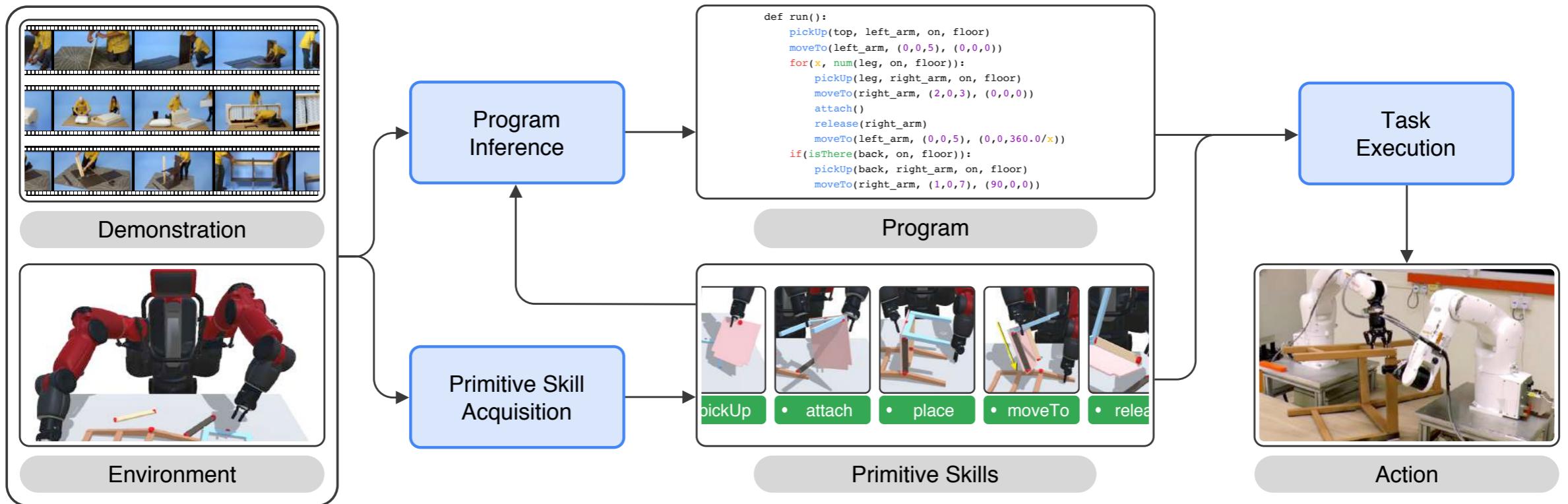
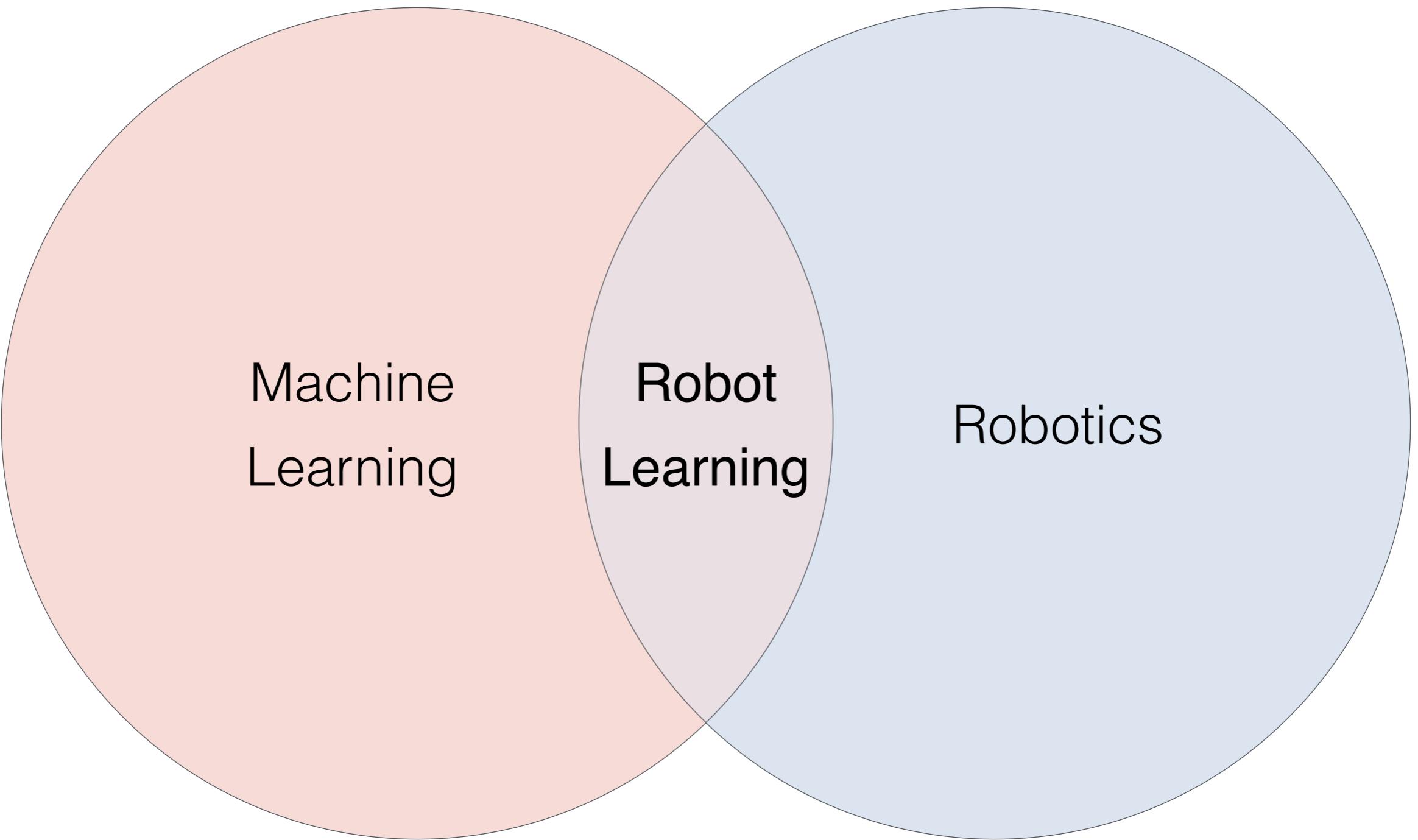


# Program-Guided Framework for Interpreting and Acquiring Complex Skills with Learning Robots



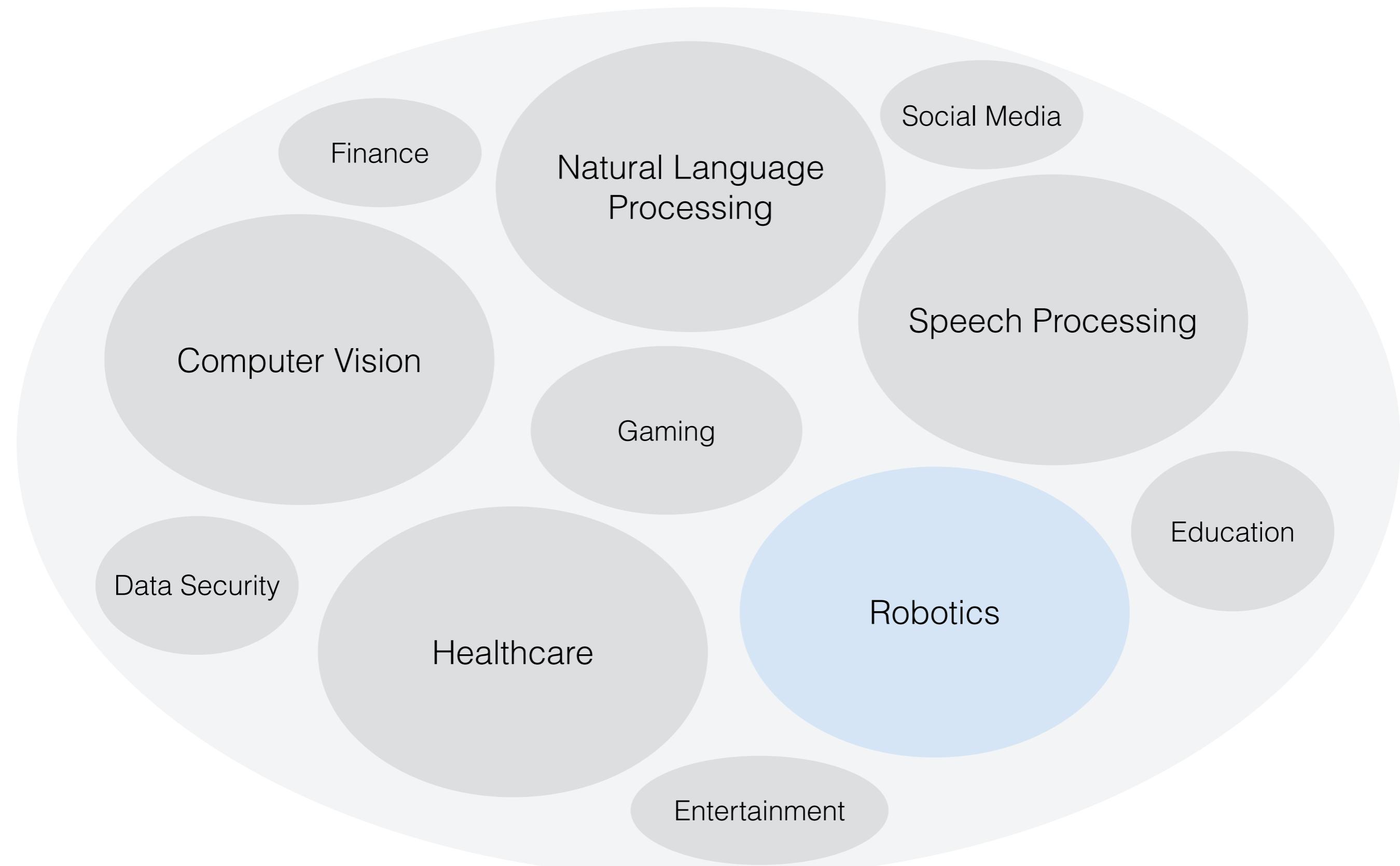
Shao-Hua Sun (孫紹華)

