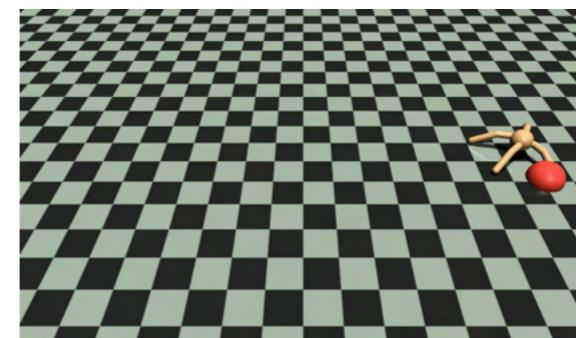
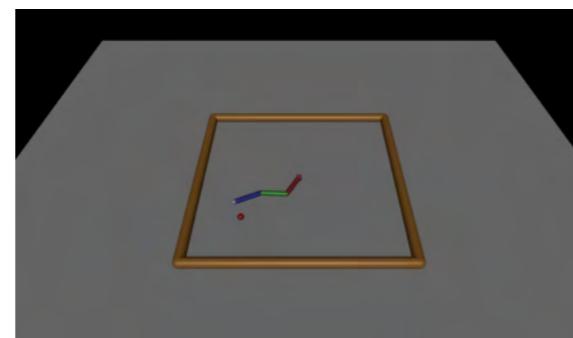
Ant



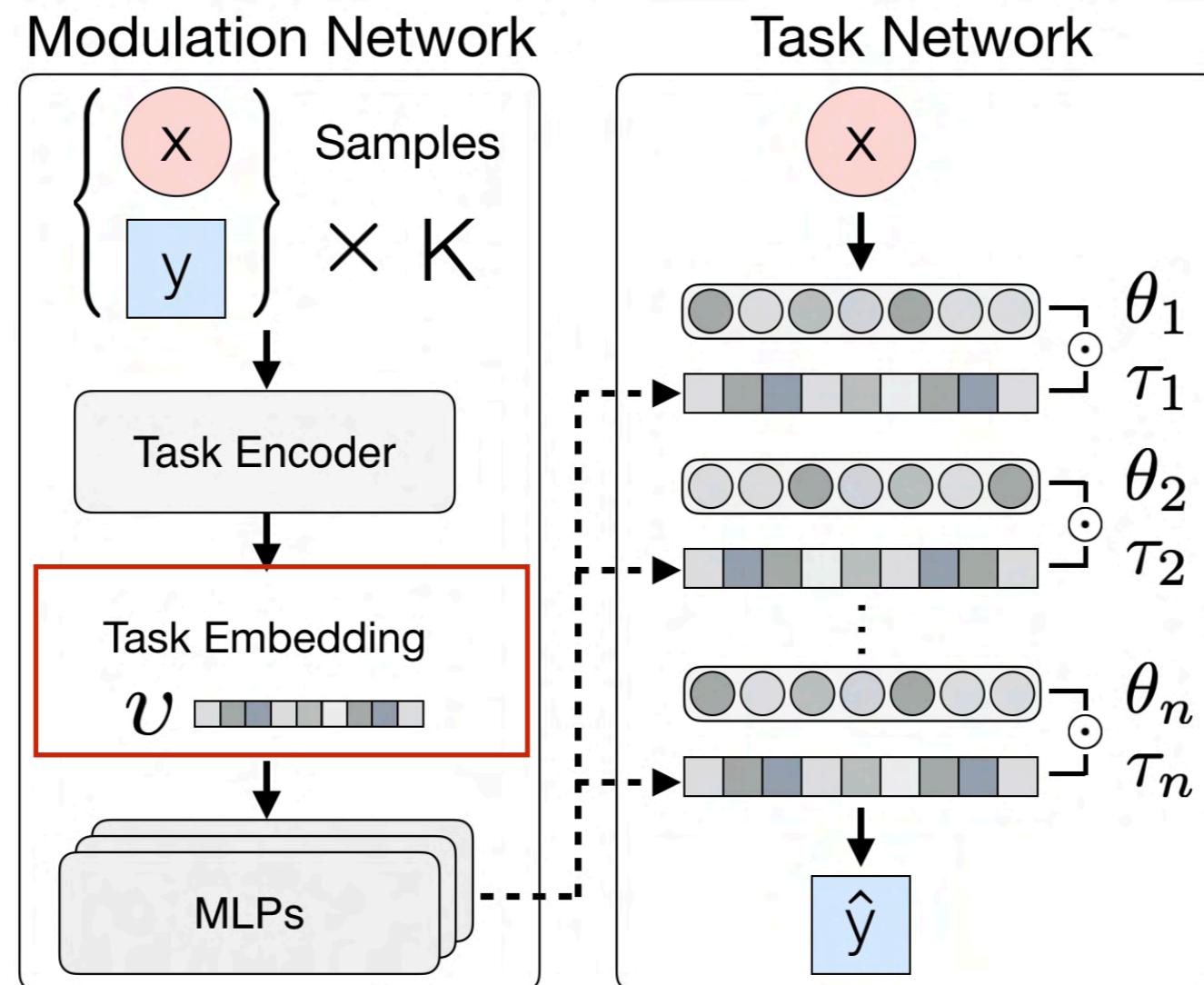
ProMP



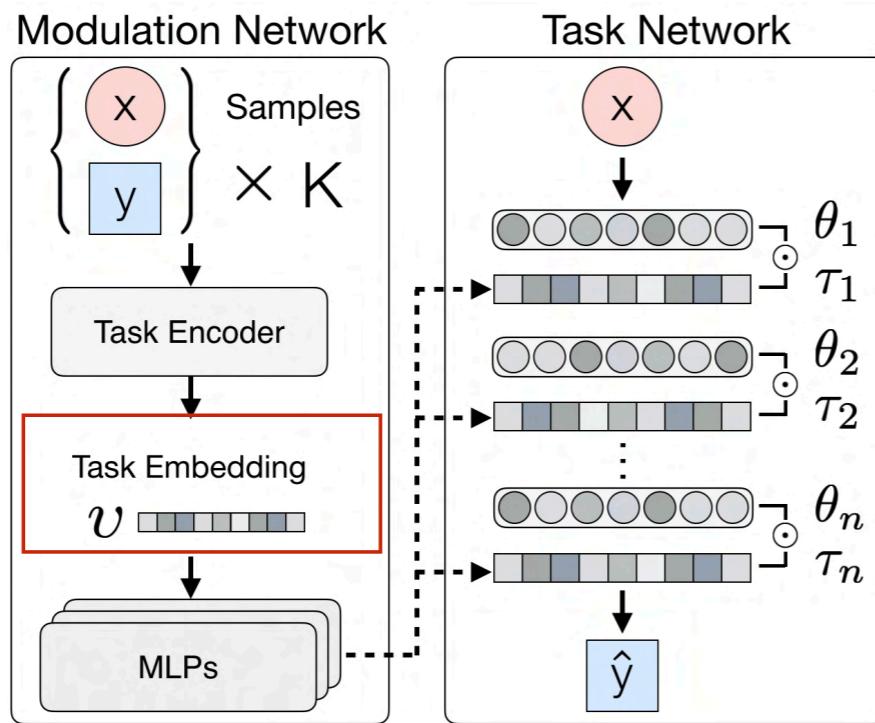
Ours



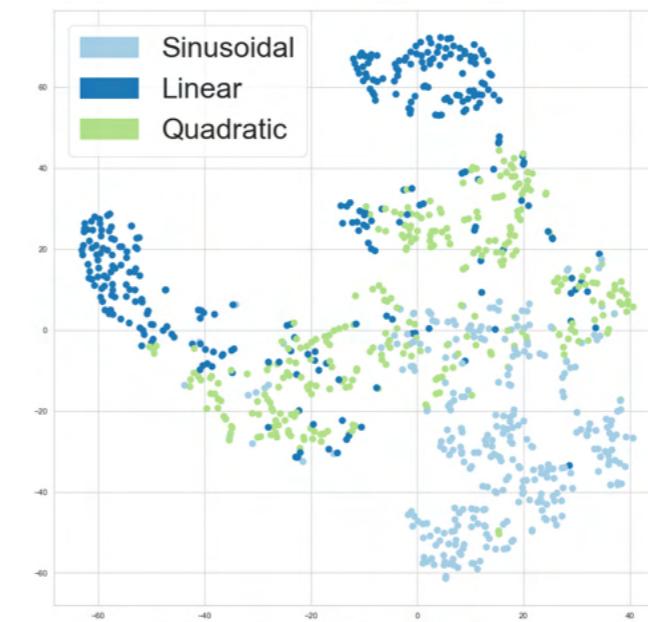
# Learned Task Embedding (tSNE plot)



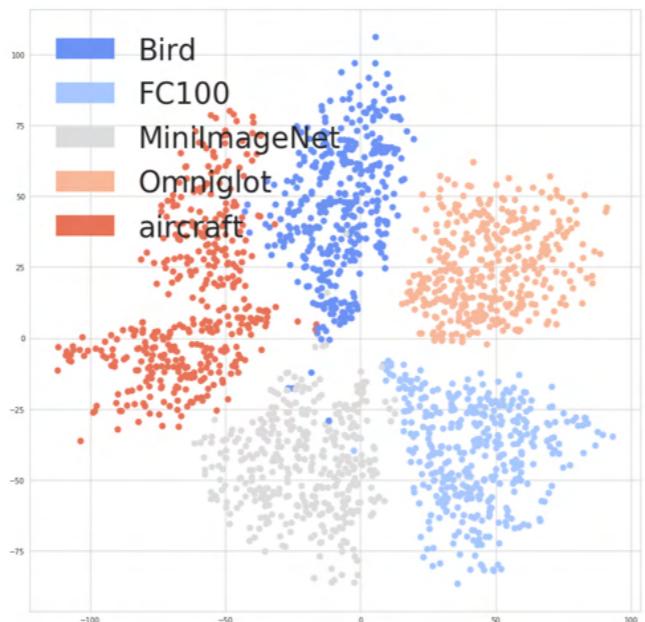
# Learned Task Embedding (tSNE plot)