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



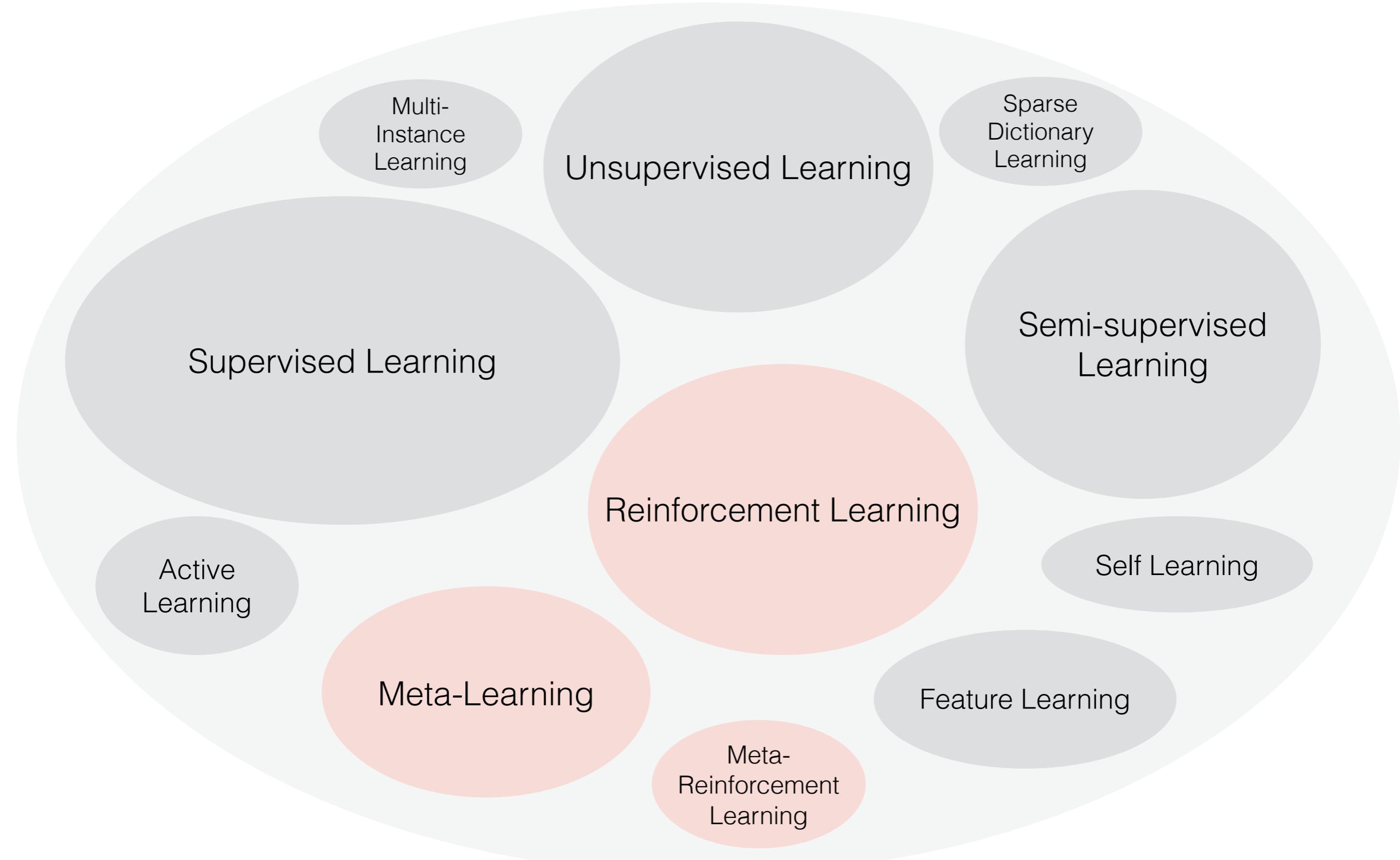
# Applications of AI

---



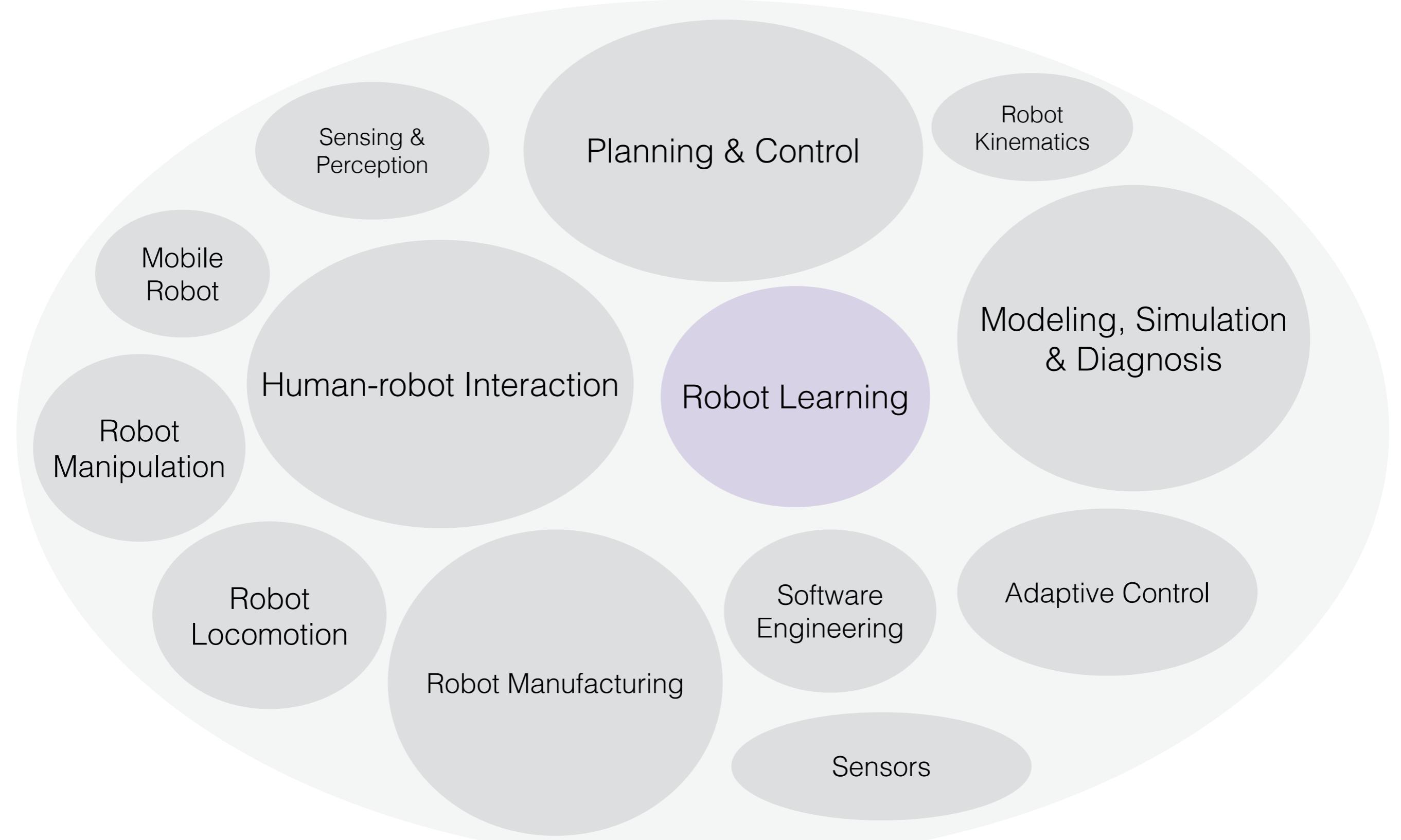
# Learning Problems

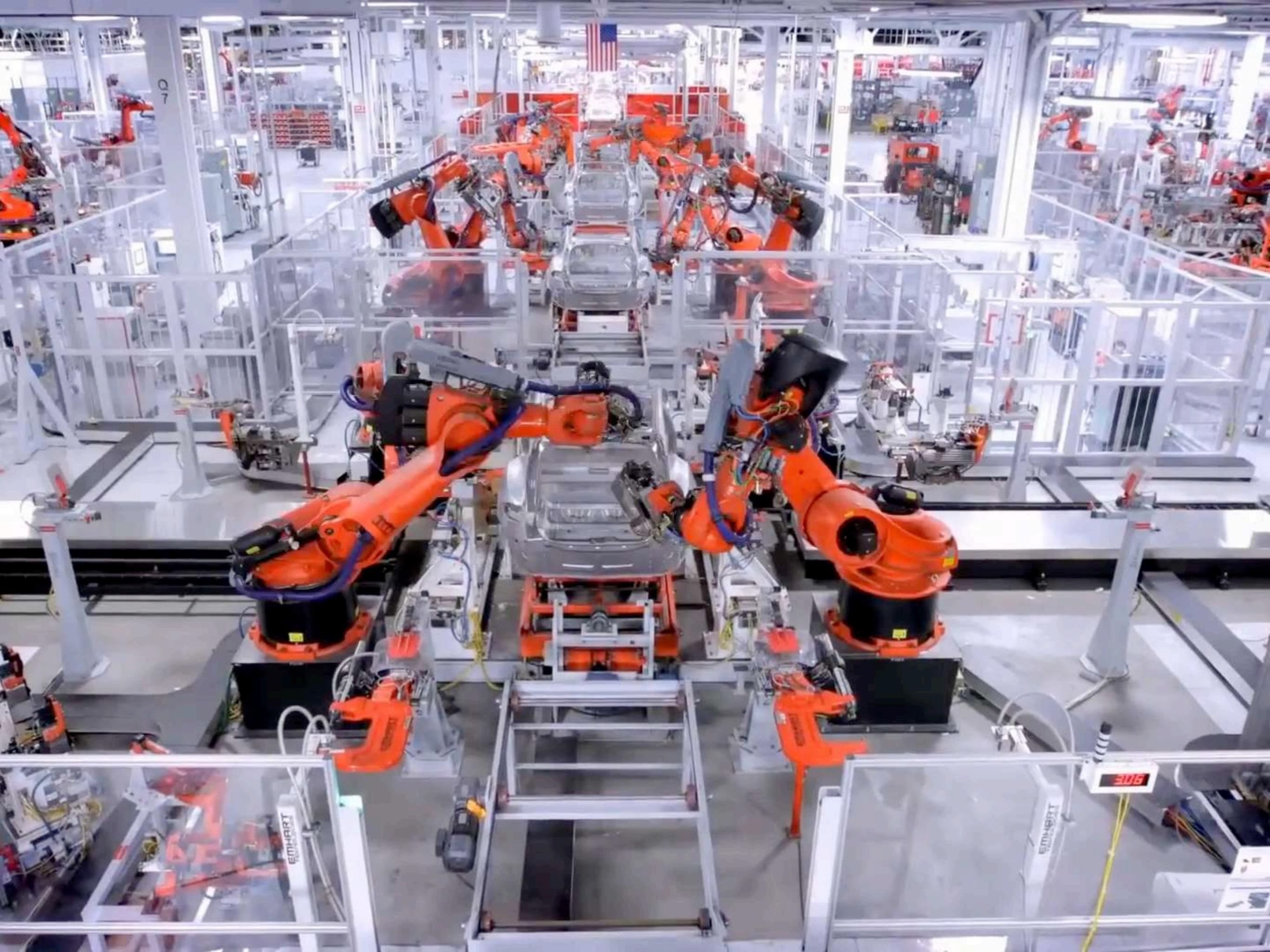
---



# Robotics

---



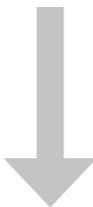


# Robot Learning

## Environment

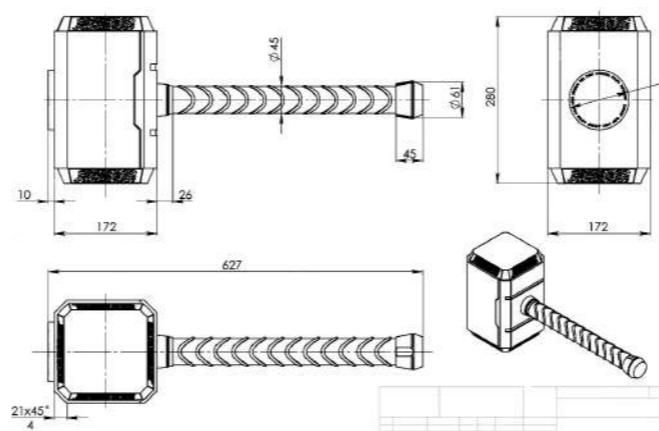


Structured



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]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	299×299	<b>10.9 M</b>	<b>2.35 B</b>	<b>78.6</b>	<b>94.2</b>
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	<b>80.8</b>	<b>95.3</b>
ResNeXt-101 (64 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	320×320	<b>145.8 M</b>	<b>42.3 B</b>	<b>82.7</b>	<b>96.2</b>
NASNet-A (6 @ 4032)	331×331	<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. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image size reported in the table. Model size for [25] calculated from open-source implementation.

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Instance Segmentation



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

## Visual Question Answering

Method	VQA v2 test-dev			VQA v2 test-old				
	All	Yes/no	Number	Other	All	Yes/no	Number	Other
Proposed model	—	—	—	—	25.08	24.98	25.17	25.17
Proposed model (without visual reasoning)	—	—	—	—	44.26	47.01	31.55	27.37
LSTN Language only (fixed model) [1]	—	—	—	—	54.23	73.08	35.18	41.83
DeepLSTM Q & V (as reported in [1])	—	—	—	—	42.27	74.62	38.28	33.38
MCB [1] (as reported in [1])	—	—	—	—	65.73	62.07	61.00	37.02
USPM-CAPS [1]	—	—	—	—	67.77	70.00	40.57	40.57
DRONIC	—	—	—	—	58.77	31.80	46.29	38.30
RCNN-UNISD-UNIC	—	—	—	—	68.18	54.51	45.30	70.01
Proposed model	—	—	—	—	62.07	76.20	36.46	52.62
ResNet features T=7, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
Image features from bottom up attention, adaptive K, single network	66.38	83.38	41.17	57.41	66.73	83.71	43.77	57.20
ResNet features T=7, ensemble	69.87	86.88	46.99	68.86	76.74	94.00	46.64	61.35

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

He et al. Mask R-CNN

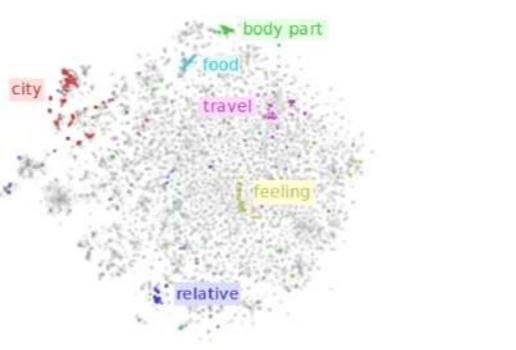
Teney et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

## 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 compétitive avec nos produits", a déclaré Kevin M. Keniston, chef du confort des passagers de l'Airbus de l'Europe.
GNMT	"La raison pour laquelle Boeing fait cela est de créer plus de sièges pour rendre son avion plus compétitif avec nos produits", a déclaré Kevin Keniston, chef du confort des passagers chez Airbus.
Human	"Boeing fait ça pour pouvoir 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 États-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 États-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

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

## Word Embeddings



On word embeddings - Part 1 by Ruder

## Named Entity Recognition

covertSki to site indexPoliticsSubscribeLogInSubscribeLogInToday's **PageAdvertisementSupported** **ORG** byf B.I. Agent **Peter Strzok** **PERSON** . Who Collected Trump **PERSON** in Texts, Is Freedmag/Peter Strzok, a top counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President **Trump** **PERSON** were uncovered, was fired. **Craig J. Keehan** **XX PERSON** for **The New York Times** Adam Goodman **ORG** and Michael S. Schmidt **ORG** **PERSON** **13 CARDINAL** , 2016WASHINGTON **CARDINAL** , **POLITICIAN** **PERSON** , senior counterintelligence agent who disparaged President **Trump** **PERSON** in inflammatory text messages and helped oversee the **Hillary Clinton** **PERSON** email and **Russia** **GPE** investigation, has been fired for violating bureau policies, Mr. **Strzok** **PERSON** , a lawyer for Monday DATE Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former **FBI** **GPE** lawyer. Lisa Page — b. PERSON, assuring the **Russia** **GPE** investigation as an illegitimate "witch hunt." Mr. **Strzok** **PERSON** , who rose over 20 years inquiry along with writing the texts, Mr. **Strzok** **PERSON** was accused of sending a highly sensitive search warrant to his personal email account. The **FBI** **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 on Twitter using Images and Text

## Question Answering

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

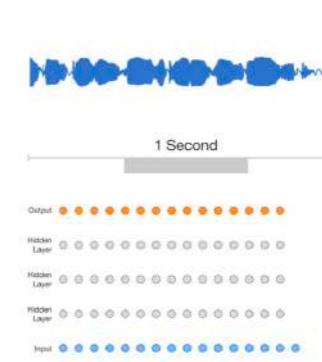
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## Speech Recognition

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-puhpu	BLSTM	11.3	8.0
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpu	BLSTM+ResNet+LACE	9.7	7.4
27k	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3
-	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 Dord 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	96.2
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



Figure 5: Mask 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).

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

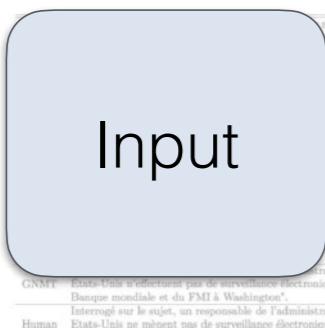
## Visual Question Answering

Method	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual	Nodes	All	Visual	Nodes
Other cross-domain models in training sets [-]	-	-	-	25.08	27.00	27.07
LSTM Language only (old model) [-]	-	-	-	44.26	47.40	55.85
Deep LSTM (Quoc et al. [27] or reported in [-])	-	-	-	54.22	73.00	58.18
MCB [-] (as reported in [-])	-	-	-	42.27	78.42	38.29
SPM+LSTM-FPN [-]	-	-	-	65.73	63.07	61.00
DRNA [-]	-	-	-	67.57	61.80	50.47
DRNA+FCN [-]	-	-	-	68.77	61.80	50.39
DRNA+UNet-UNNC [-]	-	-	-	68.18	64.30	52.20
Proposed model	-	-	-	68.18	64.30	52.20
Baseline features T=7, single network	42.07	76.20	56.48	52.62	42.27	79.32
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T=7, ensemble	66.38	81.38	43.17	57.41	66.73	81.71
Image features from bottom-up attention, adaptive K, ensemble	69.87	86.00	46.89	58.64	70.74	86.64

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

## Machine Translation



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

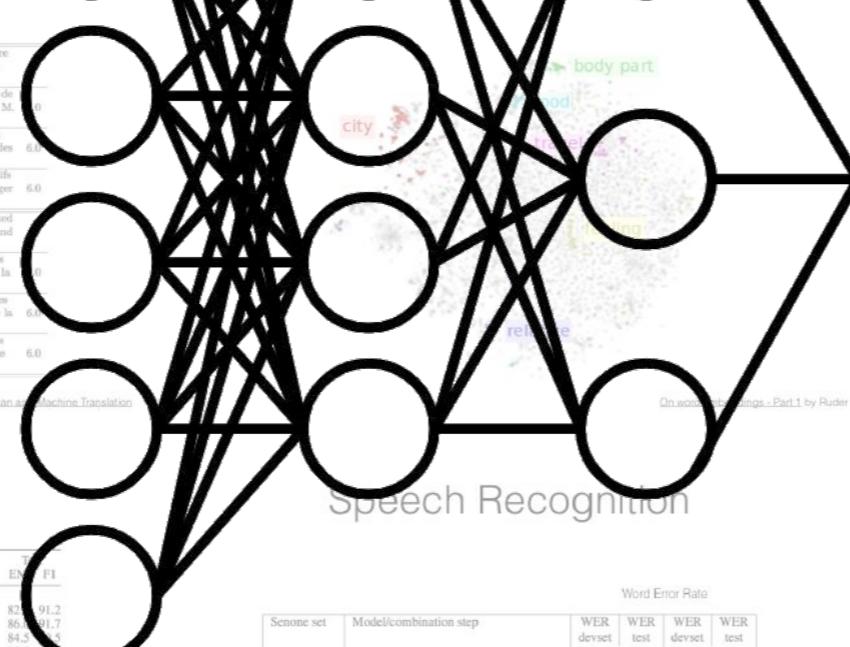
## Question Answering

System	Dev EM	Test F1	Dev F1
Leaderboard (Oct 8th, 2018)			
Human	-	82	91.2
#1 Ensemble - nlnet	-	86	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
R.M. Reader (Ensemble)	81.2	87.9	82.3
Ours	80.8	88.5	-
BERTBASE (Single)	84.1	90.9	-
BERTLARGE (Single)	85.8	91.8	-
BERTLARGE (Ensemble)	84.2	91.1	85.1
BERTLARGE (Sgl.+TriviaQnA)	84.2	91.1	91.8
BERTLARGE (Ens.+TriviaQnA)	86.2	92.2	87.4
Ours	85.2	92.3	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 some challenges - Part 1 by Ruder

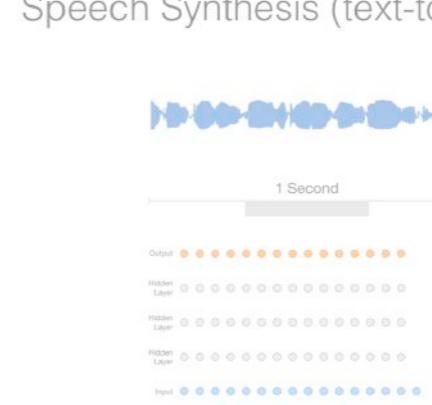
## Named Entity Recognition



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

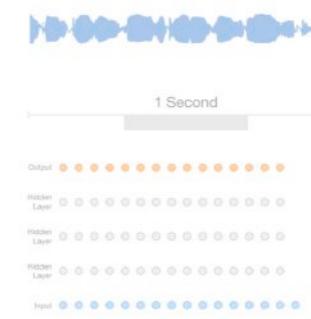
## Speech Recognition

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-puhpuh	BLSTM	11.3	8.0
9k-puhpuh	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpuh	BLSTM+ResNet+LACE	9.7	7.4
27k	BLSTM+ResNet+LACE	9.7	7.3
-	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

# 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	96.2
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 each of the models. Model sizes are from [25] calculated from open-source implementation.

## Instance Segmentation



Figure 5: Mask 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 v1 test-Set			VQA v2 test-Set		
	All	Visual	Nodes	All	Visual	Nodes
Other cross-domain models in training set (–)	—	—	—	25.08	27.00	27.07
LSTM Language only (old model) (–)	—	—	—	44.26	47.40	51.85
Deep LSTM Q-vqa (1.27) (as reported in [1])	—	—	—	54.22	73.04	58.18
MCB (1.28) (as reported in [1])	—	—	—	62.27	78.42	58.29
SPM+CAPS (1.3)	—	—	—	65.73	82.07	61.00
DRQA (1.3)	—	—	—	67.95	84.40	61.07
DRNA (1.3)	—	—	—	68.77	81.80	58.59
RCNN-UNITER-UNICOC (1.3)	—	—	—	69.18	84.30	62.20
Proposed model	—	—	—	69.21	79.32	76.77
Baseline features T=7, single network	42.07	76.20	56.44	52.62	42.21	79.32
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T=7, ensemble	66.38	83.38	43.17	57.41	66.73	83.71
Image features from bottom-up attention, adaptive K, ensemble	69.07	86.00	46.09	58.64	70.74	86.64

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

Image



## Question Answering

### Input Question:

Where do water droplets collide with ice crystals to form precipitation?

### Input Paragraph:

... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. ...

### Output Answer:

within a cloud

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				
BERTBASE (Single)	80.8	88.5	-	-
BERTLARGE (Single)	84.1	90.9	-	-
BERTLARGE (Ensemble)	85.8	91.8	-	-
BERTLARGE (Sgl.+TriviaQnA)	84.2	91.1	85.1	91.8
BERTLARGE (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

## Speech Recognition

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

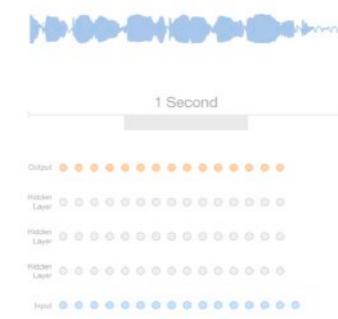
Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

## 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

# 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	95.5
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 multiple-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image size reported in the table. Model size for [25] calculated from open-source implementation.

## Instance Segmentation



Figure 5: Mask 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).

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Model	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual	Nodes	All	Visual	Nodes
Other most common answer in training set [-]	-	-	-	25.08	27.00	27.07
LSTM Language only (old model) [-]	-	-	-	44.26	47.40	55.85
Deep LSTM (Q score 1.07) (as reported in [-])	-	-	-	54.22	73.00	58.18
MCB [-] (as reported in [-])	-	-	-	62.27	78.42	58.29
SPMEL-FPN [-]	-	-	-	65.73	63.07	61.00
DRNA [-]	-	-	-	67.77	64.00	59.07
DRNA	-	-	-	68.77	61.80	59.39
DRN-UNet-UNetC	-	-	-	69.18	64.30	59.20
Proposed model	-	-	-	70.01	-	-
Baseline features T=7, single network	42.07	76.20	56.44	52.62	42.27	79.32
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T=7, ensemble	66.38	81.38	43.17	57.41	66.73	81.71
Image features from bottom-up attention, adaptive K, ensemble	69.07	86.00	46.00	58.64	70.34	86.64

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

## English sentence

France is never cold  
in September

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".