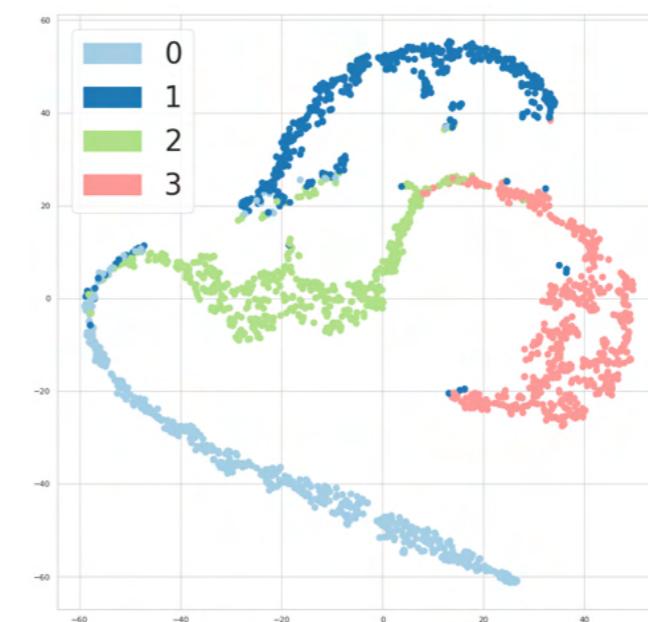
## 3-mode Regression



## 5-mode Classification



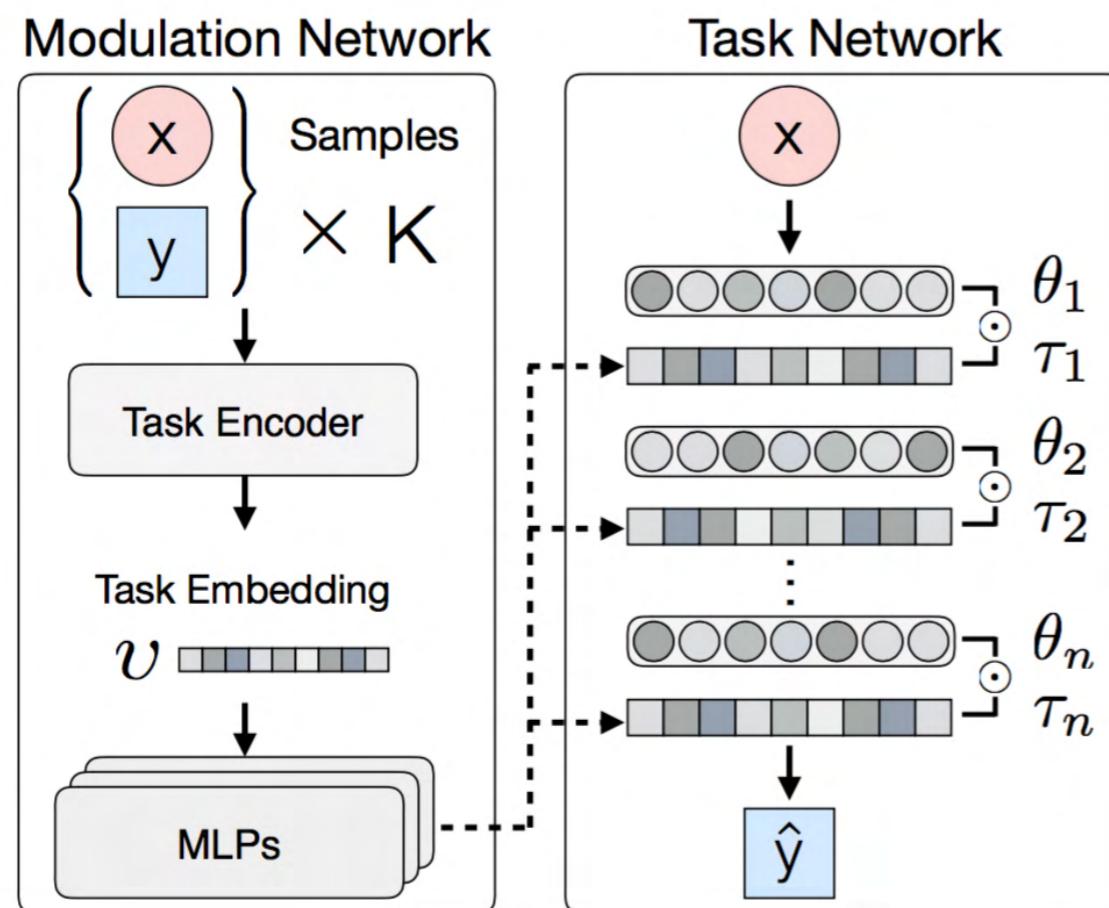
## 4-mode Reacher



# Takeaway

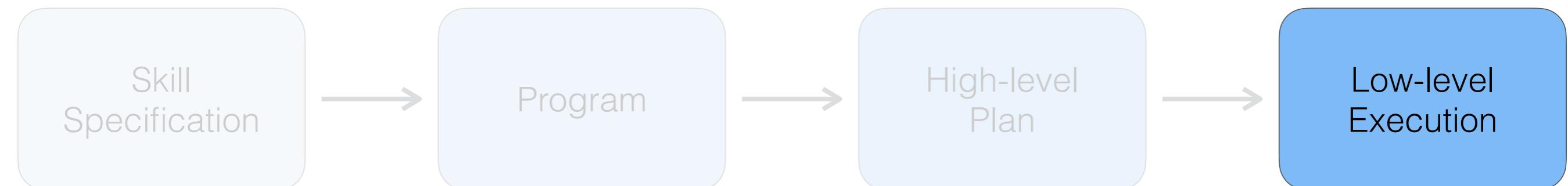
---

- MAML struggles at learning from multimodal task distributions
- We propose **multimodal MAML** to alleviate the issue