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

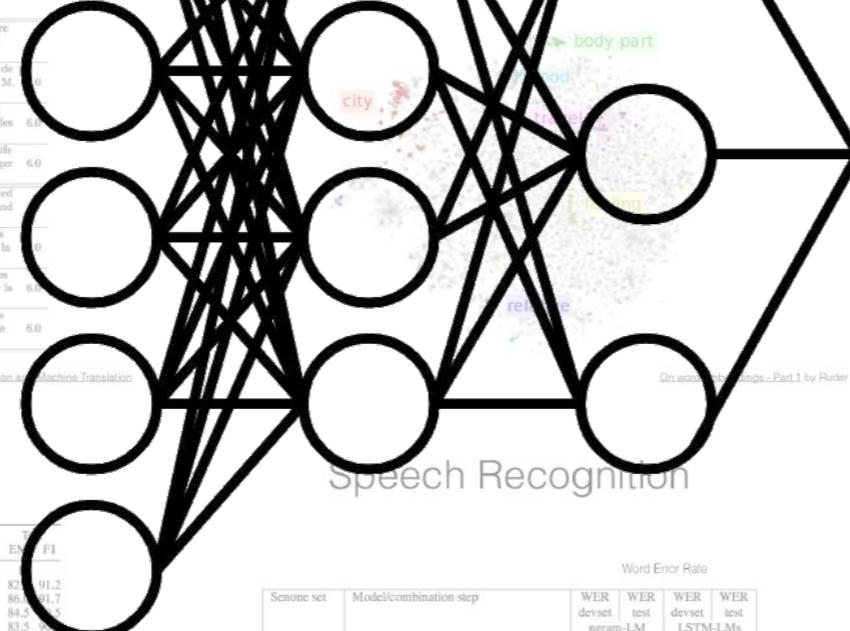
## Question Answering

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	-	-	-	-
BERTBASE (Single)	80.8	88.5	-	-
BERTLARGE (Single)	84.1	90.9	-	-
BERTLARGE (Ensemble)	85.8	91.8	-	-
BERTLARGE (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
BERTLARGE (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



Word Error Rate

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

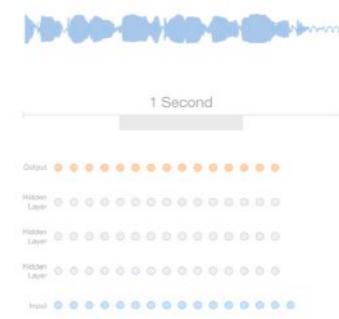
## Speech Recognition

## 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 Synthesis (text-to-speech)



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



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).

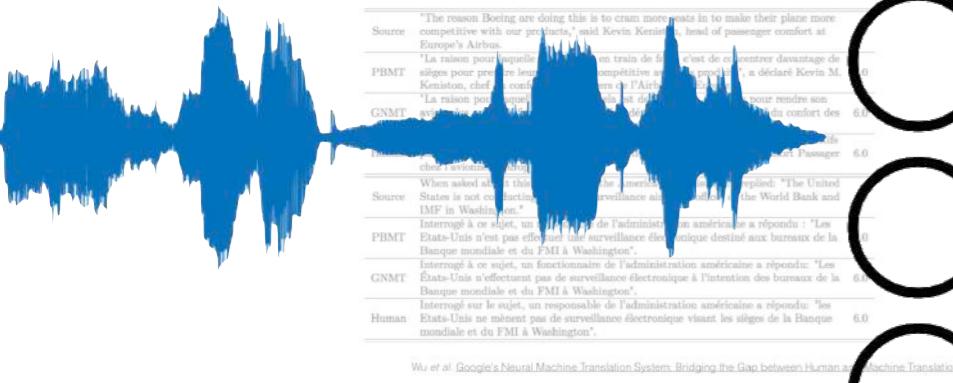
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

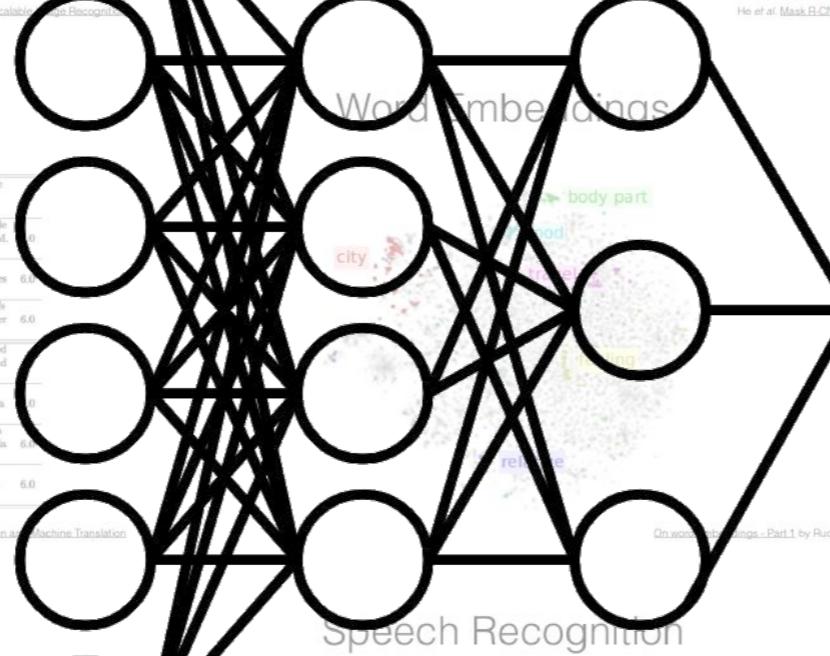
Method	VQA v2 test-Dev			VQA v2 test-Old		
	All	Visual	Nodes	All	Visual	Nodes
Other cross-domain models in training set (–)	—	—	—	25.06	37.00	35.85
LSTM Language only (old model) (–)	—	—	—	44.26	47.40	35.85
Deep LSTM Q-vqa (1.27) (as reported in [1])	—	—	—	54.22	73.00	38.18
MCB (1.25) (as reported in [1])	—	—	—	62.27	78.42	38.28
SPM+CAPS (1.25)	—	—	—	65.73	82.07	61.00
DRNA	—	—	—	67.57	84.40	61.07
DRNA	—	—	—	68.77	81.80	62.59
DRNA-UNID-UNIC	—	—	—	69.18	84.30	62.20
Proposed model	—	—	—	69.21	79.32	70.01
Baseline features T=7, single network	42.07	76.20	36.48	52.62	62.27	36.77
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T=7, ensemble	66.38	83.38	43.17	57.41	66.73	83.71
Image features from bottom-up attention, adaptive K, ensemble	69.07	86.00	46.89	68.64	69.35	84.35

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

## Waveform Translation



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



## Text

This is a supervised learning method

## Question Answering

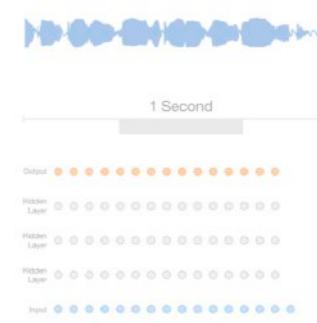
System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)	-	-	-	-
Human	-	-	82.1	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#2 Single - QANet	-	-	82.5	89.3
Published	-	-	-	-
BIDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	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>LAIRG</sub> (Single)	84.1	90.9	-	-
BERT <sub>LAIRG</sub> (Ensemble)	85.8	91.8	-	-
BERT <sub>LAIRG</sub> (Sgl.+TriviaQnA)	84.2	91.1	85.1	91.8
BERT <sub>LAIRG</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.

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-puhpuh	BLSTM	11.3	8.0	9.2	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpuh	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpuh	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8
-	Confusion network combination	-	-	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

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

# 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	96.2
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



Figure 5: Mask 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).

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

He et al. Mask R-CNN

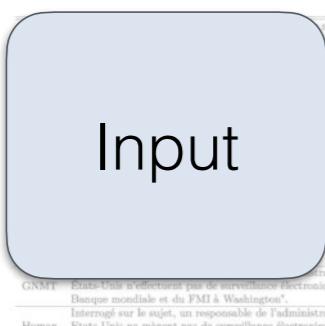
## Visual Question Answering

Method	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual	Nodes	All	Visual	Nodes
Other most common answer in training set [-]	-	-	-	25.08	27.00	27.07
LSTM Language only (old model) [-]	-	-	-	44.26	47.40	55.85
Deep LSTM (Quoc et al. [27] or reported in [1])	-	-	-	54.22	73.00	58.18
MCB [-] (as reported in [1])	-	-	-	62.27	78.42	58.29
SPM+LSTM-FPN [-]	-	-	-	65.73	63.07	61.00
DRNA [-]	-	-	-	67.57	61.80	50.47
DRNA+FCN [-]	-	-	-	68.77	61.80	50.39
DRNA+UNet-UNNC [-]	-	-	-	68.18	64.30	52.20
Proposed model	-	-	-	68.18	64.30	52.20
Baseline features T=7, single network	42.07	76.20	36.48	52.62	42.27	79.32
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T=7, ensemble	66.38	81.38	43.17	57.41	66.73	81.71
Image features from bottom-up attention, adaptive K, ensemble	69.87	86.00	46.89	68.64	70.35	82.39

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

## Machine Translation



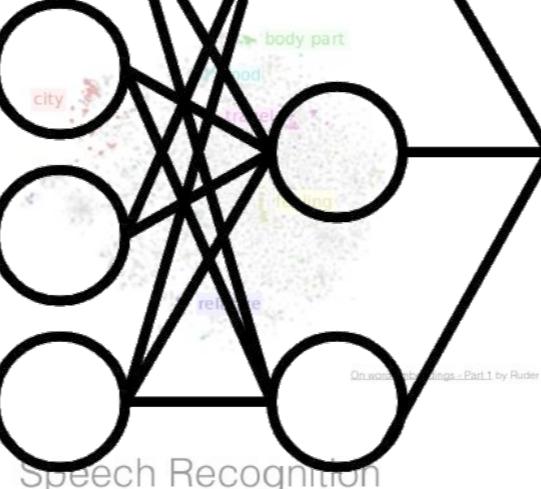
## Question Answering

System	Dev EM	Test F1	EM	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				
BERTBASE (Single)	80.8	88.5	-	-
BERTLARGE (Single)	84.1	90.9	-	-
BERTLARGE (Ensemble)	85.8	91.8	-	-
BERTLARGE (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
BERTLARGE (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

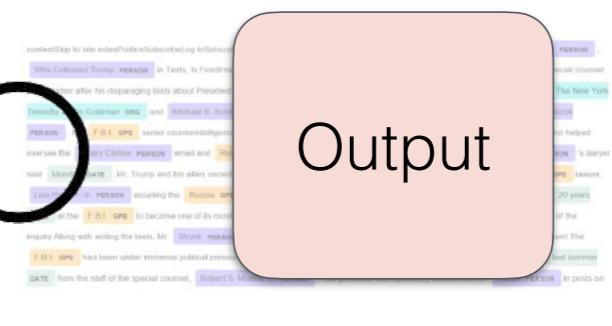


Senone set	Model/combination step	WER		WER	
		devset	test	devset	test
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.2	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	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

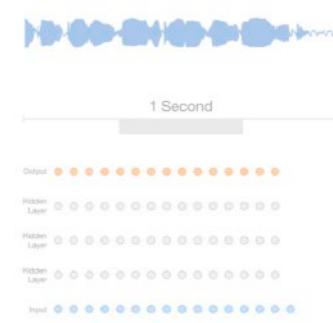
## Speech Recognition

## 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

# Supervised Learning

## Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Jarrett et al. [27]	224x224	~1.0M	~1.0B	74.8	92.2
NASNet-A (5@15.38)	299x299	19.9M	2.35 B	78.6	94.2
Inception V3 [50]	299x299	10.6M	5.72 B	78.8	94.4
Xception [9]	299x299	10.6M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299x299	10.6M	13.2 B	80.1	95.1
NASNet-A (7@1920)	299x299	10.6M	4.93 B	80.8	95.3
PolyNet [69]	331x331	92 M	34.7 B	80.9	95.6
DPN-131 [8]	329x320	79.5 M	32.0 B	81.5	95.7
SENet [25]	320x320	145.8 M	42.3 B	82.7	96.0
NASNet-A (6@4032)	331x331	88.9 M	23.8 B	82.7	96.2

Table 1. 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 perform a forward pass. Note that some models may contain multiple accumulate operations are calculated for the image size reported in the paper. Model size for 128 is calculated from open-source implementation.

## Instance Segmentation



Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Method	VQA v1 test-dev			VQA v2 test-dev		
	All	Visual	Nodes	All	Visual	Nodes
Other cross-domain models in training set (1)	—	—	—	25.08	27.00	27.07
LSTM Language only (old model) (1)	—	—	—	44.26	47.40	55.85
Deep LSTM Q score (1) (as reported in [1])	—	—	—	54.22	73.00	58.18
MCB (1) (as reported in [1])	—	—	—	62.27	78.42	58.29
SPMC-LPN (1)	—	—	—	65.73	63.07	61.00
DRNA (1)	—	—	—	67.77	61.80	53.07
DRNA (2)	—	—	—	68.77	61.80	53.09
DRNA-UNet-UNNC	—	—	—	69.18	64.30	55.20
Proposed model	—	—	—	70.01	70.20	70.01
Baseline features T+7, single network	42.07	76.20	36.48	52.62	42.21	79.32
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T+7, ensemble	66.38	81.38	43.17	57.41	66.73	81.71
Image features from bottom-up attention, adaptive K, ensemble	69.07	86.00	46.86	76.74	76.64	81.55

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



Input

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".

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	-	-	-	-
BERTBASE (Single)	80.8	88.5	-	-
BERTLARGE (Single)	84.1	90.9	-	-
BERTLARGE (Ensemble)	85.8	91.8	-	-
BERTLARGE (Sgl.+TriviaQAn)	84.2	91.1	85.1	91.8
BERTLARGE (Ens.+TriviaQAn)	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.

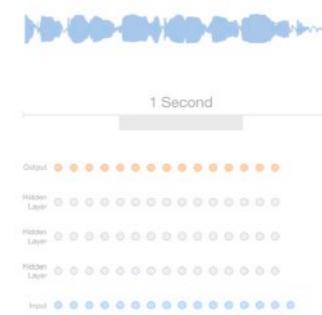
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## 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-puhpuh	BLSTM	11.3	8.0
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpuh	BLSTM+ResNet+LACE	9.7	7.4
9k-puhpuh	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

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. (%)
Jarrett et al. [27]	224x224	~1.0M	~1.0B	74.8	92.2
NASNet-A (5@ 15.38)	299x299	19.9M	2.35 B	78.6	94.2
Inception V3 [50]	299x299	10.6M	5.72 B	78.8	94.4
Xception [9]	299x299	10.6M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299x299	10.6M	13.2 B	80.1	95.1
NASNet-A (7@ 1920)	299x299	10.6M	4.93 B	80.8	95.3
PolyNet [69]	331x331	92 M	34.7 B	80.9	95.6
DPN-131 [8]	329x320	79.5 M	32.0 B	81.5	95.7
SENet [25]	320x320	145.8 M	42.3 B	82.7	96.0
NASNet-A (6@ 4032)	331x331	88.9 M	23.8 B	82.7	96.2

Table 1. 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 perform a forward pass. Note that some models may contain multiple accumulate operations are calculated for the image size reported in the paper. Model size for 128 is calculated from open-source implementation.

## Instance Segmentation



Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

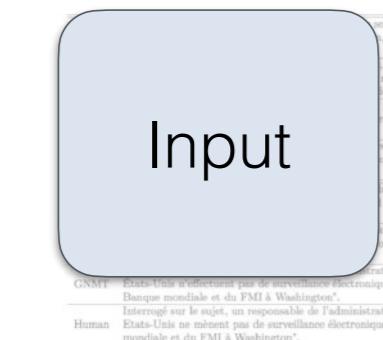
Method	VQA v1 test-dev			VQA v2 test-dev		
	All	Visual	Nodes	All	Visual	Nodes
Other cross-domain models in training set (1)	—	—	—	25.08	27.00	27.07
LSTM Language only (old model) (1)	—	—	—	44.26	47.40	55.85
Deep LSTM Q-vqa (1) (as reported in [1])	—	—	—	54.22	73.08	58.18
MCB (1) (as reported in [1])	—	—	—	42.27	78.42	38.29
SPMC-LPN (1)	—	—	—	65.73	63.07	61.00
DRNA (1)	—	—	—	67.77	61.80	50.47
DRNA (2)	—	—	—	68.77	61.80	50.39
DRNA-UNet-UNNC	—	—	—	68.18	64.30	55.20
Proposed model	—	—	—	68.18	64.30	55.20
Baseline features T+7, single network	42.07	76.20	36.48	52.62	42.27	79.32
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T+7, ensemble	66.38	81.38	43.17	57.41	66.73	81.71
Image features from bottom-up attention, adaptive K, ensemble	69.87	86.00	46.86	76.74	76.64	81.35

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



Zoph et al. Learning Transferable Architectures for Scalable Image Recognition



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

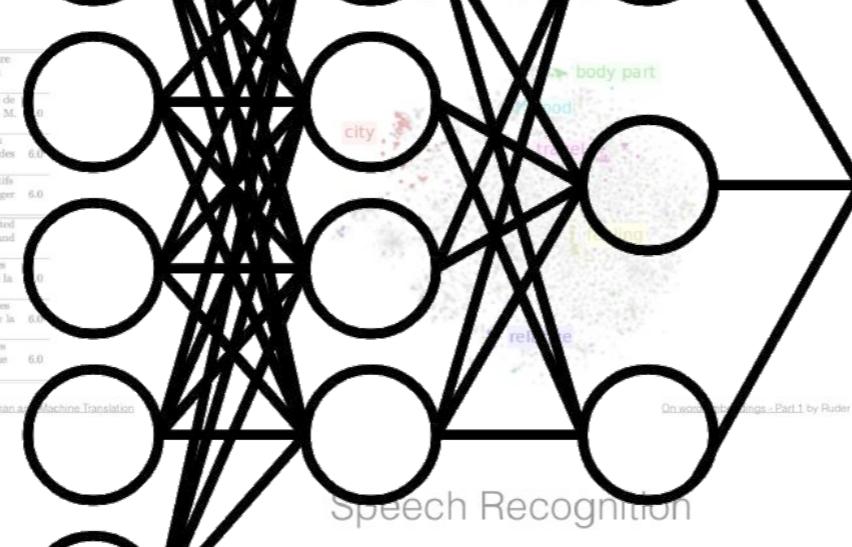
## Question Answering

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				
BERTBASE (Single)	80.8	88.5	-	-
BERTLARGE (Single)	84.1	90.9	-	-
BERTLARGE (Ensemble)	85.8	91.8	-	-
BERTLARGE (Sgl.+TriviaQnA)	84.2	91.1	85.1	91.8
BERTLARGE (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.

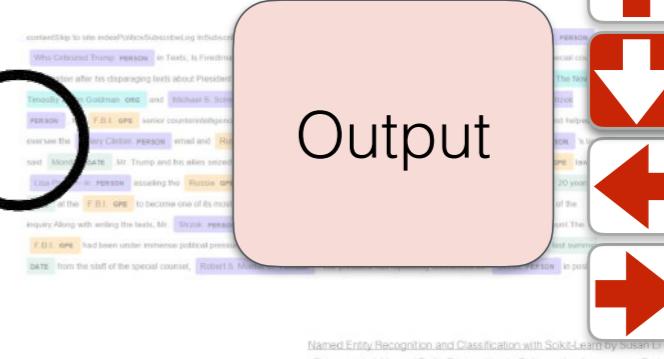
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## Word Embeddings



On word embeddings - Part 1 by Ruder

## Named Entity Recognition



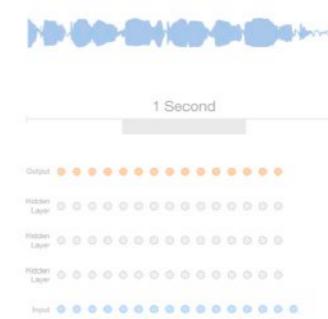
Named Entity Recognition and Classification with Scikit-learn by Susan Li Esteves et al. Named Entity Recognition in Twitter using Images and Text

## 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-puhpuh	BLSTM	11.3	8.0
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpuh	BLSTM+ResNet+LACE	9.7	7.4
9k-puhpuh	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. (%)
Jarrett et al. [27]	224x224	11.0M	1.0B	74.8	92.2
NASNet-A (5@ 15.38)	299x299	10.9M <sup>a</sup>	2.35 B	78.6	94.2
Inception V3 [60]	299x299	23.8M <sup>b</sup>	5.72 B	78.8	94.4
Xception [9]	299x299	22.8M <sup>b</sup>	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299x299	55.8M <sup>b</sup>	13.2 B	80.1	95.1
NASNet-A (7@ 1920)	299x299	22.6M <sup>b</sup>	4.93 B	80.8	95.3
PolyNet [69]	331x331	1.6M <sup>b</sup>	83.6 B	80.9	95.6
DPN-131 [8]	329x329	1.6M <sup>b</sup>	32.0 B	81.5	95.7
SENet [25]	320x320	1.6M <sup>b</sup>	42.3 B	82.7	95.8
NASNet-A (6@ 4032)	331x331	1.6M <sup>b</sup>	23.8 B	82.7	95.8

Table 1. Performance of architecture search and other related state-of-the-art models in ImageNet classification. Mult-Adds indicate the number of compute operations required to perform a multiplication and addition. If multiple operations are composed, the total number of operations is the sum of the individual operations. Model size for 145 is calculated from open-source implementation.

## Instance Segmentation



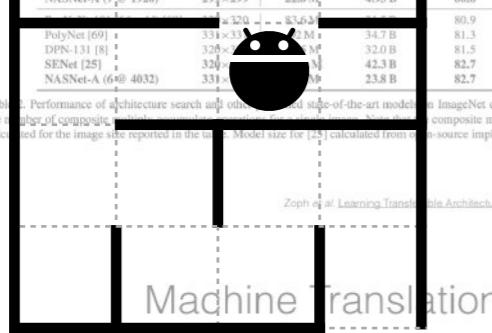
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Method	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual	Nodes	All	Visual	Nodes
Other cross-domain models in training set (1)	—	—	—	25.08	27.00	27.07
LSTM Language only model (1)	—	—	—	44.26	47.40	55.85
Deep LSTM Q-vqa (1) (as reported in [1])	—	—	—	54.22	73.08	58.18
MCB (1) (as reported in [1])	—	—	—	42.27	78.42	38.29
SPMC-LPN (1)	—	—	—	65.73	63.07	61.06
DRNA (1)	—	—	—	67.77	61.80	61.07
DRNA (2)	—	—	—	68.77	61.80	59.98
DRNA (3)	—	—	—	68.18	64.35	65.20
DRNA-UNet-UNNC	—	—	—	68.18	64.35	65.20
Proposed model	—	—	—	68.18	64.35	65.20
Baseline features T+7, single network	42.07	76.20	56.48	52.62	42.27	79.32
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T+7, ensemble	66.38	81.38	43.17	57.41	66.73	81.71
Image features from bottom-up attention, adaptive K, ensemble	69.87	86.00	46.89	76.74	86.64	81.55

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



Input

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".

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				
BERTBASE (Single)	80.8	88.5	-	-
BERTLARGE (Single)	84.1	90.9	-	-
BERTLARGE (Ensemble)	85.8	91.8	-	-
BERTLARGE (Sgl.+TriviaQnA)	84.2	91.1	85.1	91.8
BERTLARGE (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.

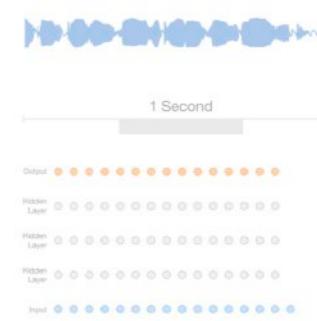
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

## 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-puhpuh	BLSTM	11.3	8.0	9.2	6.3
9k-puhpuh	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpuh	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
27k	BLSTM+ResNet+LACE	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. (%)
Jasguliu V2 [1]	299×299	10.9 M <sup>b</sup>	2.35 B	74.8	92.2
NASNet-A (5 <sup>a</sup> @ 15.38)	299×299	10.9 M <sup>b</sup>	2.35 B	78.6	94.2
Inception V3 [50]	299×299	23.8 M <sup>b</sup>	5.72 B	78.8	94.4
Xception [9]	299×299	22.8 M <sup>b</sup>	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M <sup>b</sup>	13.2 B	80.1	95.1
NASNet-A (7 <sup>a</sup> @ 1920)	299×299	22.6 M <sup>b</sup>	4.93 B	80.8	95.3
PolyNet [69]	331×331	1.6 M <sup>b</sup>	3.47 B	80.9	95.6
DPN-131 [8]	329×329	—	32.0 B	81.5	95.7
SENet [25]	320×320	—	42.3 B	82.7	95.8
NASNet-A (6 <sup>a</sup> @ 4032)	331×331	—	23.8 B	82.7	95.8

Table 1: Performance of architecture search and other related state-of-the-art models in ImageNet classification. Mult-Adds indicate the number of compute operations required to process one input image. Models which composite multiple accuracy operations are calculated for the image size reported in the paper. Model size for 145<sup>a</sup> calculated from open-source implementation.

## Instance Segmentation



Figure 5: Mask 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).

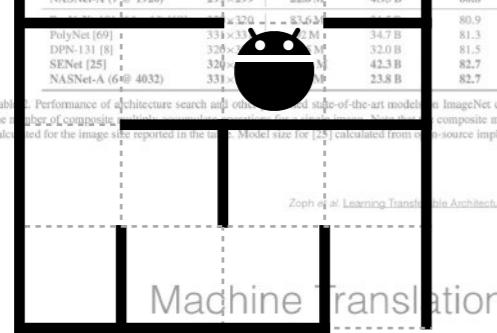
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

## Visual Question Answering

Method	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual Nodes	Others	All	Visual Nodes	Others
Other cross-domain models in training set [1]	—	—	—	25.08	36.08	37.77
LSTM Language only (old model) [1]	—	—	—	44.26	47.40	55.85
Deep LSTM Q-vqa [1] (revised in [1])	—	—	—	54.22	73.08	58.18
MCB [1] (as reported in [1])	—	—	—	42.27	78.42	38.29
SPM+CPN [1]	—	—	—	65.73	63.07	61.06
DRNA [1]	—	—	—	67.77	61.80	61.07
DRNA+ [1]	—	—	—	68.77	61.80	59.98
RCNN-UNet-UNNC	—	—	—	68.18	64.39	65.20
Proposed model	—	—	—	70.01	70.01	70.01
Baseline features T=7, single network	42.07	76.20	36.48	52.62	42.27	79.32
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20
Baseline features T=7, ensemble	64.38	63.38	43.17	57.47	64.73	63.71
Image features from bottom-up attention, adaptive K, ensemble	69.87	86.00	46.89	76.74	76.64	81.35

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



## Input

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".

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	80.8	88.5	-	-
BIDAF+ELMo (Ensemble)	81.1	92.0	-	-
BERT <sub>BASE</sub> (Single)	80.8	88.5	-	-
BERT <sub>BASE</sub> (Ensemble)	81.1	92.0	-	-
BERT <sub>LARGE</sub> (Single)	85.8	91.8	-	-
BERT <sub>LARGE</sub> (Ensemble)	84.2	91.1	85.1	91.8
BERT <sub>LARGE</sub> (Sgl+TriviaQnA)	84.2	91.1	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

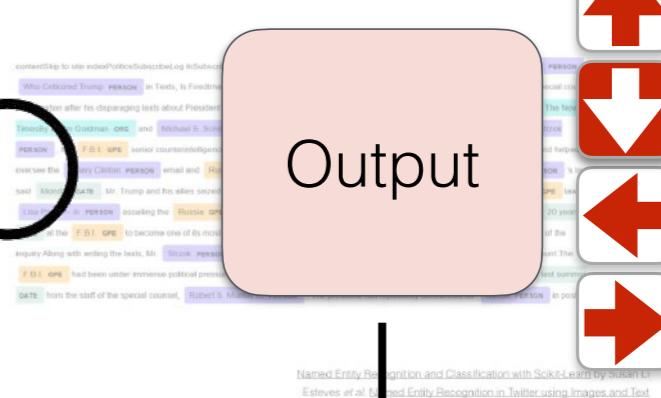


## Speech Recognition

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-puhpuh	BLSTM	11.3	8.0
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2
9k-puhpuh	BLSTM+ResNet+LACE	9.7	7.4
9k-puhpuh	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.8
27k	BLSTM+ResNet+LACE	10.0	7.5
9k-puhpuh	BLSTM+ResNet+LACE+CNN-BLSTM	7.1	5.7
9k	+ LSTM rescoring	7.3	5.2
9k	+ ngram rescoring	7.2	5.2
9k	+ backchannel penalty	7.2	5.1

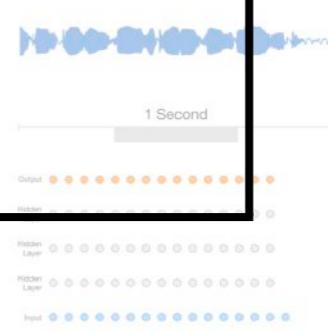
Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

## Named Entity Recognition



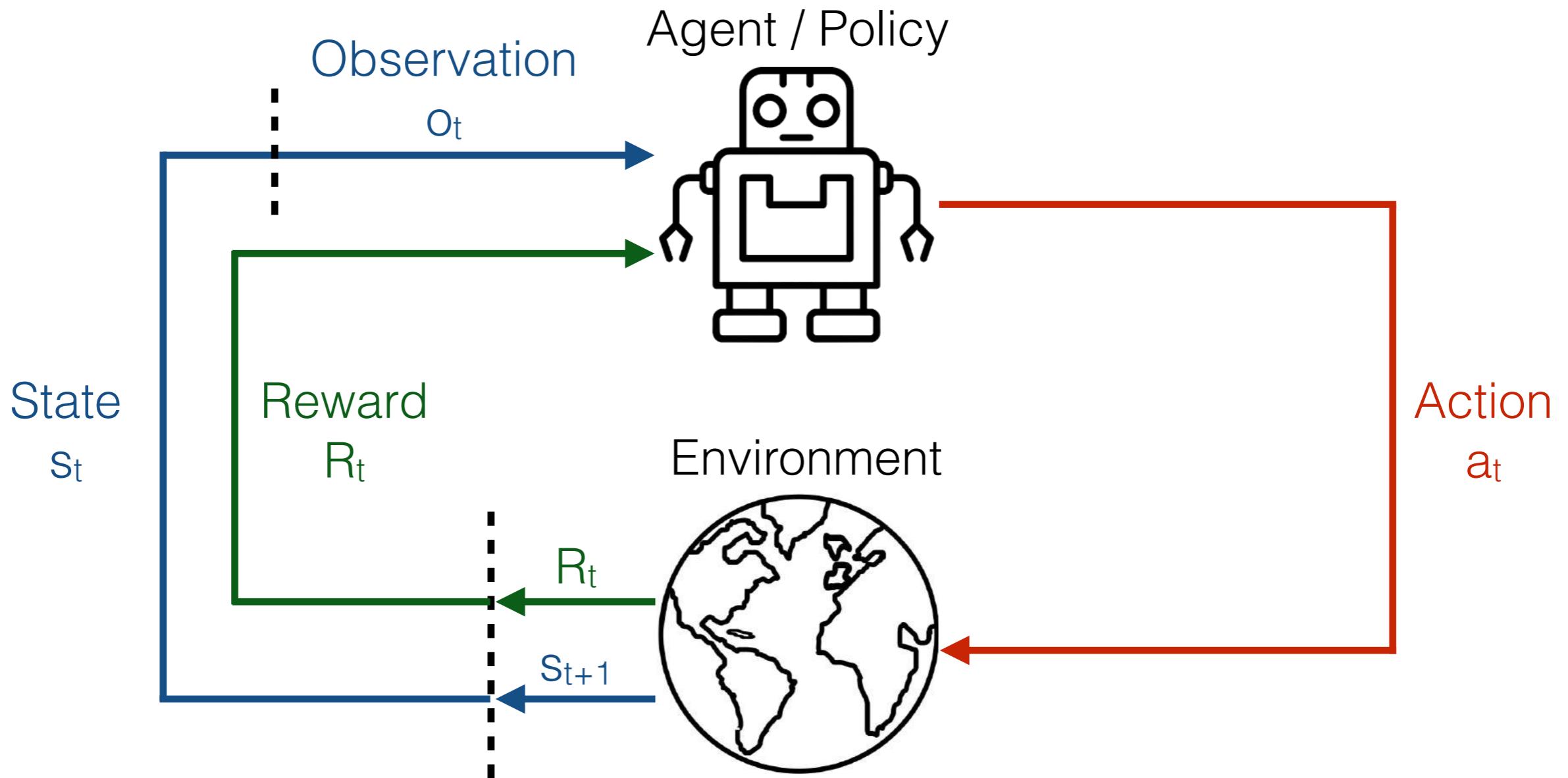
Named Entity Recognition and Classification with Scikit-learn by Susan Li Esteves et al. Named Entity Recognition in Twitter using Images and Text

## Speech Synthesis (text-to-speech)



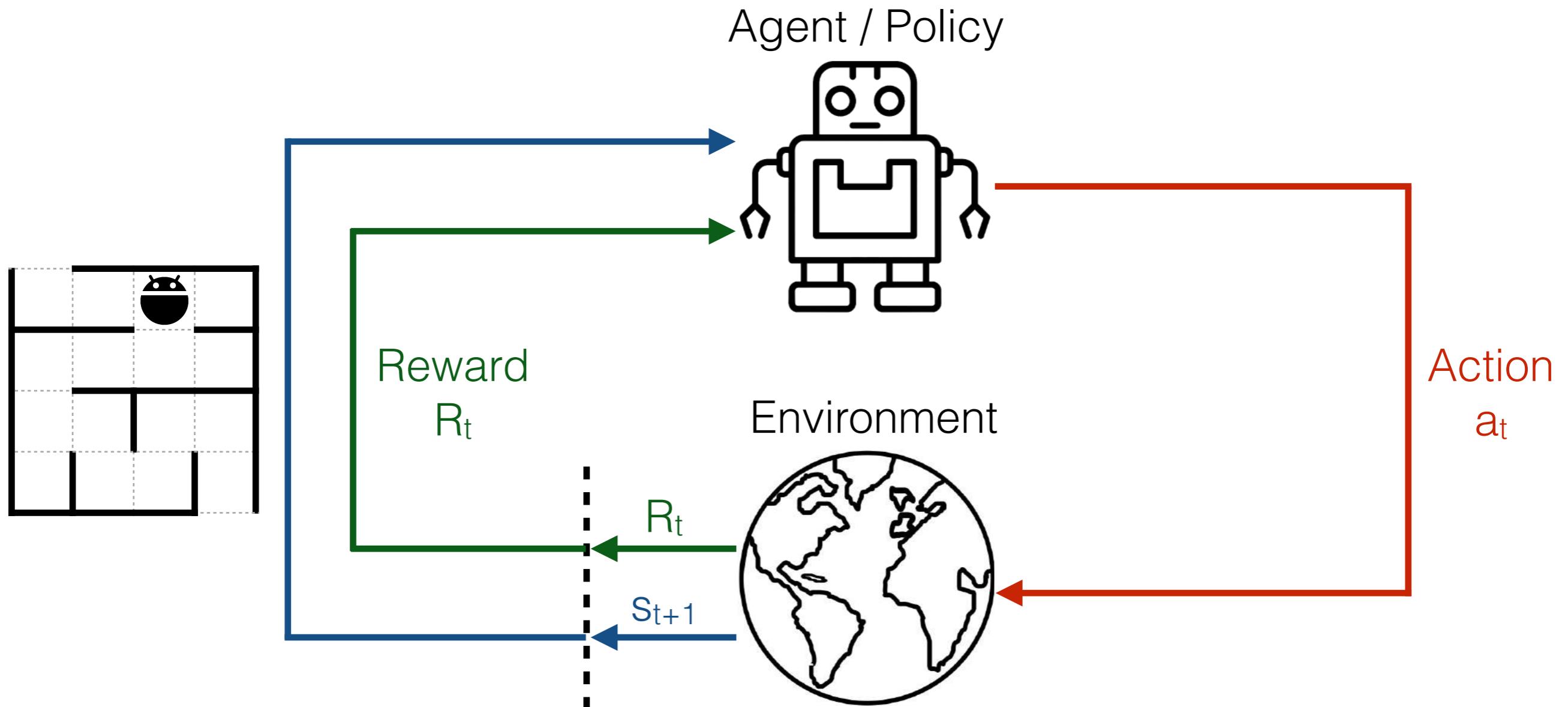
Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