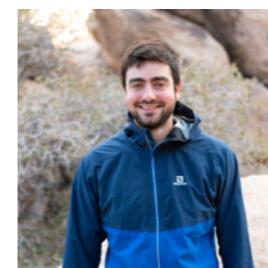


# Skill-based Meta-Reinforcement Learning

ICLR 2022



Taewook Nam



Karl Pertsch

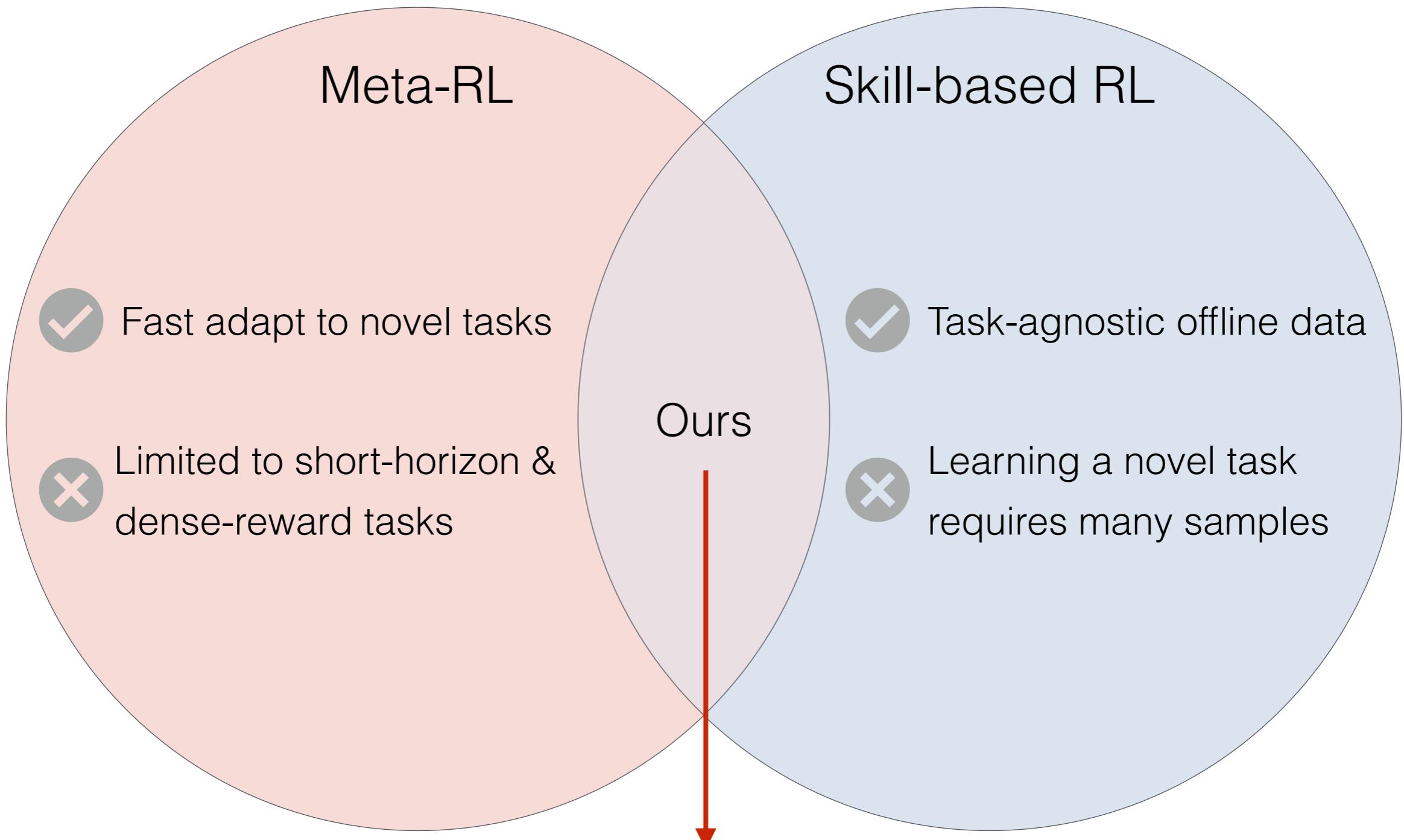


Sung Ju Hwang



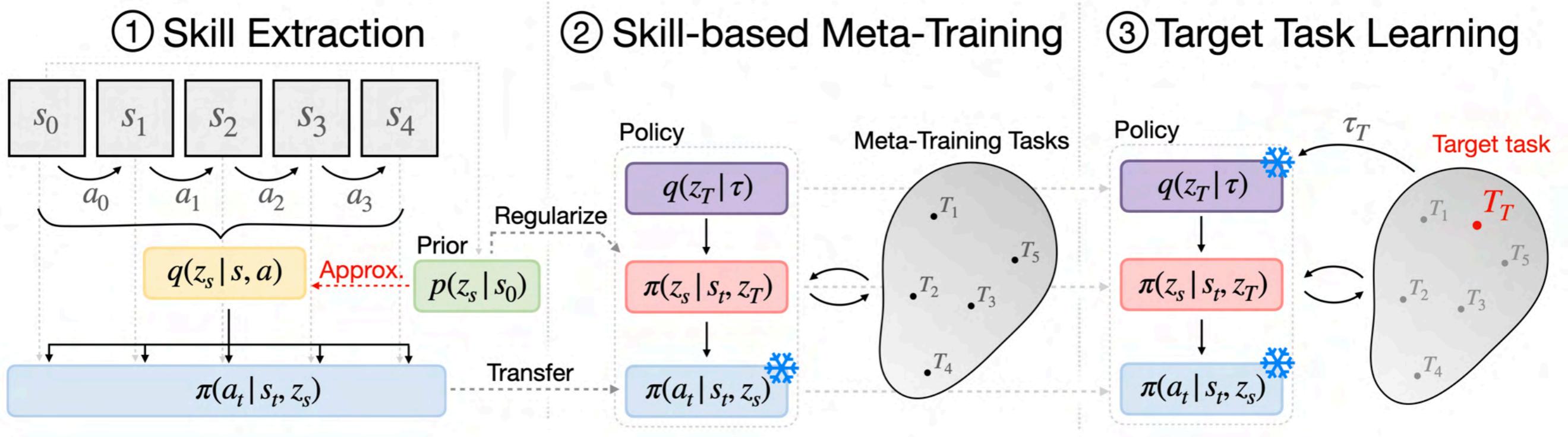
Joseph J. Lim

# Meta-RL with Skills

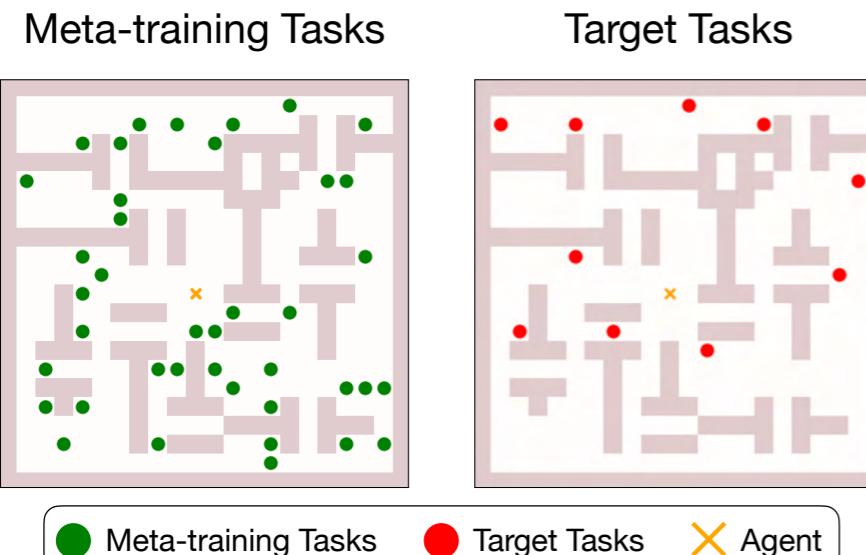


- Learn skills from task-agnostic offline data
- Meta-learn on long-horizon, sparse-reward tasks

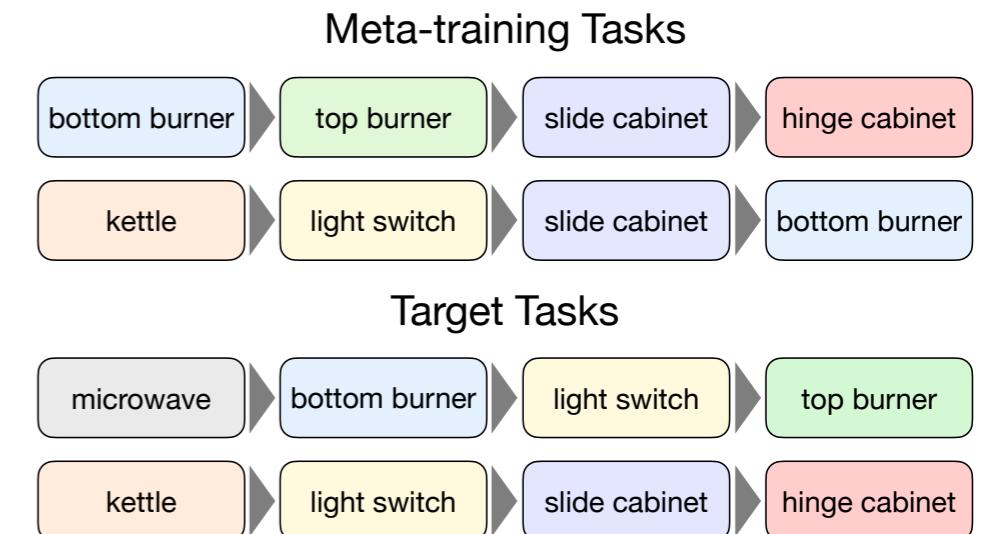
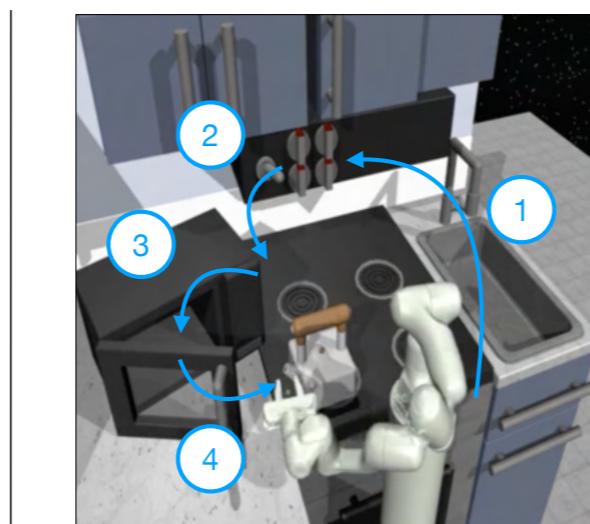
# SiMPL: Skill-based Meta Policy Learning



# Environments

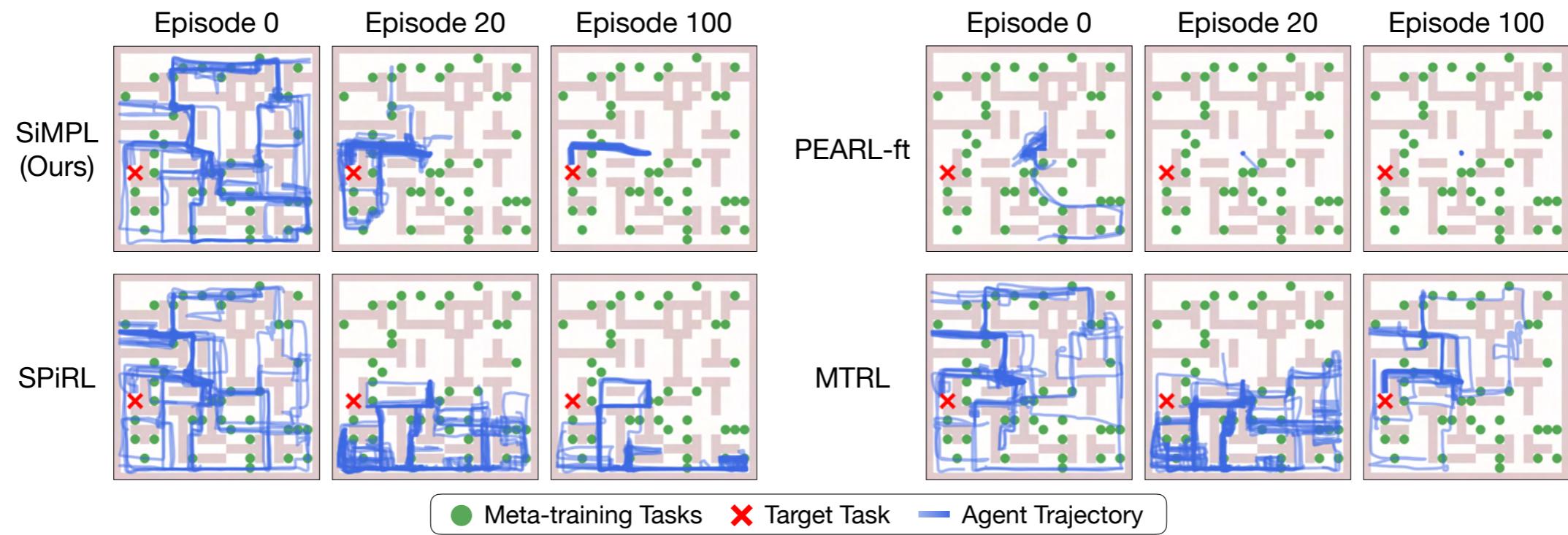
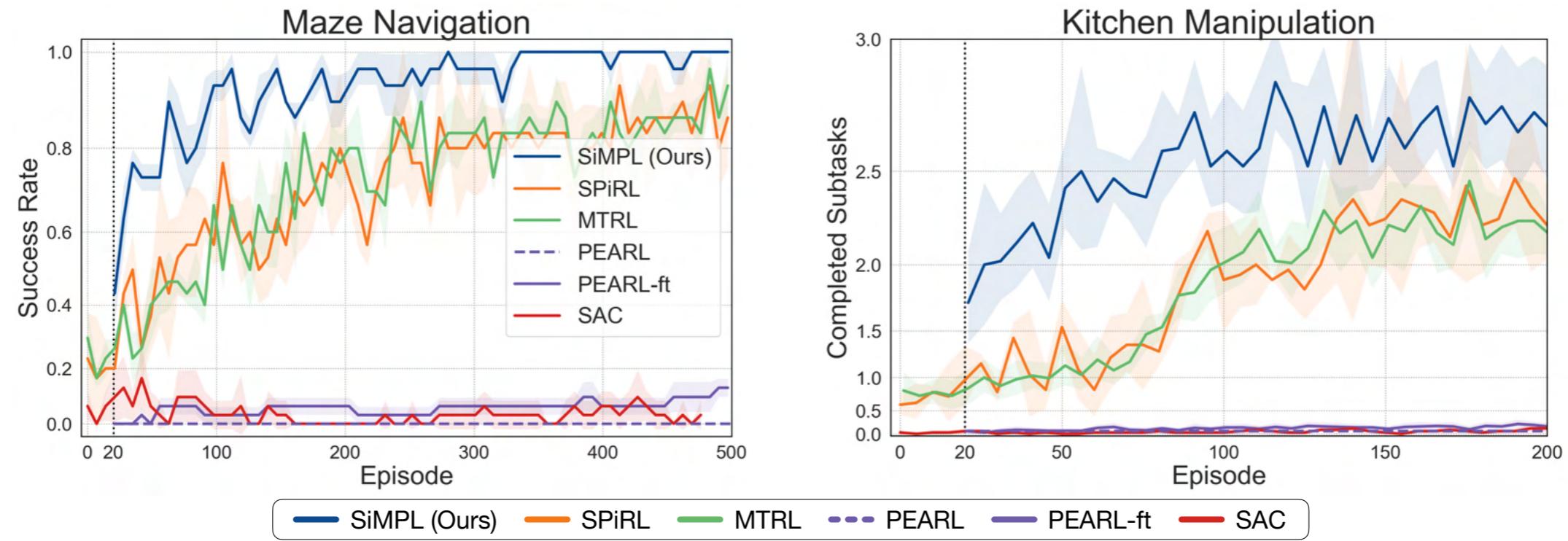