# Robot Learning via Reinforcement Learning



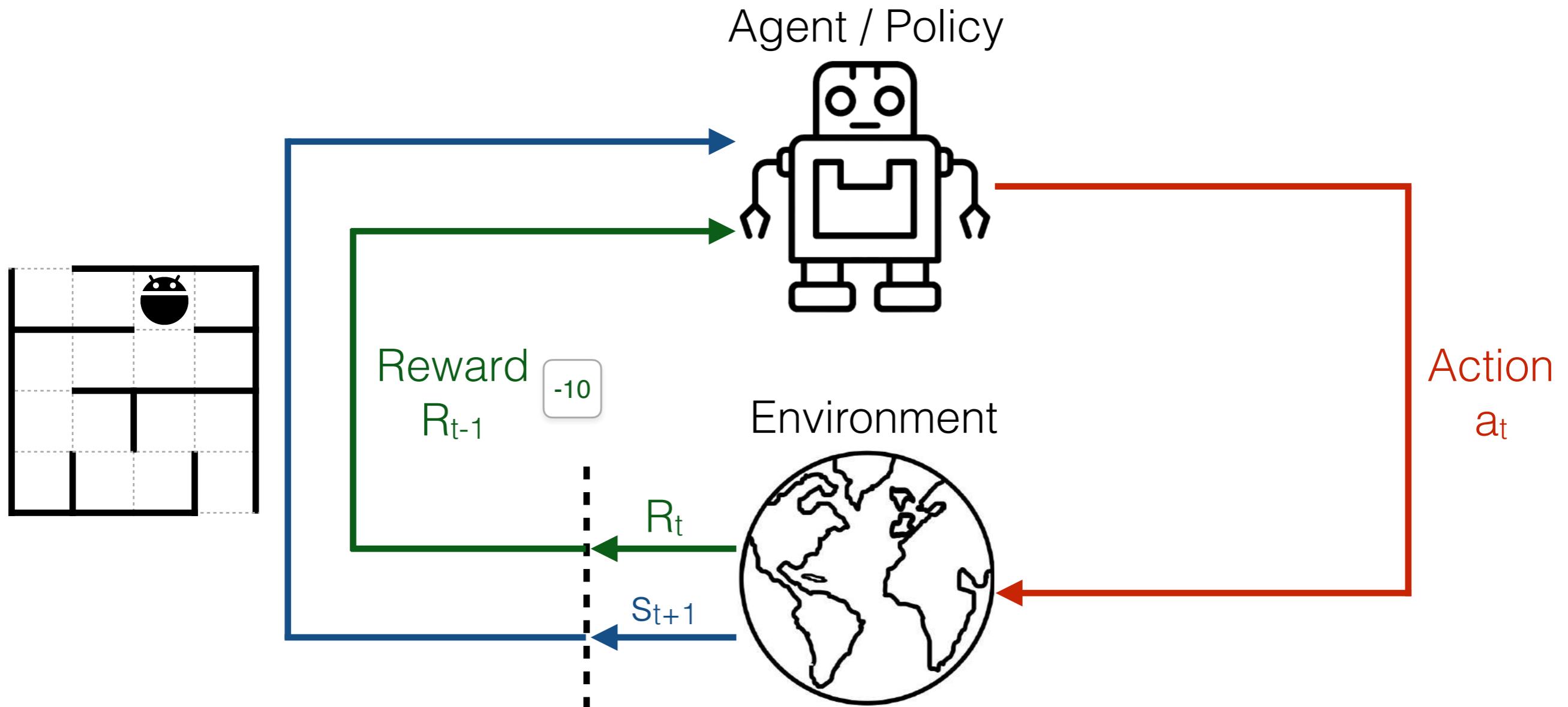
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



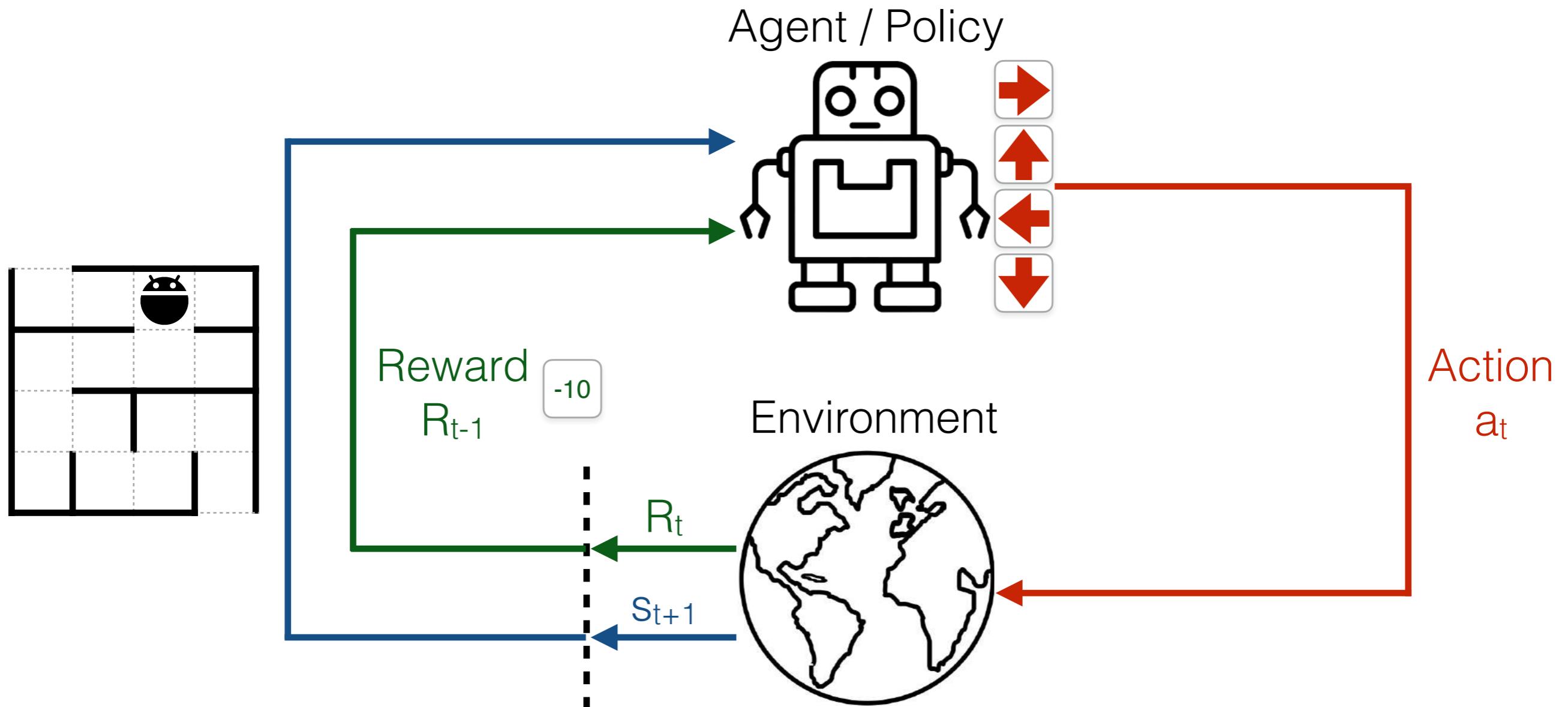
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



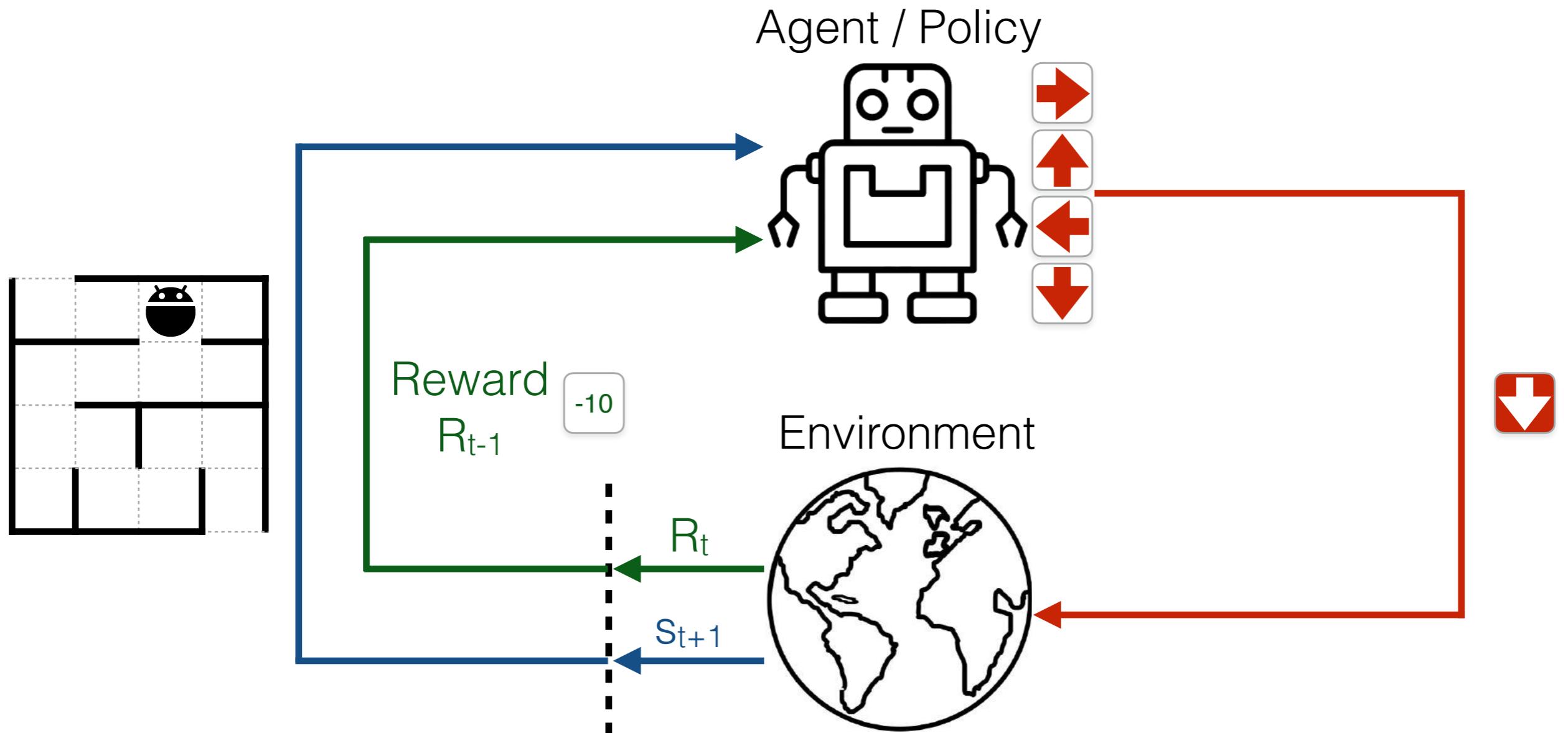
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



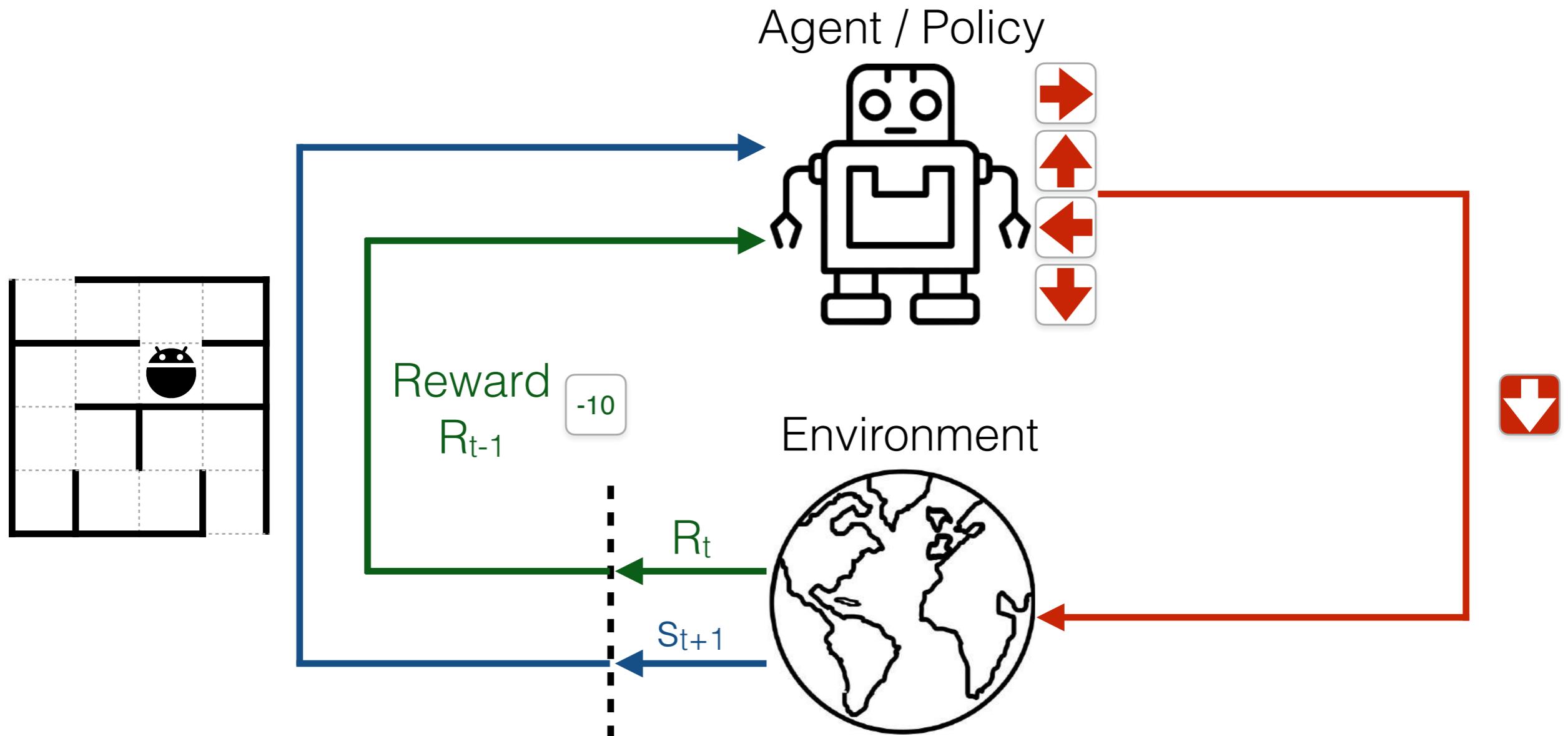
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



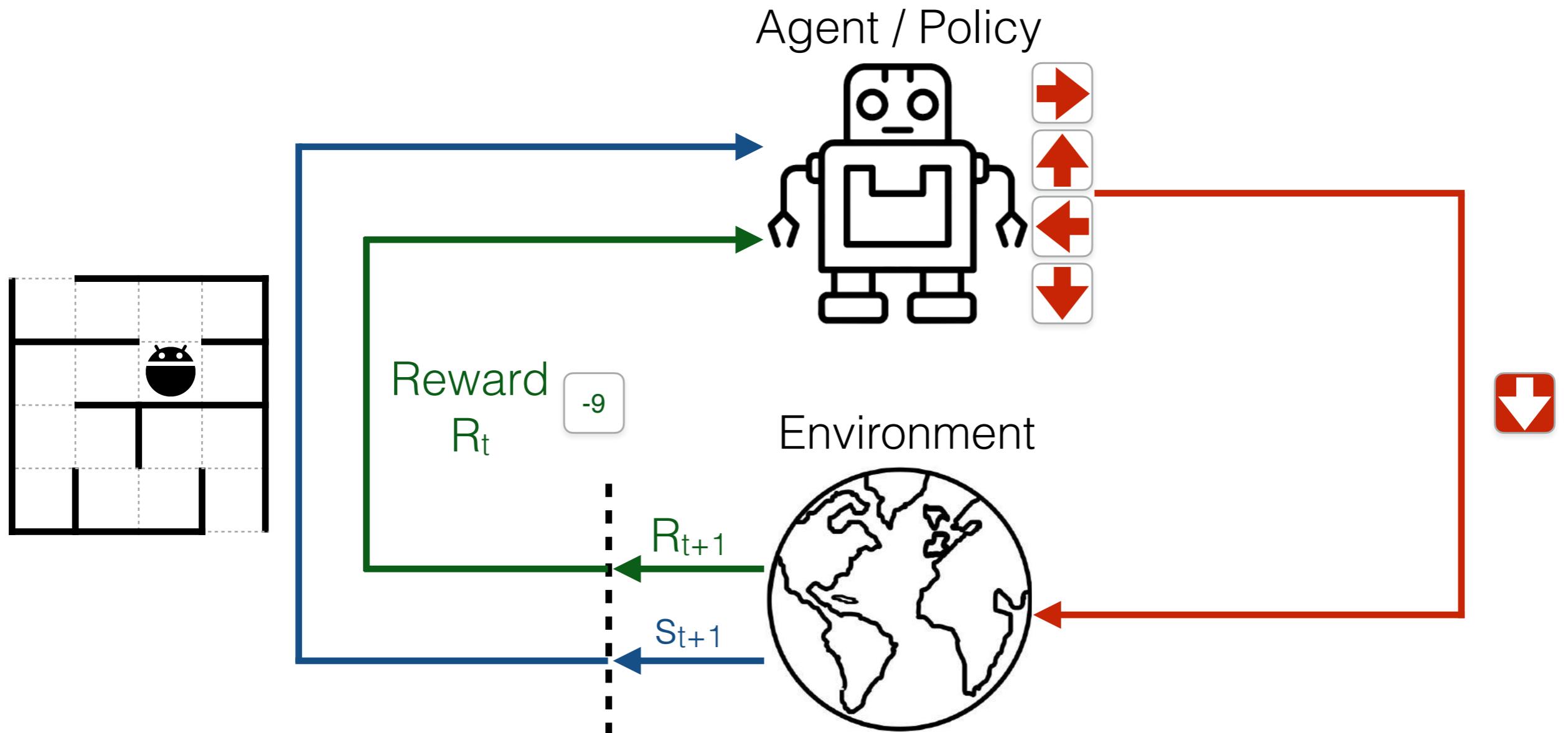
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



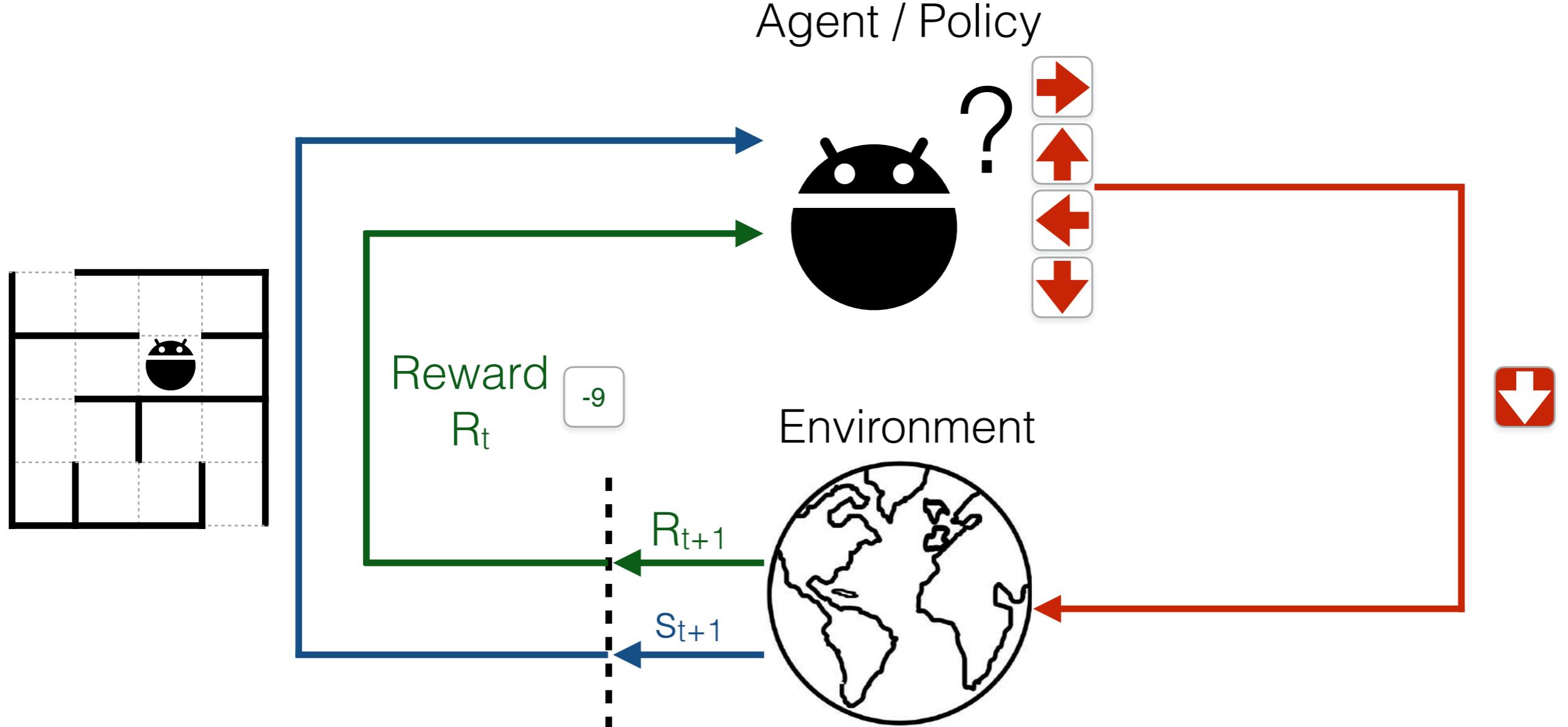
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



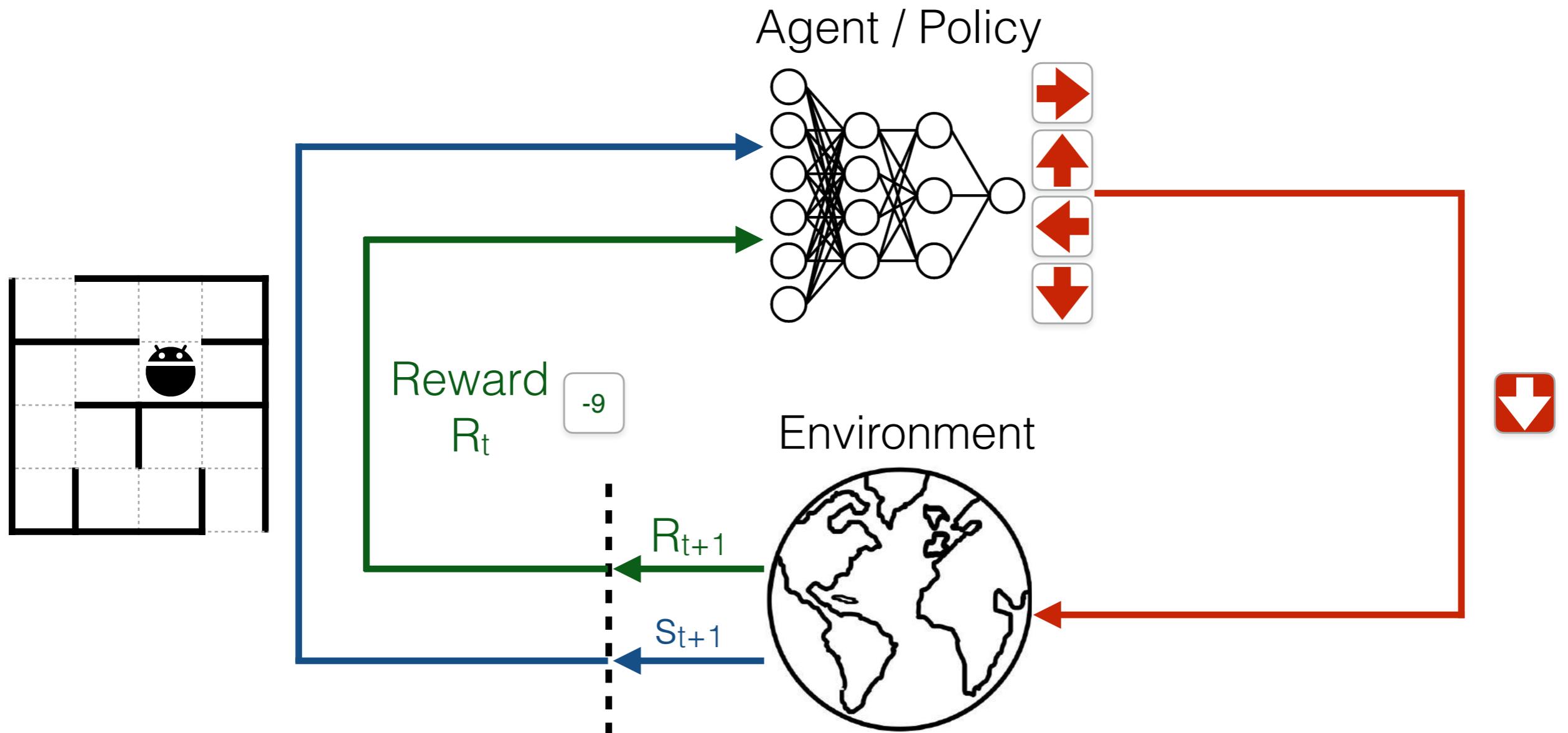
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# Robot Learning via Reinforcement Learning



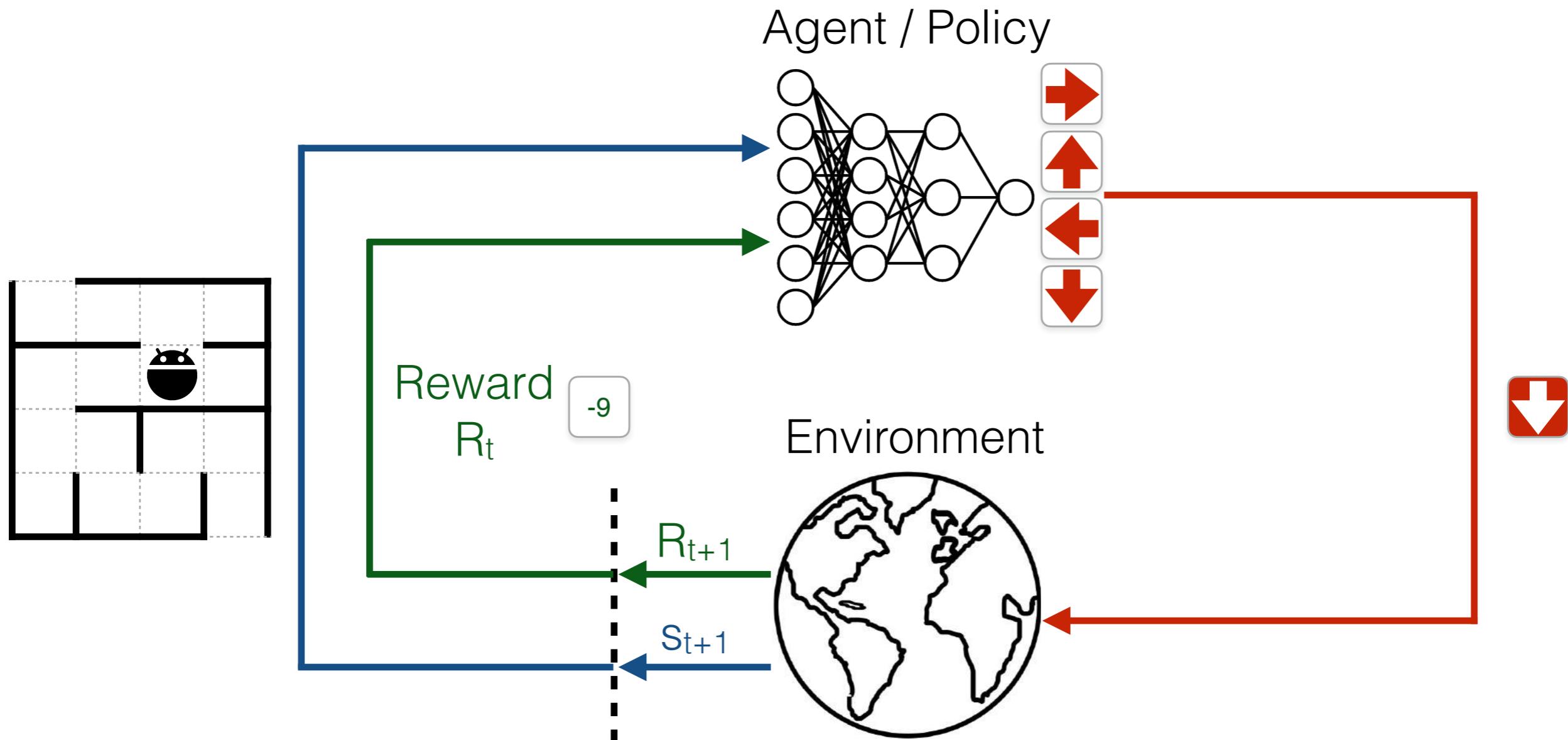
Goal: maximize  $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