(a) Maze Navigation



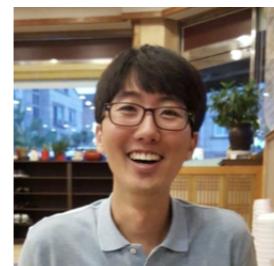
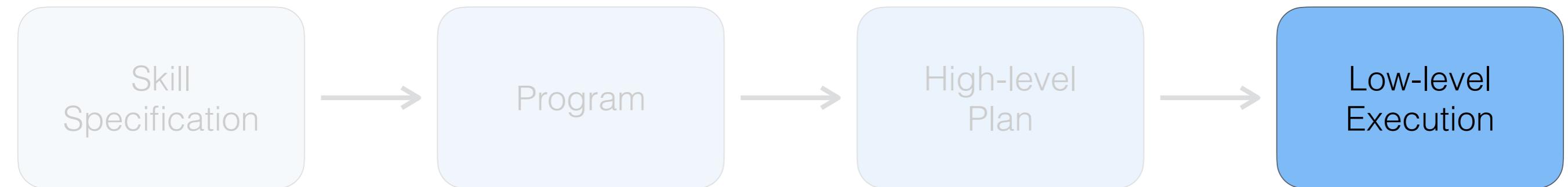
(b) Kitchen Manipulation

# Results

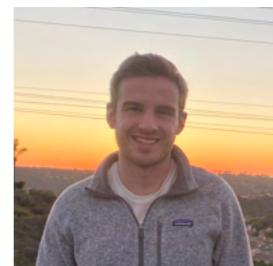


# Generalizable Imitation Learning from Observation via Inferring Goal Proximity

NeurIPS 2021



Youngwoon Lee

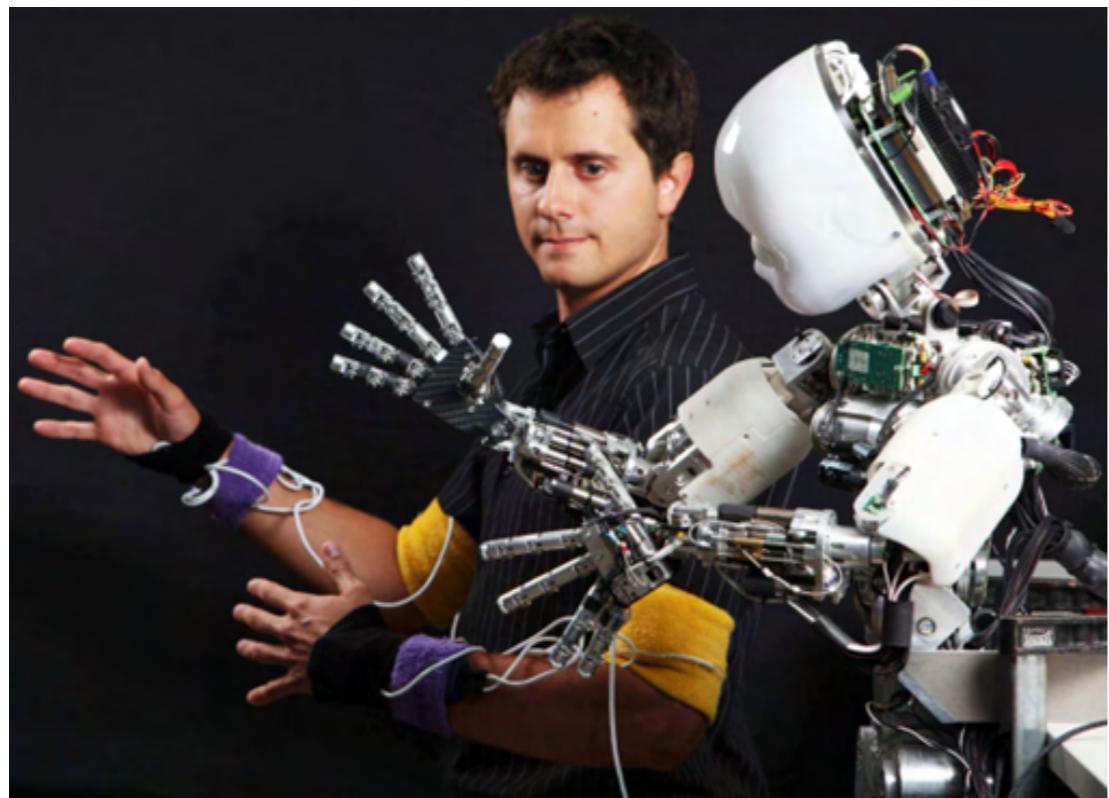


Andrew Szot



Joseph J. Lim

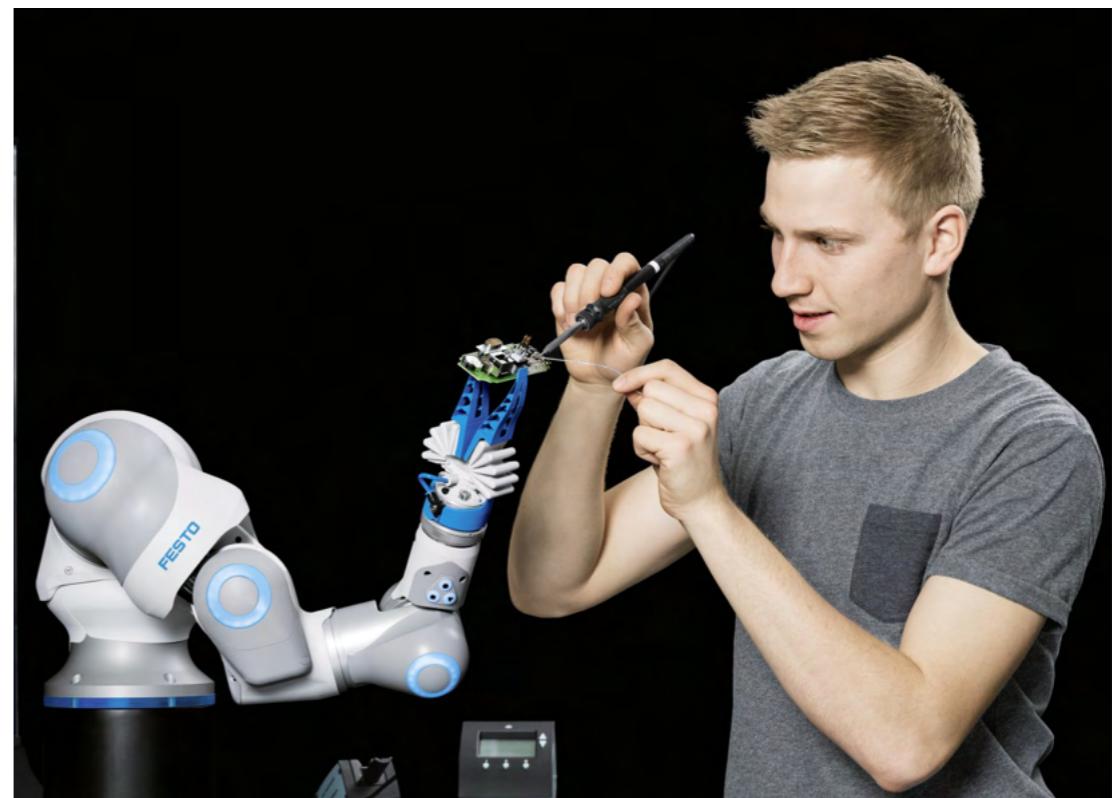
## Learning from Demonstration



with expert's actions

Demo:  $\{s_1, a_1, s_2, a_2, s_3, a_3, \dots\}$

## Learning from Observation

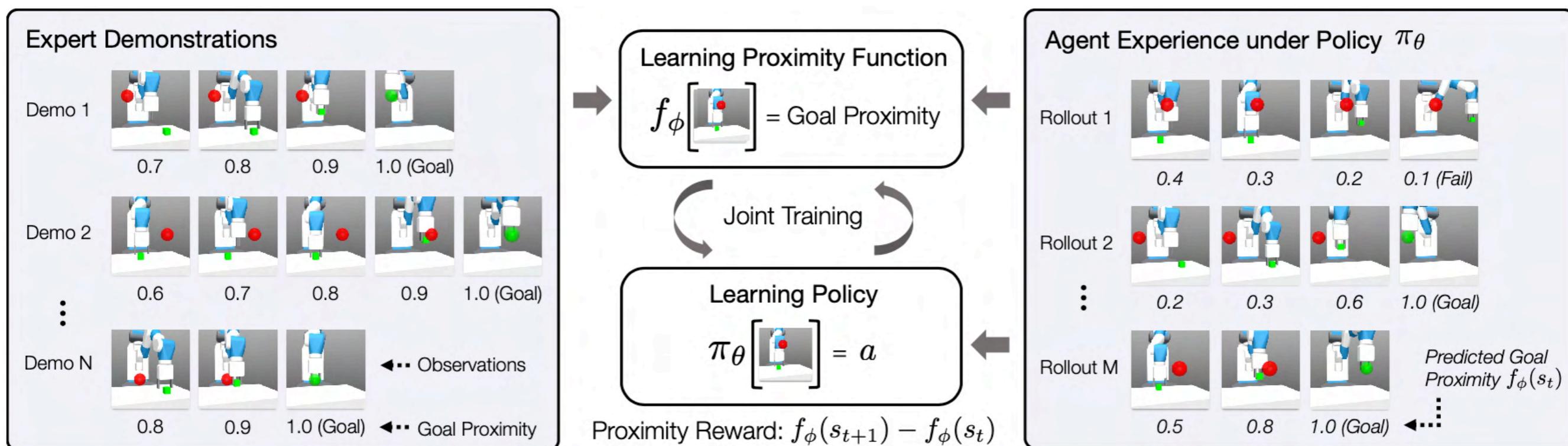


vs.

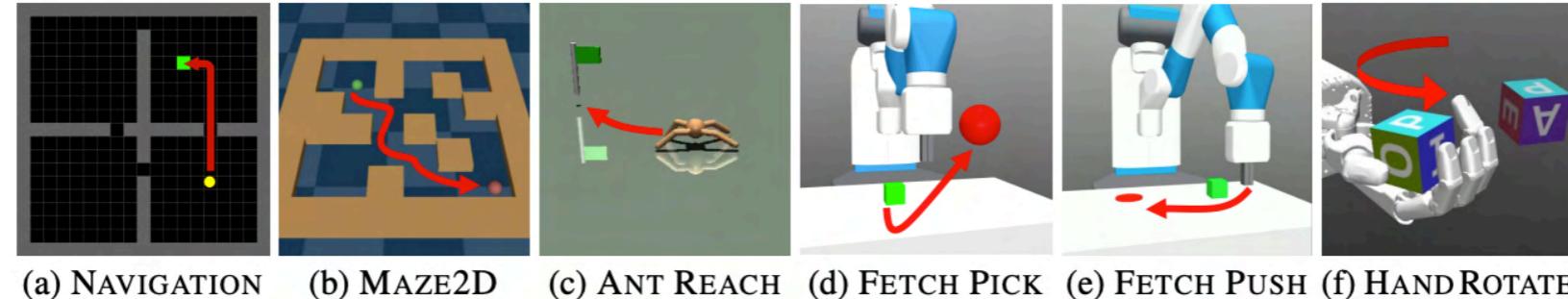
without expert's actions

Demo:  $\{s_1, s_2, s_3, \dots\}$

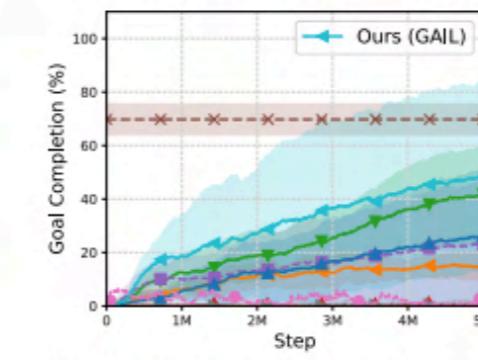
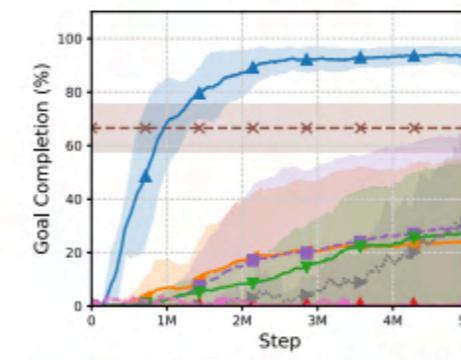
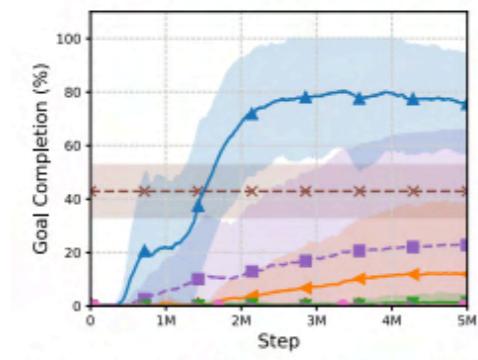
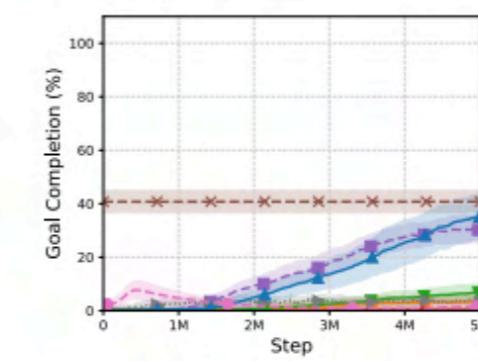
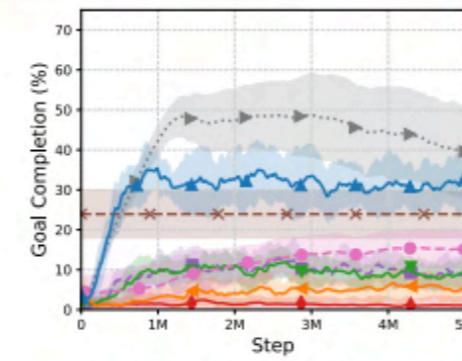
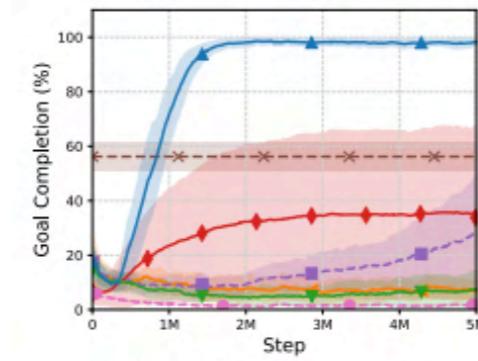
# Learning from Observation via Inferring Goal Proximity



# Experiments

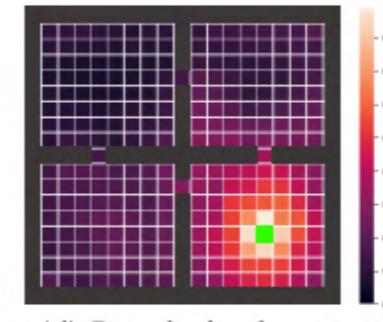
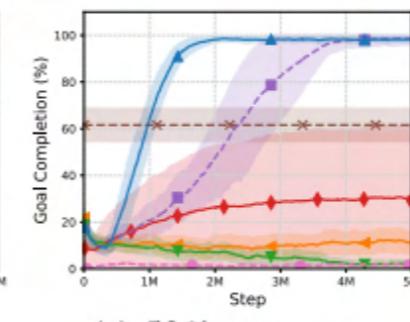
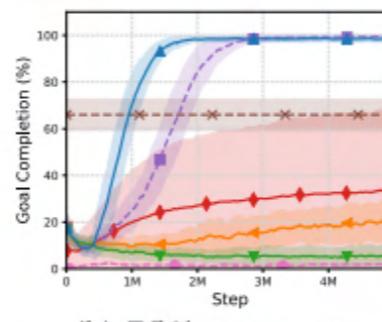
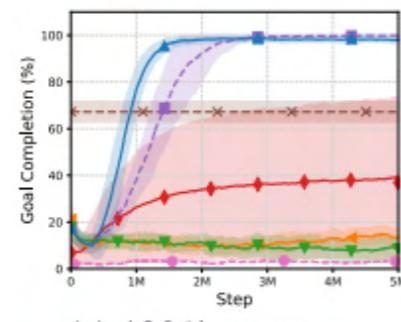


LfO   ◆ Ours   ▼ GAIfO-s   ◀ GAIfO   ◆ BCO   LfD   ● SQL   ■ GAIL   ✖ BC   LfO+reward   ▶ GoalGAIL