# 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



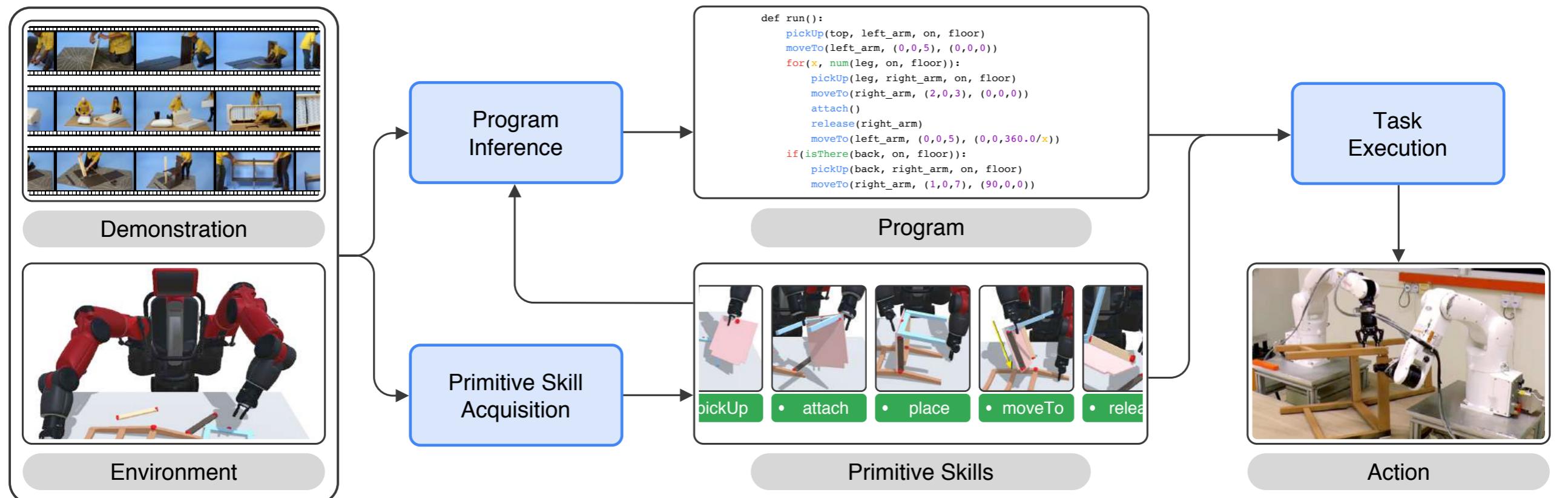
Uninterpretable

Not generalizable

Limited to short-horizon tasks

No skill-reuse

# Program-Guided Framework for Interpreting and Acquiring Complex Skills



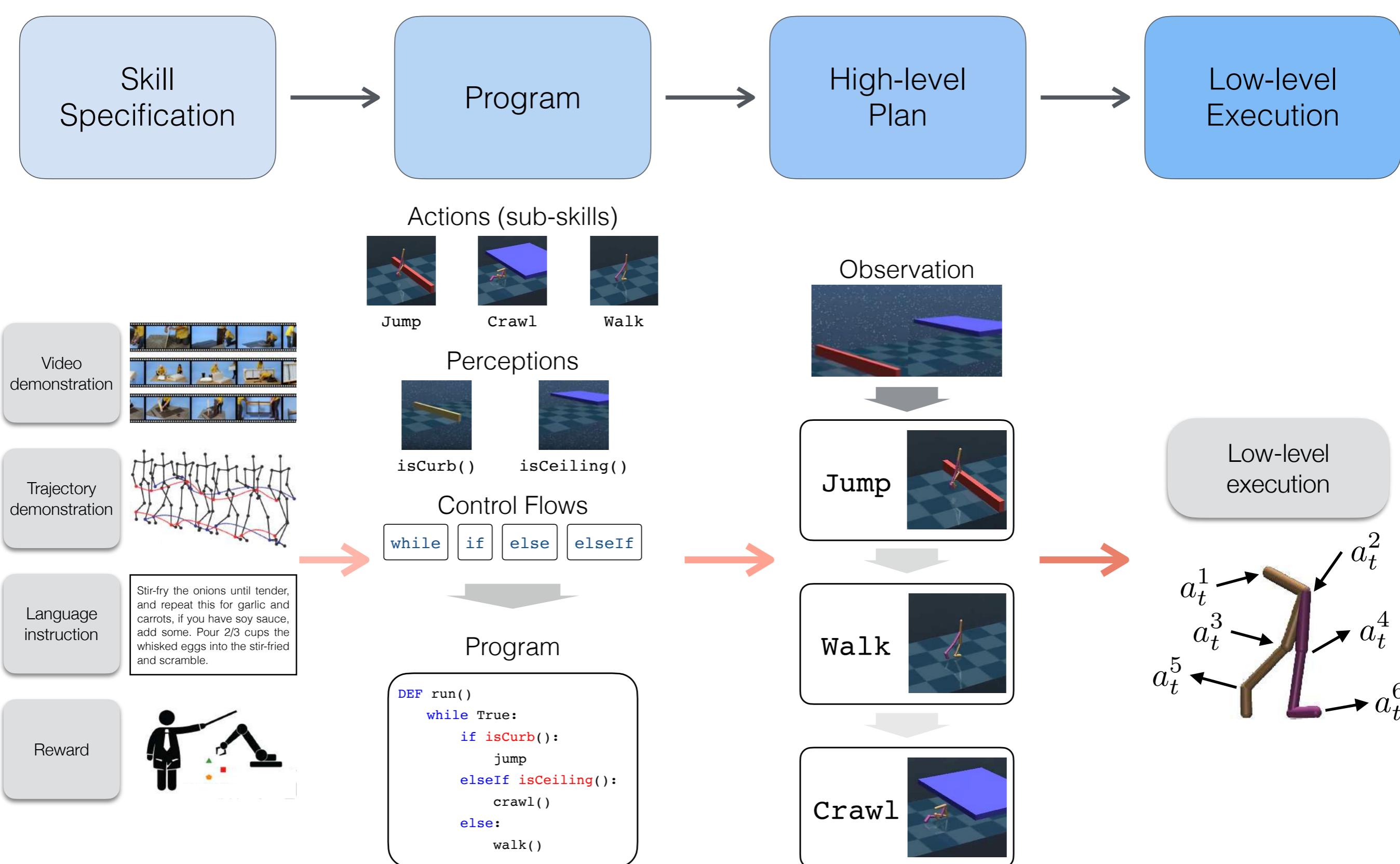
Interpretable

Programmatic / Generalizable

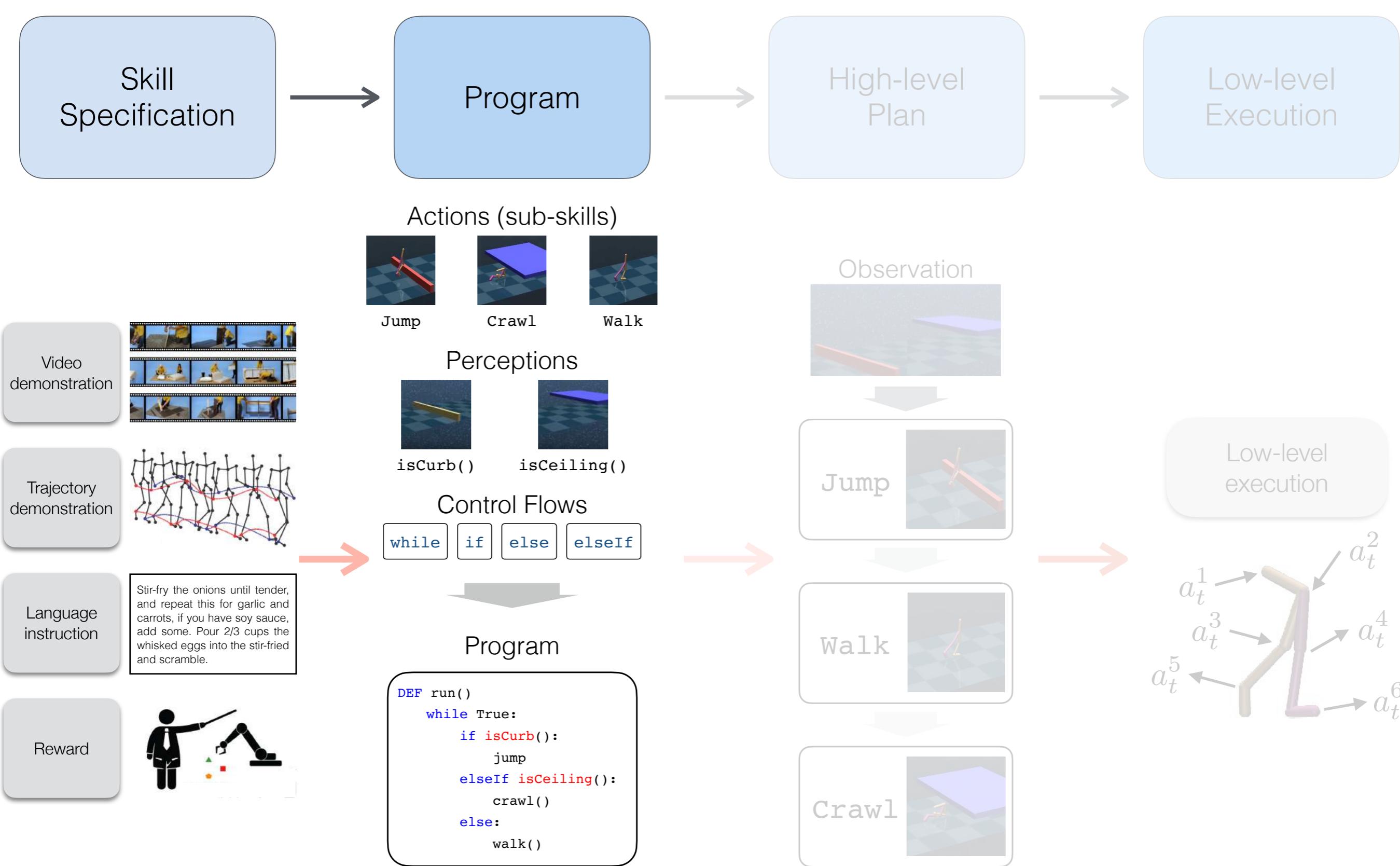
Hierarchical

Modular

# Program-Guided Framework for Interpreting and Acquiring Complex Skills

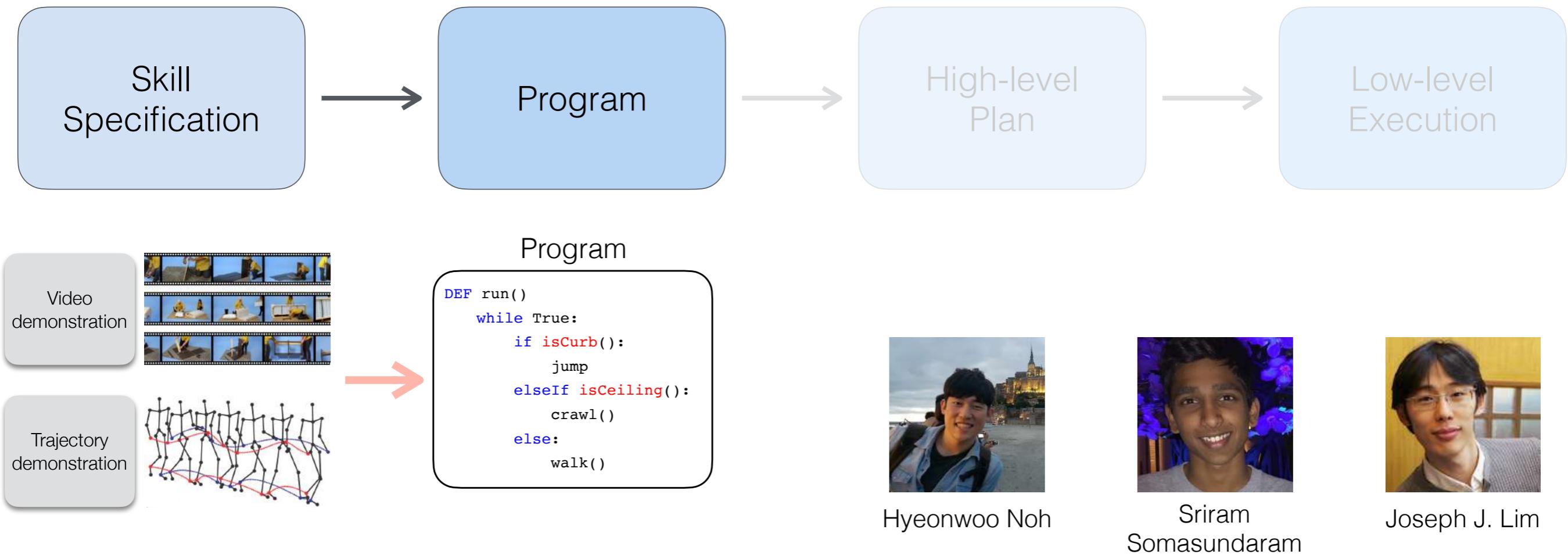


# Program Inference



# Neural Program Synthesis from Diverse Demonstration Videos

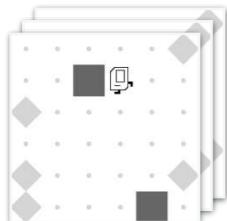
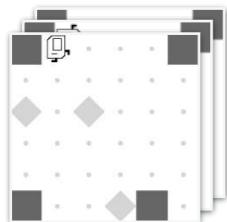
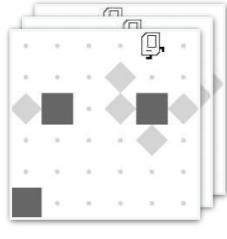
ICML 2018



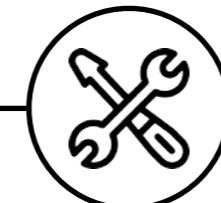
# Imitation Learning

---

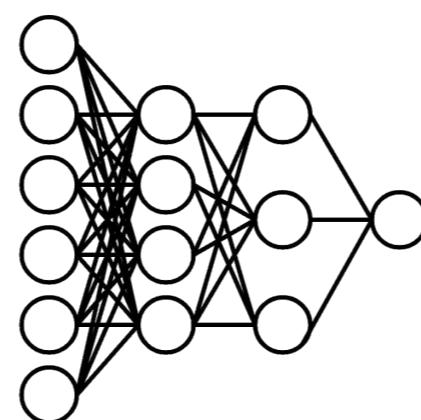
Demonstrations



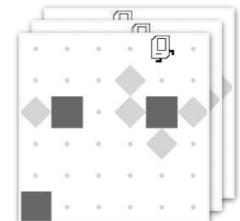
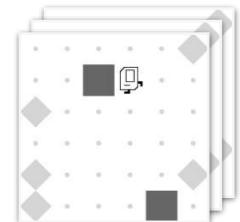
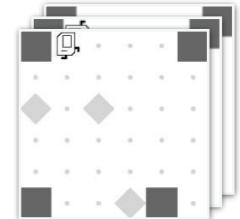
Imitate



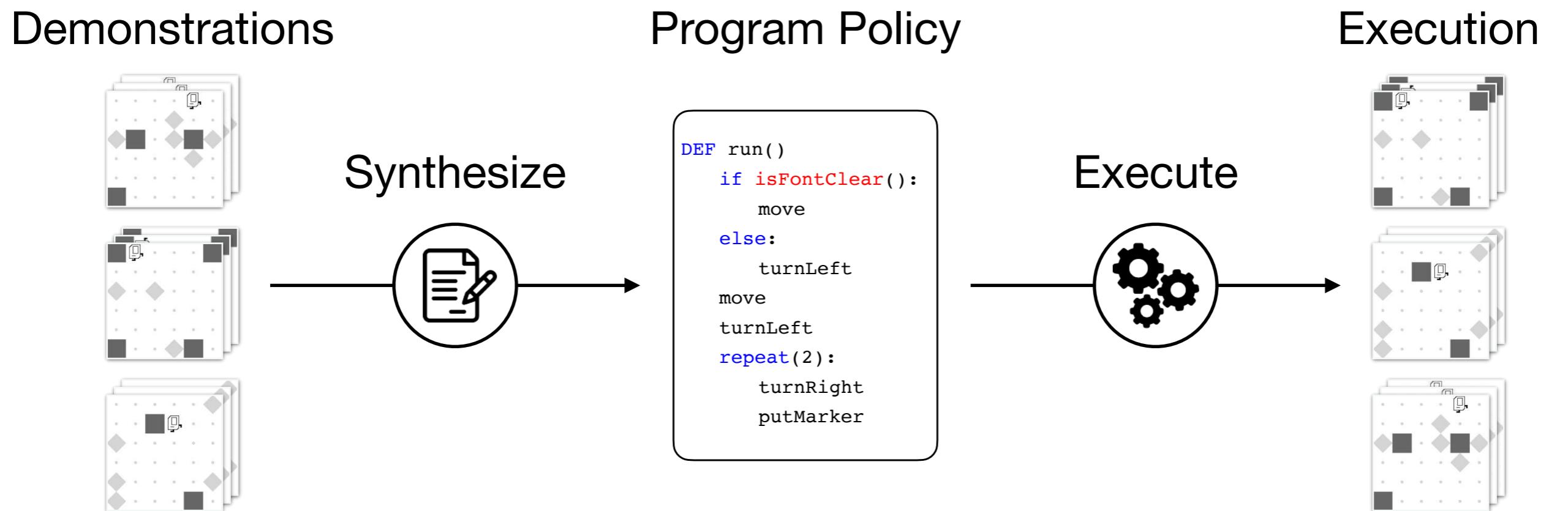
Neural Network  
Policy



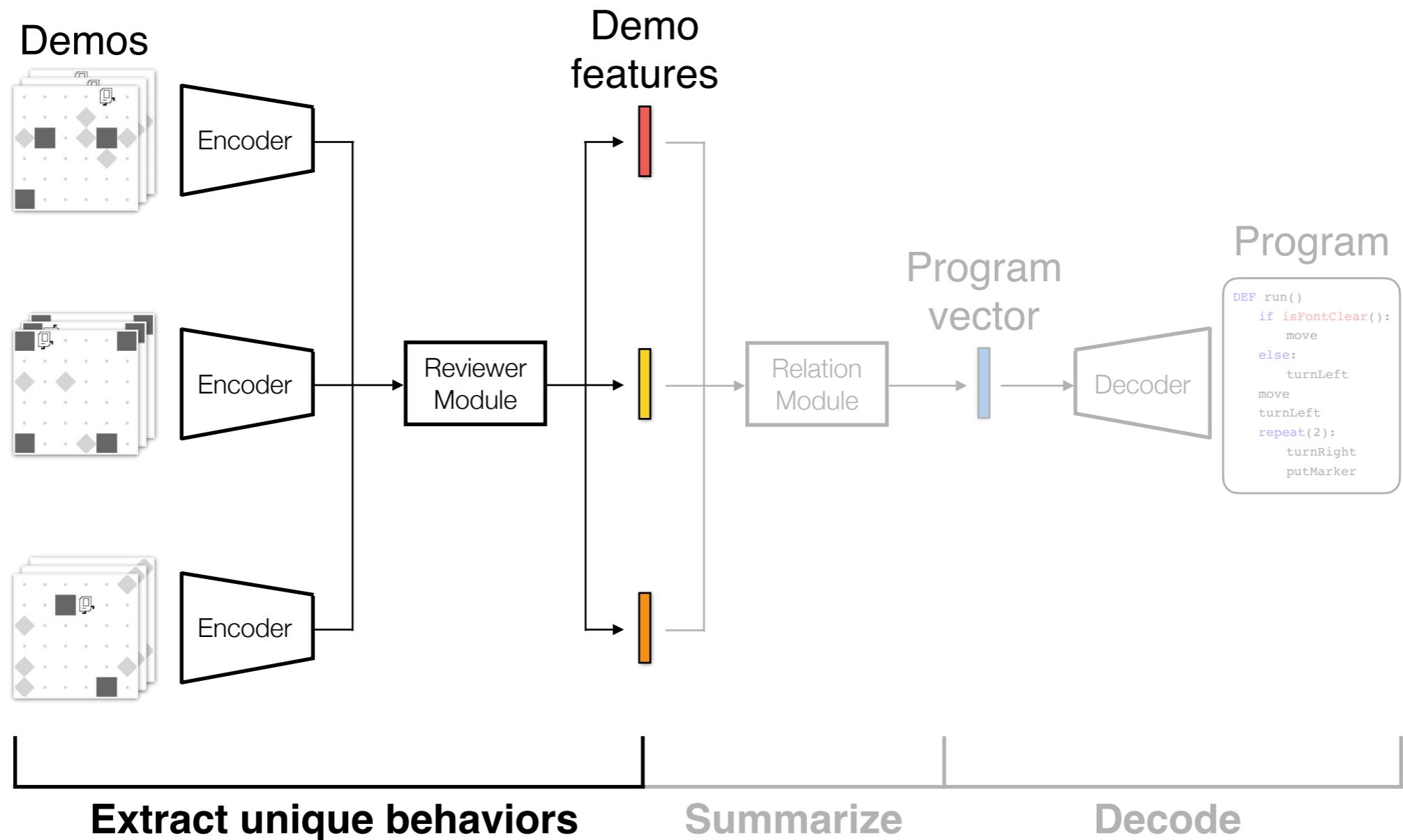
Execution



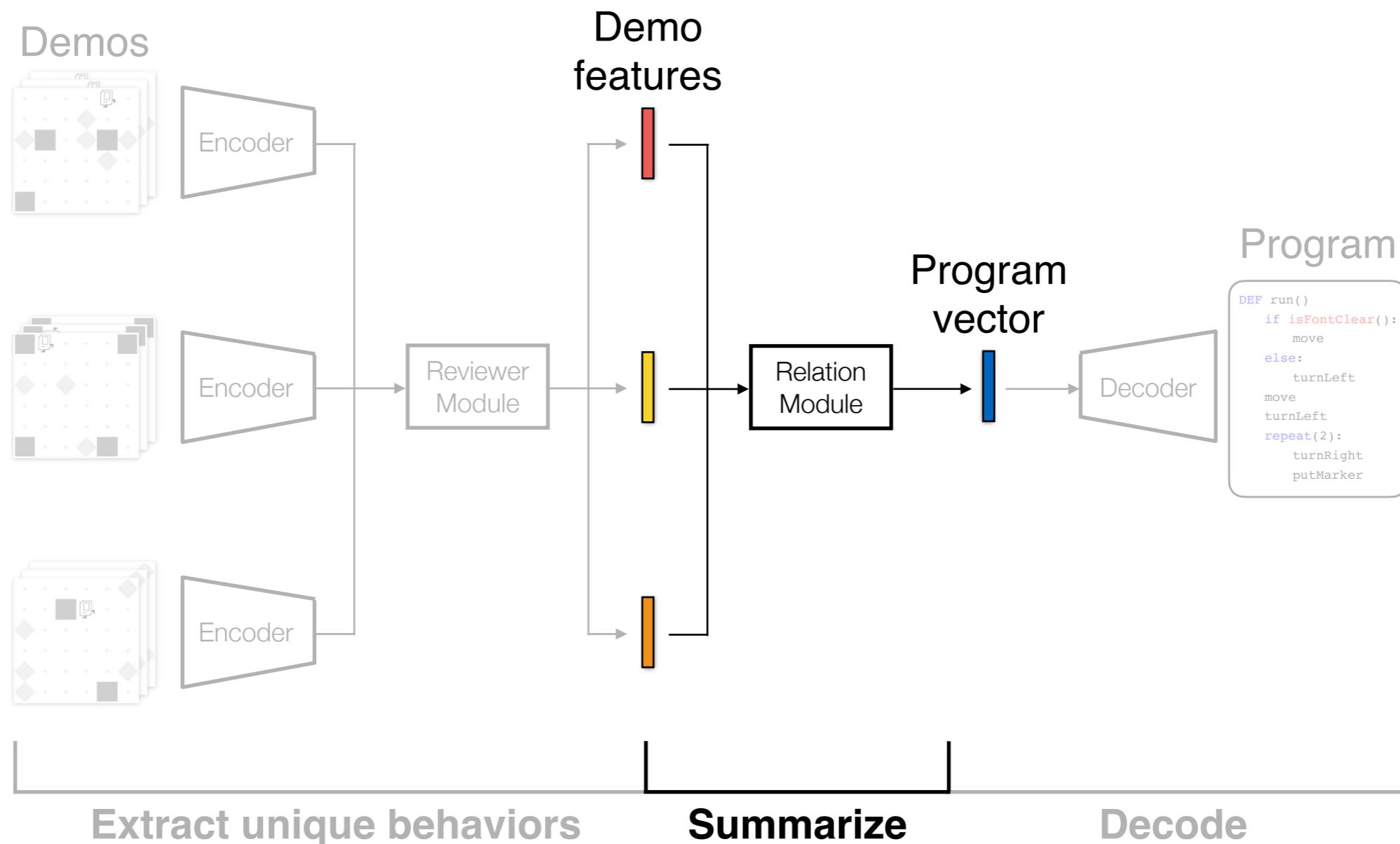
# Imitation Learning by Synthesizing Programs