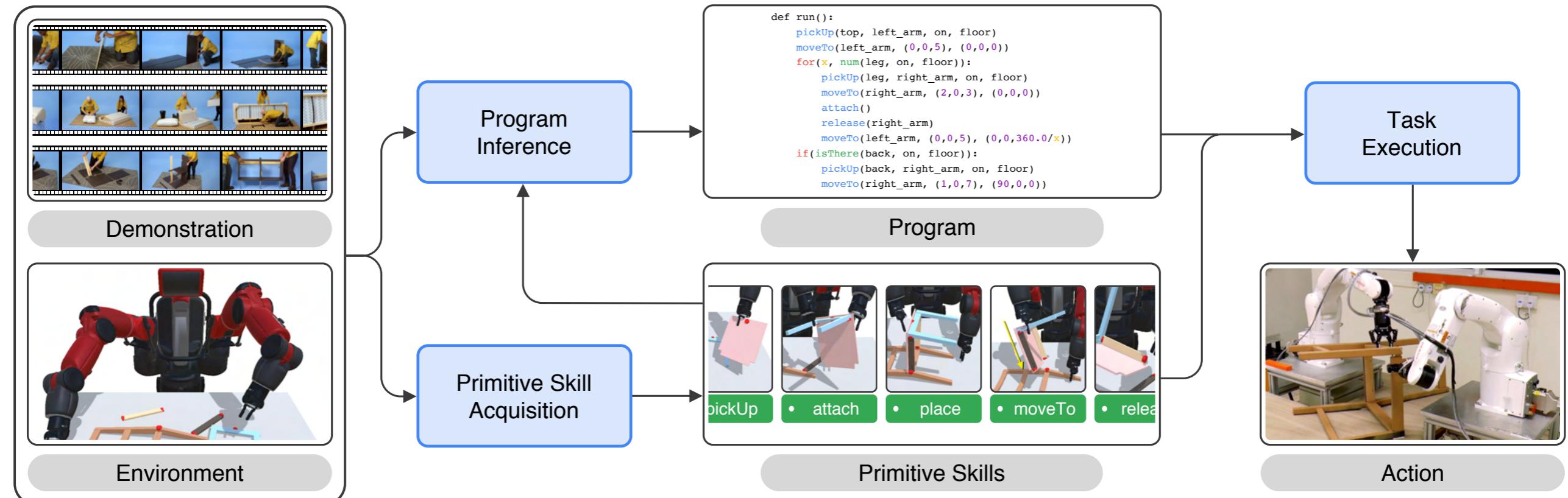


LfO   ◆ Ours   ▼ GAIfO-s   ◀ GAIfO   ◆ BCO   LfD   ● SQL   ■ GAIL   ✖ BC

LfD   ● SQL   ■ GAIL   ✖ BC



# Program-Guided Framework for Interpreting and Acquiring Complex Skills with Learning Robots



Interpretable

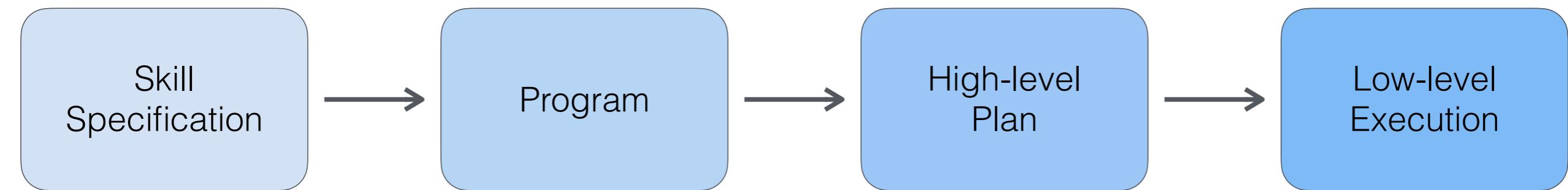
Programmatic /  
Generalizable

Hierarchical

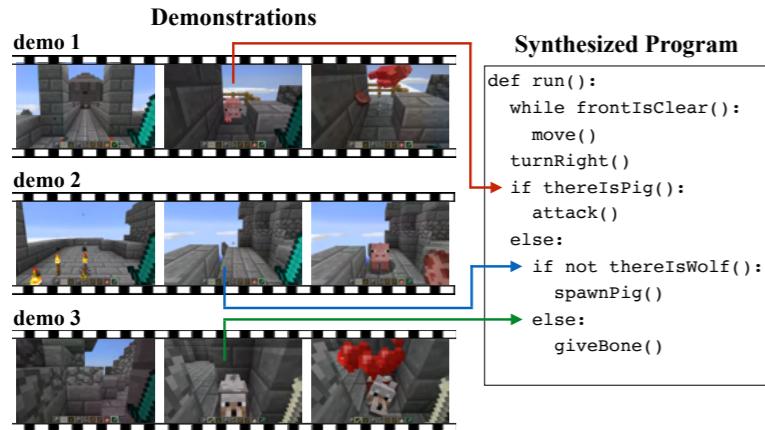
Modular

# Program-Guided Framework for Interpreting and Acquiring Complex Skills with Learning Robots

---



# Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

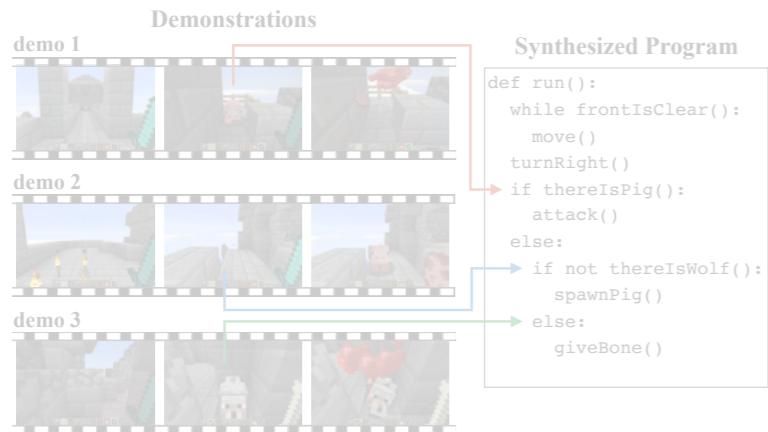
Skill  
Specification

Program

High-level  
Plan

Low-level  
Execution

## Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

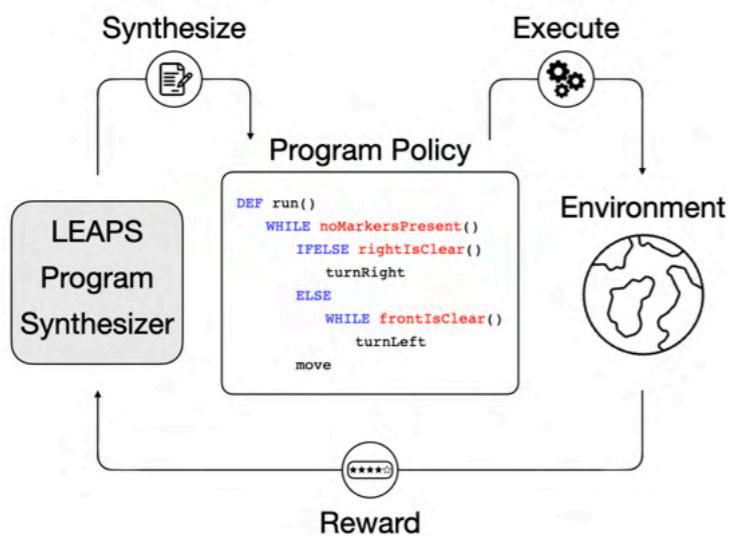
Skill Specification

Program

High-level Plan

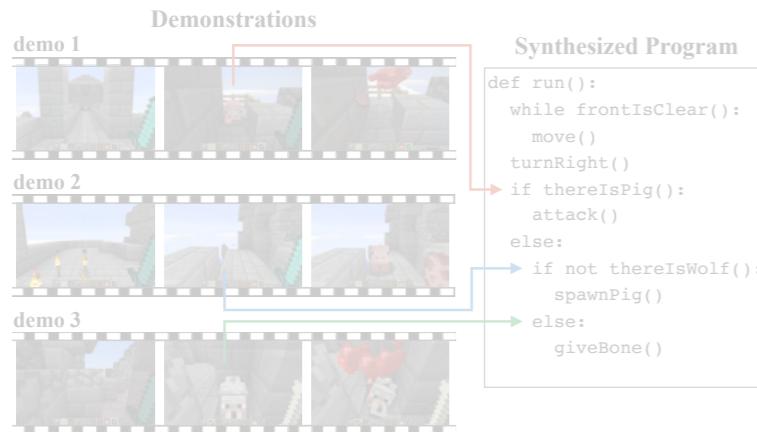
Low-level Execution

## Learning to Synthesize Programs as Interpretable and Generalizable Policies



NeurIPS 2021

## Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

Skill Specification

Program

High-level Plan

Low-level Execution

Learning to Synthesize Programs as  
Interpretable and Generalizable Policies



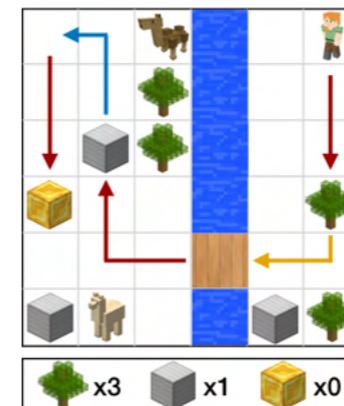
NeurIPS 2021

Program Guided Agent

**Program**

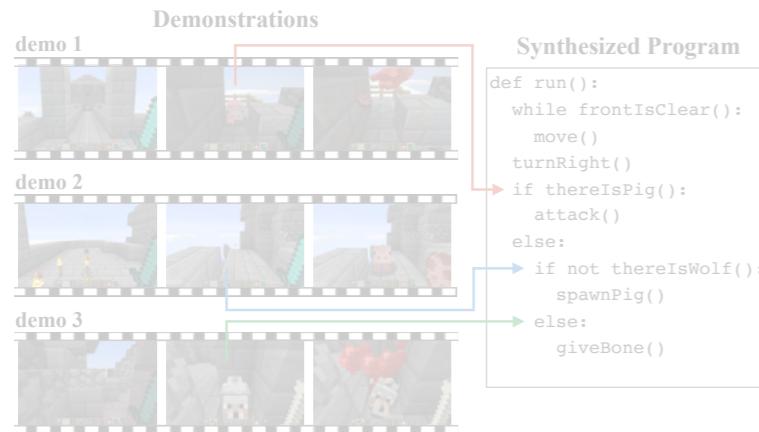
```

def Task():
    if is_there[River]:
        mine(Wood)
        build_bridge()
    if agent[Iron] < 3:
        mine(Iron)
        place(Iron, 2, 3)
    else:
        goto(4, 2)
    while env[Gold] > 0:
        mine(Gold)
    
```



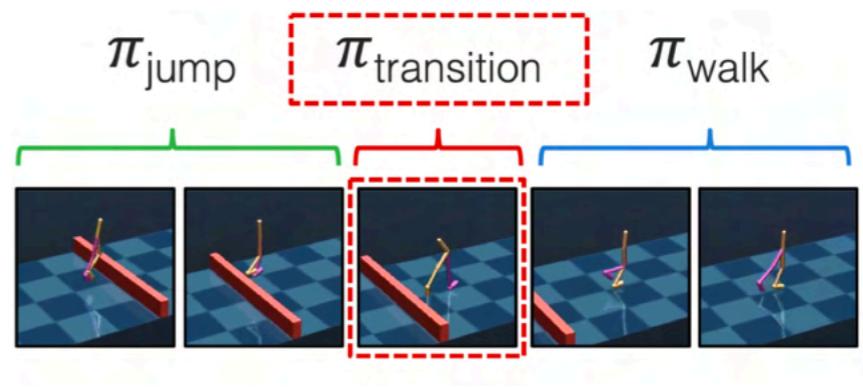
ICLR 2020 (Spotlight)

## Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

## Composing Complex Skills by Learning Transition Policies



ICLR 2019

Skill Specification

Program

High-level Plan

Low-level Execution

## Learning to Synthesize Programs as Interpretable and Generalizable Policies

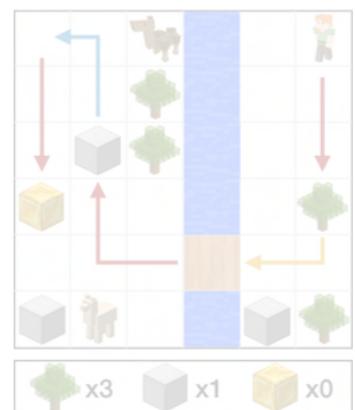


NeurIPS 2021

## Program Guided Agent

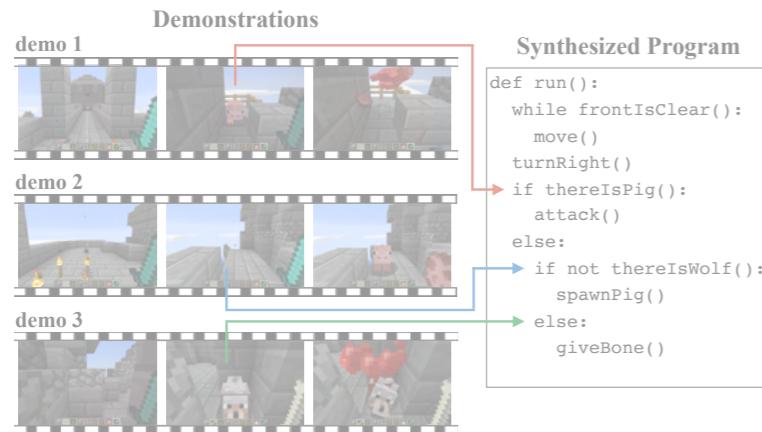
Program

```
def Task():
    if is_there[River]:
        mine(Wood)
        build_bridge()
    if agent[Iron] < 3:
        mine(Iron)
        place(Iron, 2, 3)
    else:
        goto(4, 2)
    while env[Gold] > 0:
        mine(Gold)
```

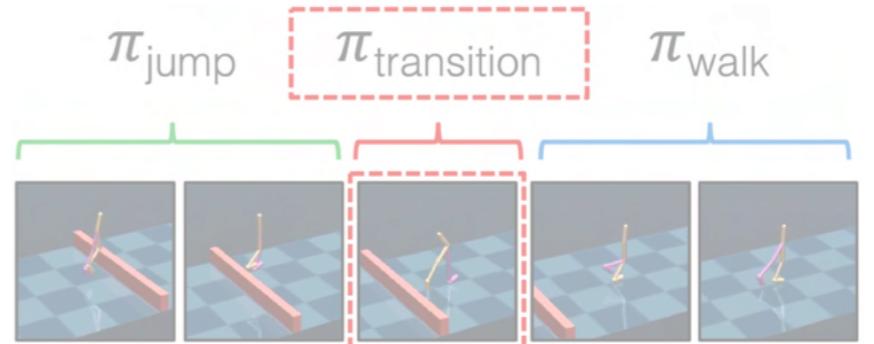


ICLR 2020 (Spotlight)

## Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018



ICLR 2019

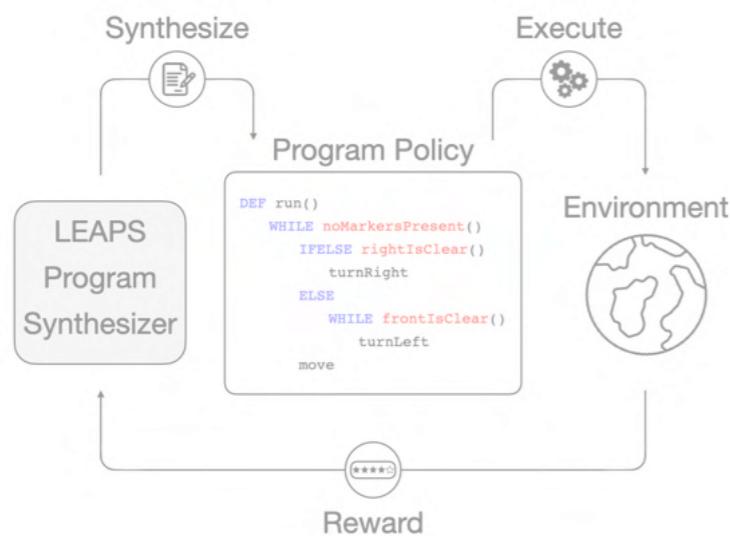
Skill Specification

Program

High-level Plan

Low-level Execution

## Learning to Synthesize Programs as Interpretable and Generalizable Policies

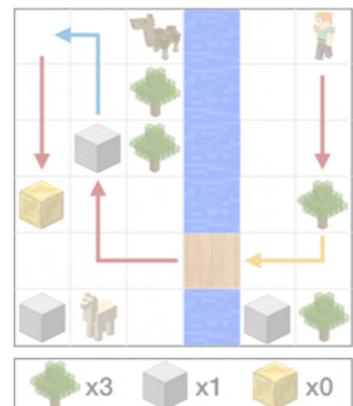


NeurIPS 2021

## Program Guided Agent

**Program**

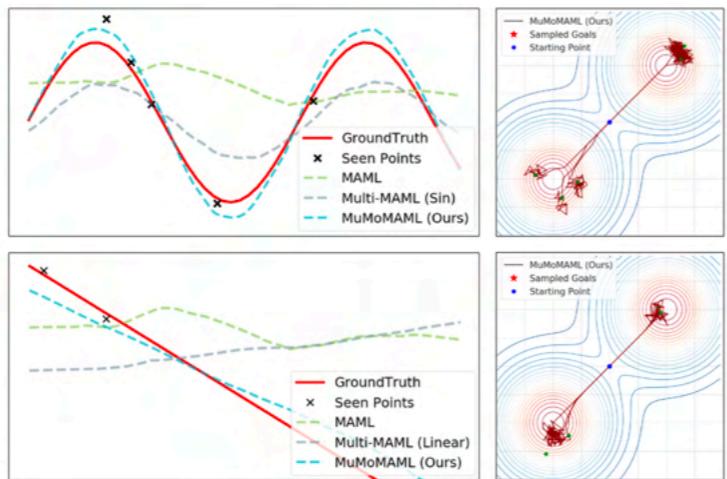
```
def Task():
    if is_there[River]:
        mine(Wood)
        build_bridge()
    if agent[Iron] < 3:
        mine(Iron)
        place(Iron, 2, 3)
    else:
        goto(4, 2)
    while env[Gold] > 0:
        mine(Gold)
```



ICLR 2020 (Spotlight)

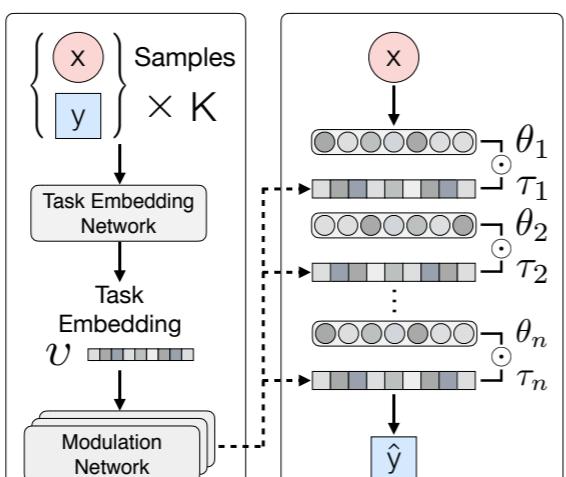
Low-level  
Execution

## Toward Multimodal Model-Agnostic Meta-Learning



Meta-learning workshop @ NeurIPS 2018

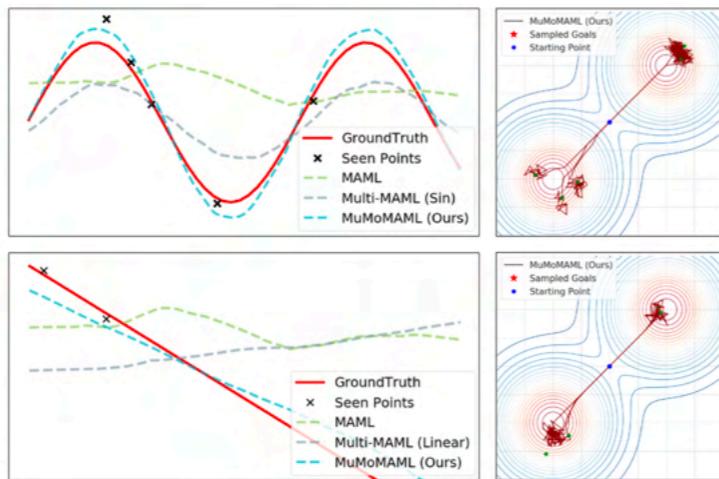
## Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation



NeurIPS 2019 (Spotlight)

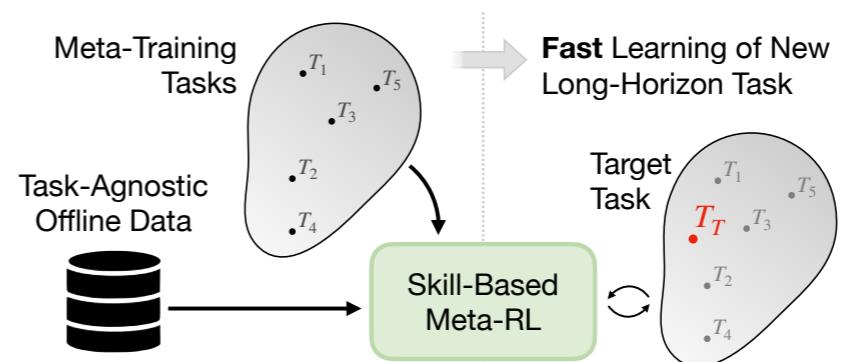
## Low-level Execution

### Toward Multimodal Model-Agnostic Meta-Learning



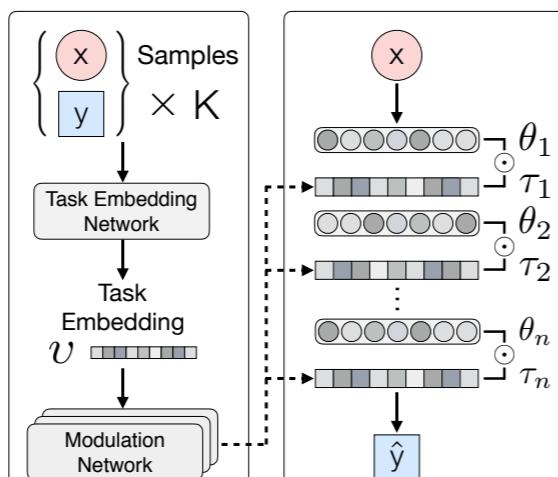
Meta-learning workshop @ NeurIPS 2018

### Skill-based Meta-Reinforcement Learning



ICLR 2022

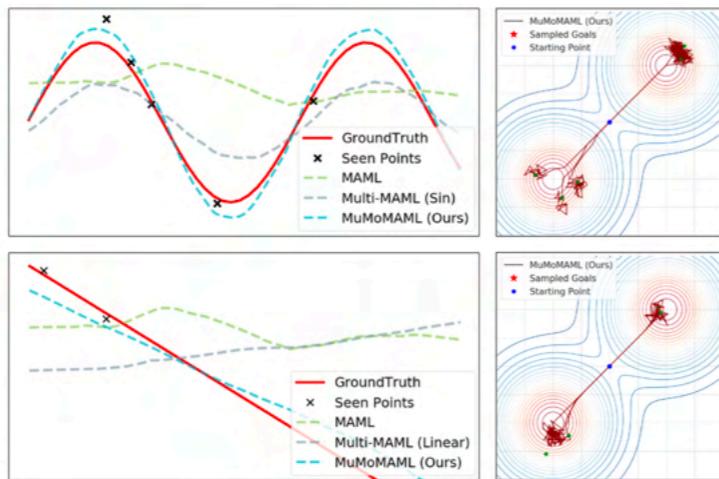
### Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation



NeurIPS 2019 (Spotlight)

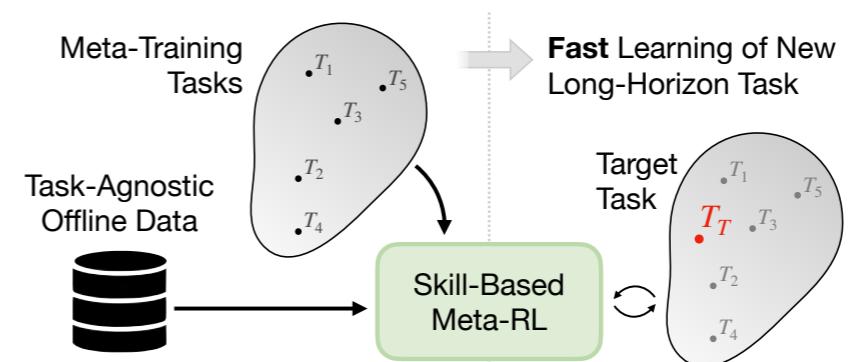
## Low-level Execution

### Toward Multimodal Model-Agnostic Meta-Learning



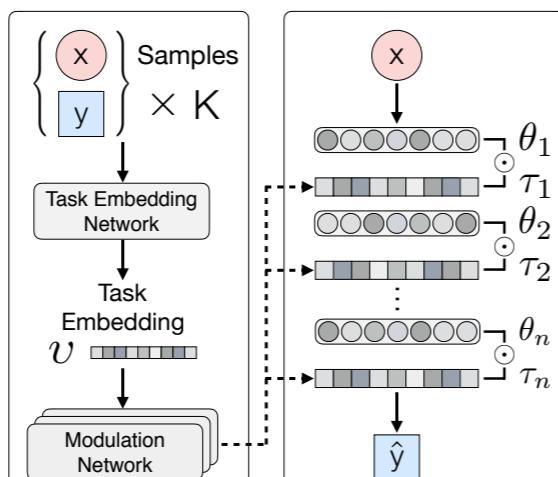
Meta-learning workshop @ NeurIPS 2018

### Skill-based Meta-Reinforcement Learning



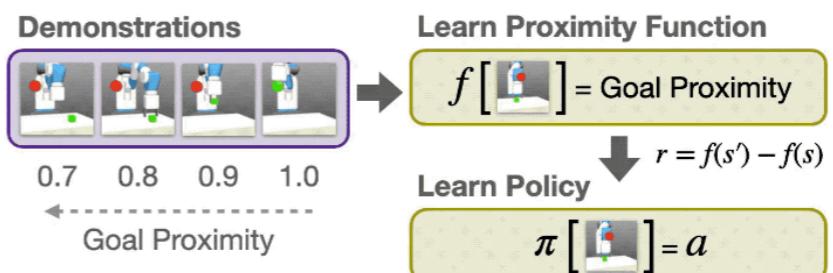
ICLR 2022

### Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation



NeurIPS 2019 (Spotlight)

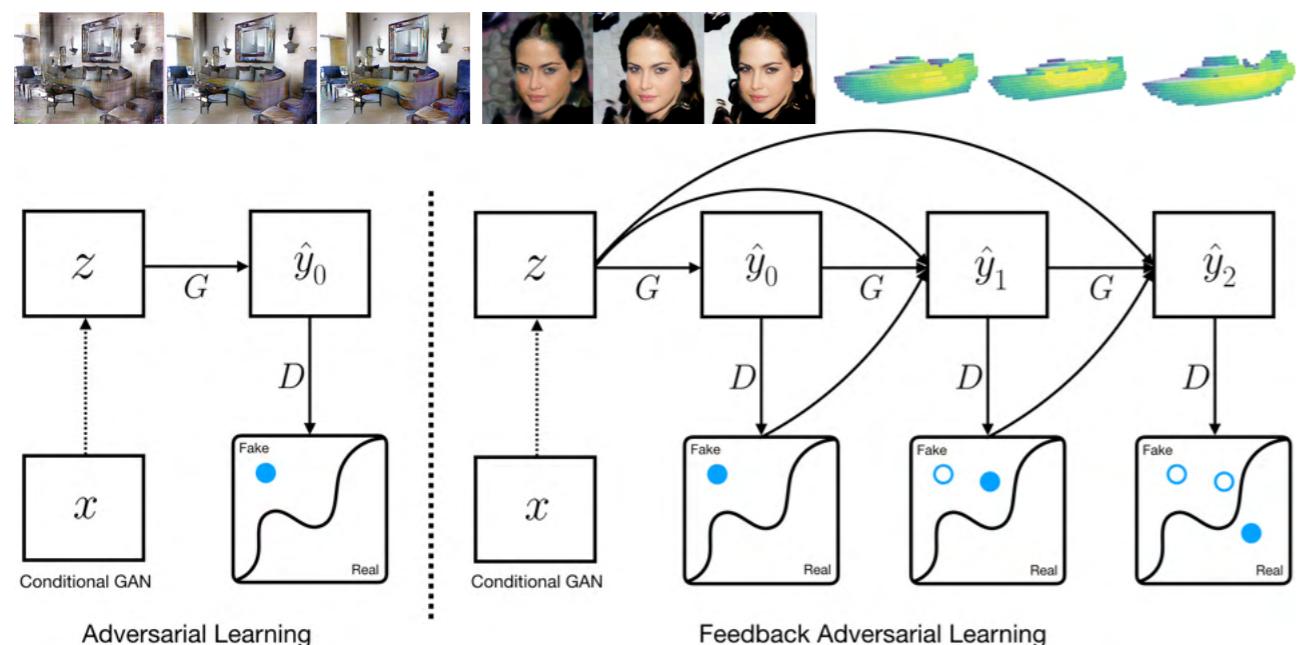
### Generalizable Imitation Learning from Observation via Inferring Goal Proximity



NeurIPS 2021

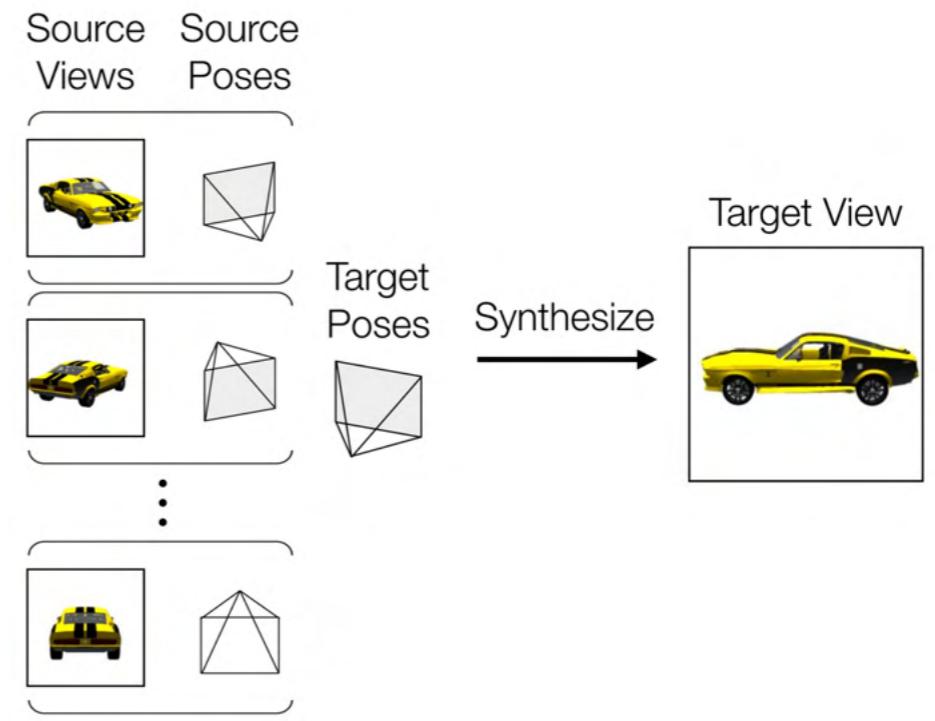
# Computer Vision / Image Synthesis / GAN

## Feedback Adversarial Learning: Spatial Feedback for Improving Generative Adversarial Networks

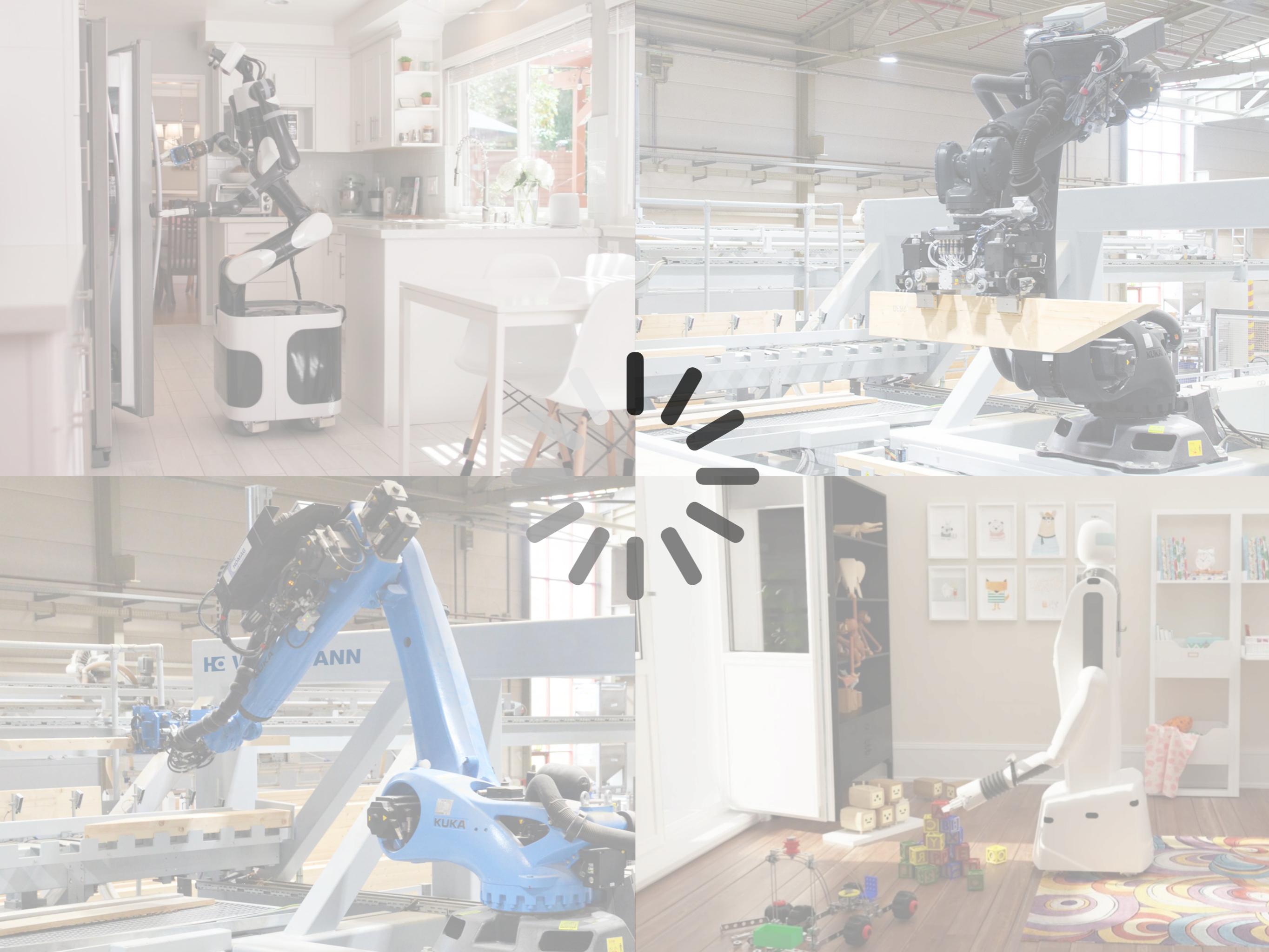


CVPR 2019

## Multi-view to Novel view: Synthesizing Views with Self-Learned Confidence



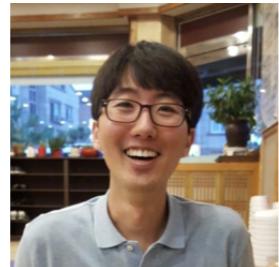
ECCV 2018





**Joseph J. Lim**

PI @ USC / CLVR



**Youngwoon Lee**

PhD @ USC / CLVR



**Karl Pertsch**

PhD @ USC / CLVR



**Jesse Zhang**

PhD @ USC / CLVR



**Minyoung Huh**

PhD @ MIT  
with Prof. Pulkit Agrawal &  
Prof. Phillip Isola



**Hexiang Hu**

Research Scientist  
@ Google



**Hyeonwoo Noh**

Research Scientist  
@ OpenAI



**Ning Zhang**

Research Scientist  
@ Facebook



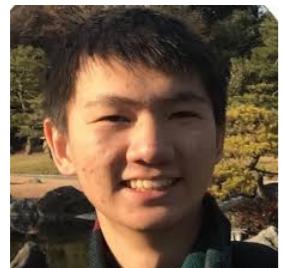
**Yuan-Hong Liao**

PhD @ U of T  
with Prof. Sanja Fidler



**Risto Vuorio**

PhD @ Oxford  
with Prof. Whiteson



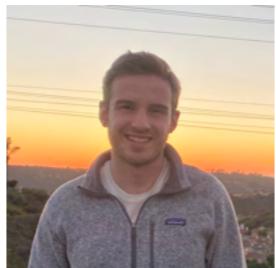
**Edward Hu**

PhD @ UPenn with Prof.  
Jayaraman



**Te-Lin Wu**

PhD @ UCLA  
with Prof. Nanyun Peng



**Andrew Szot**

PhD @ Georgia Tech  
with Prof. Zsolt Kira & Prof.  
Dhruv Batra



**Sung Ju Hwang**

Associate Professor  
@ KAIST



**Taewook Nam**

PhD @ KAIST  
with Prof. Sung Ju Hwang



**Dweep Trivedi**

Visitor @ USC / CLVR



**Sriram Somasundaram**

MS @ Stanford

# Thank You

---

## Questions?