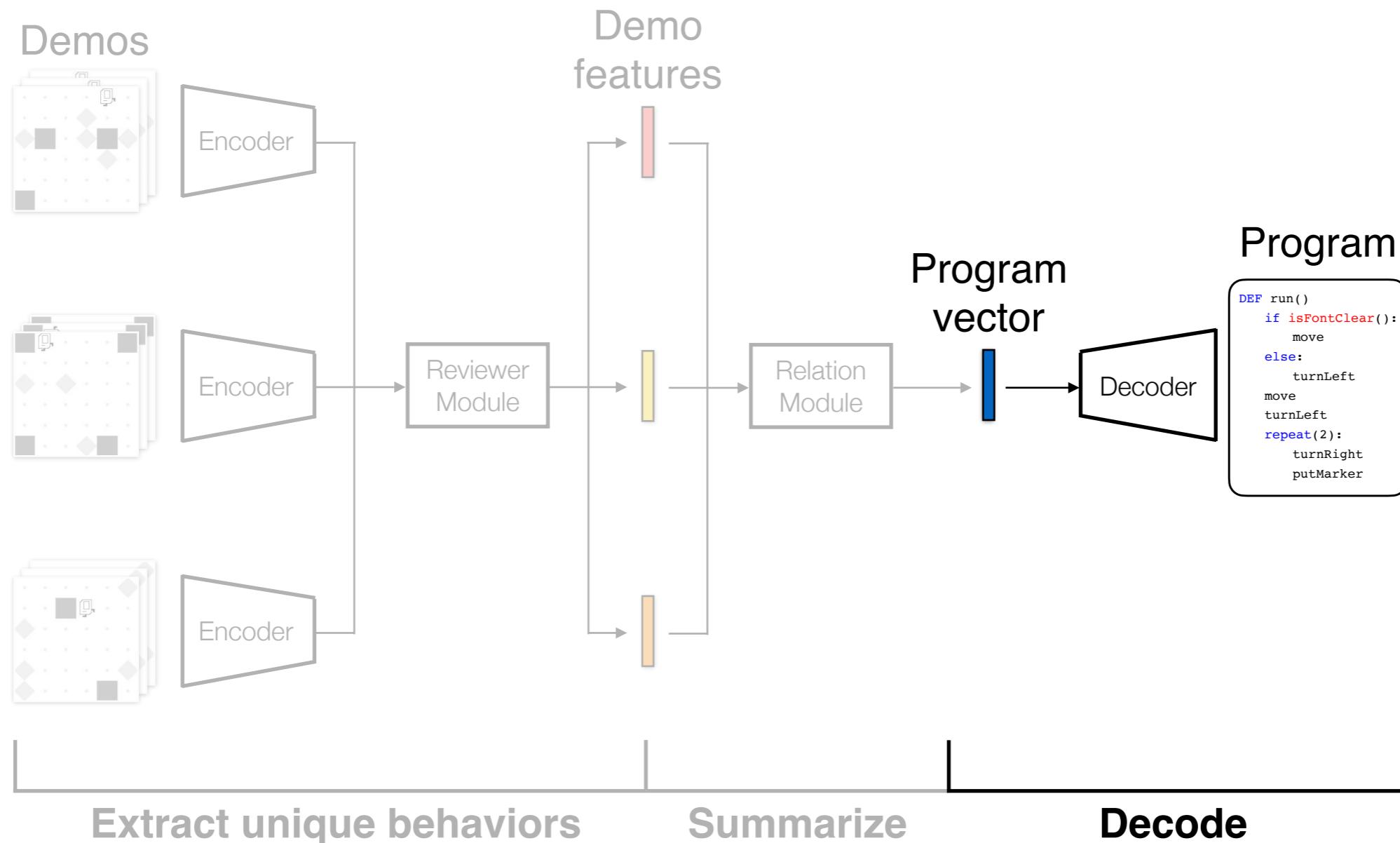
# Model Overview



# Model Overview



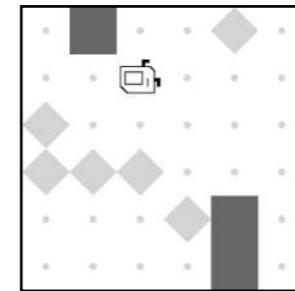
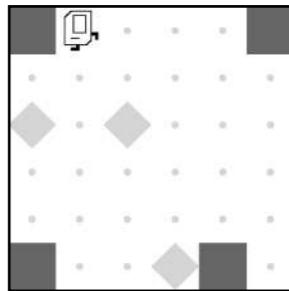
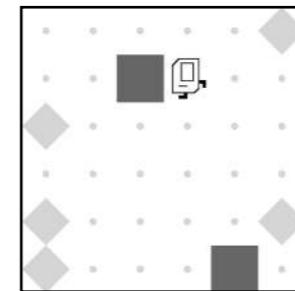
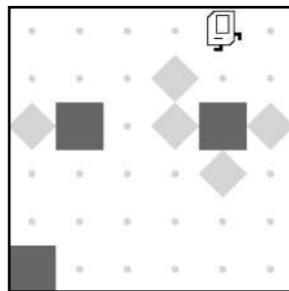
# Model Overview



# Environments

Karel

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



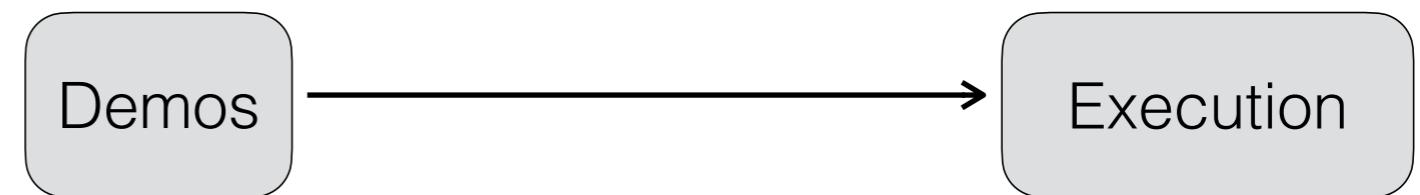
ViZDoom

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

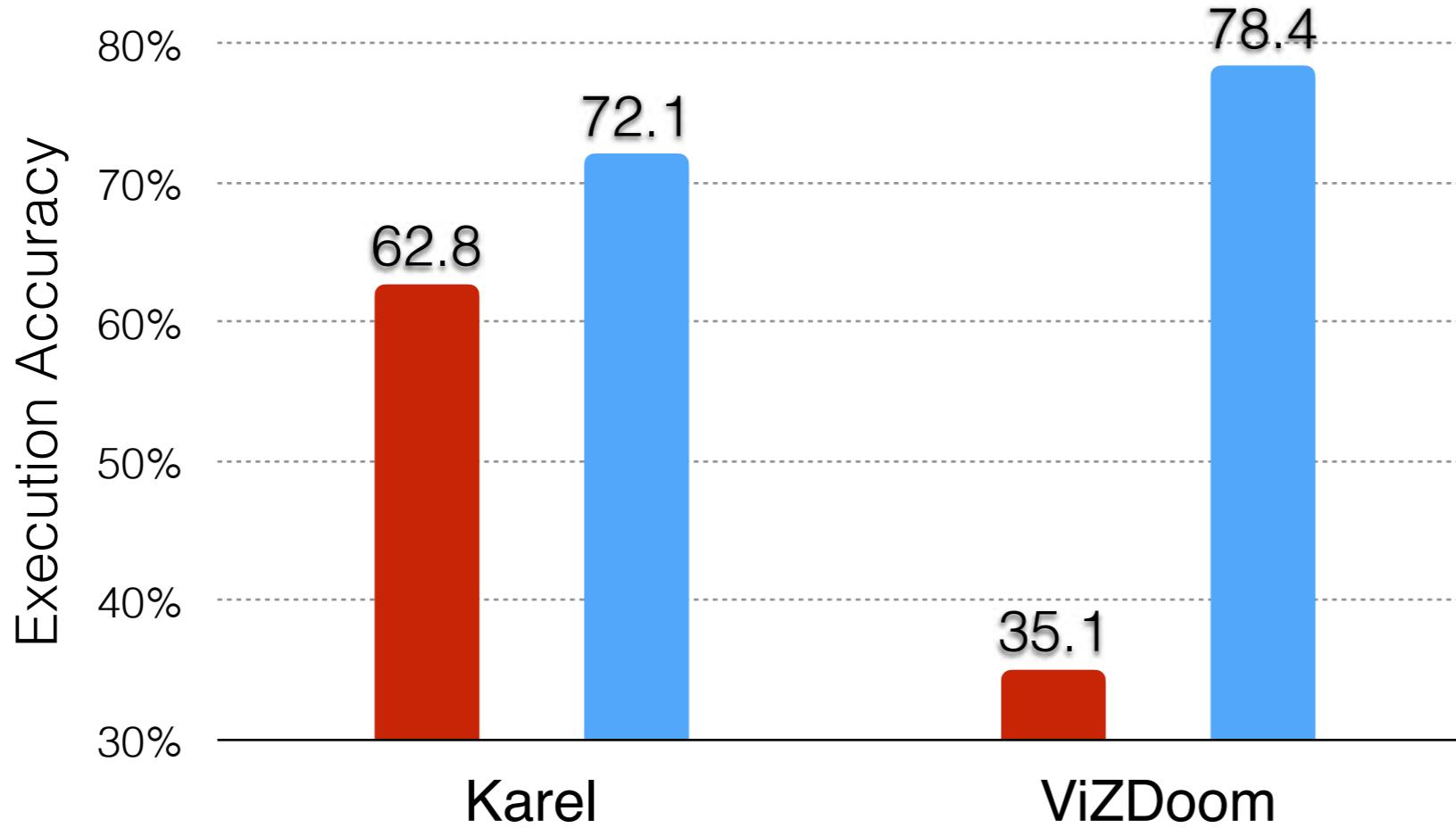
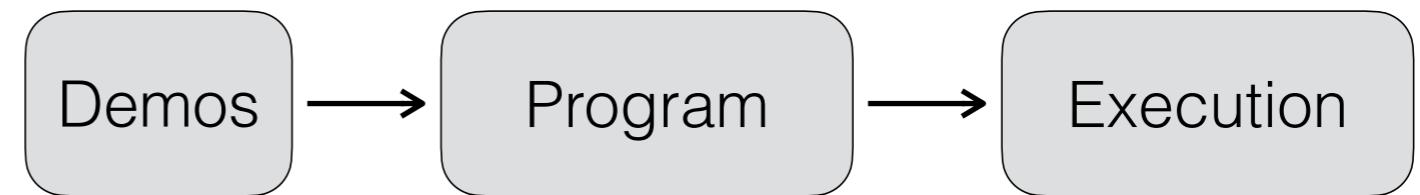


# Quantitative Results

Neural Network Policy



Program Policy

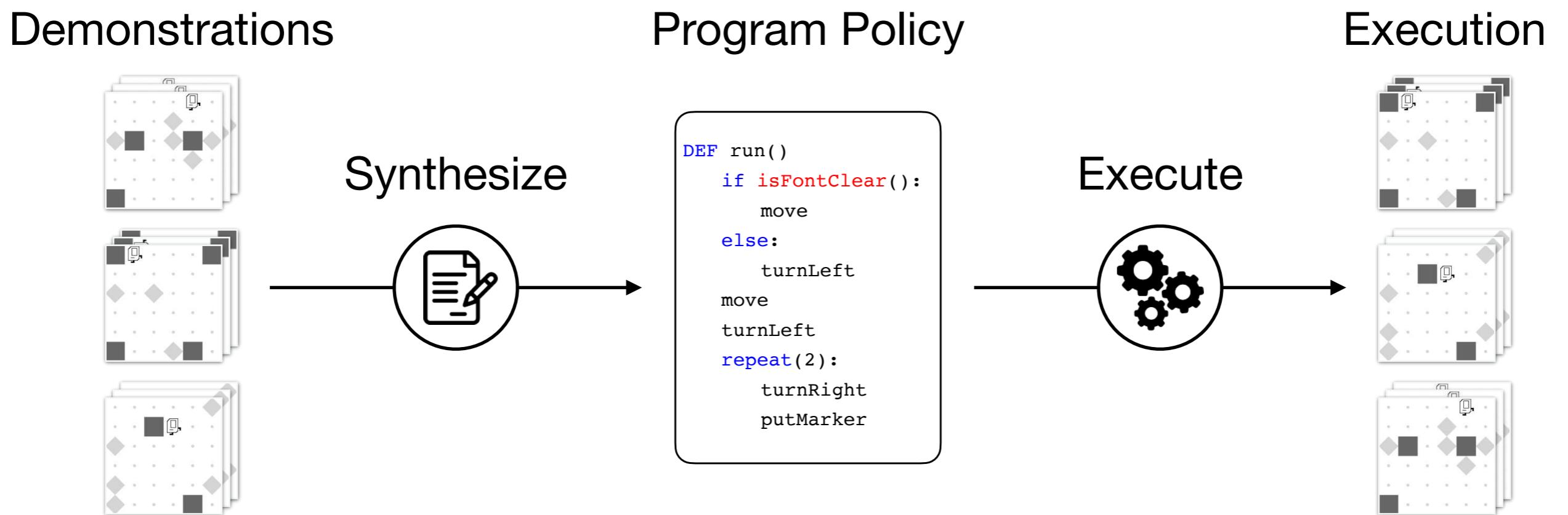


```
DEF run()
    if isFontClear():
        move
    else:
        turnLeft
        move
        turnLeft
        repeat(2):
            turnRight
            putMarker
```

# Takeaway

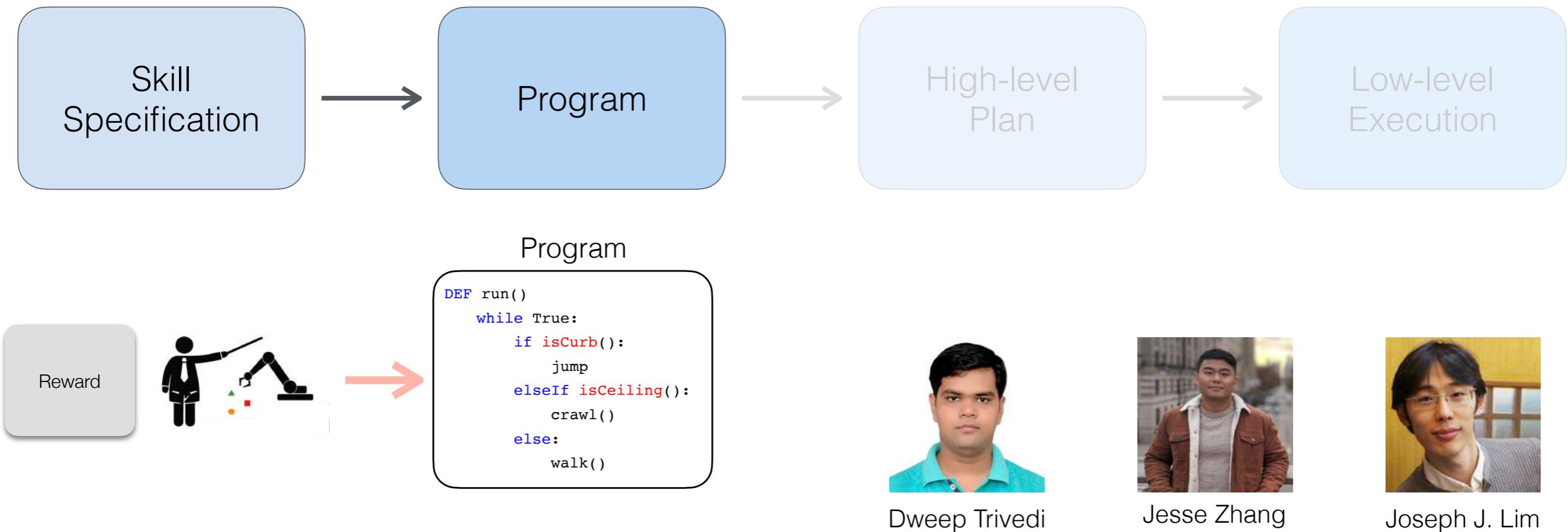
---

- Synthesize programs to imitate demonstrations

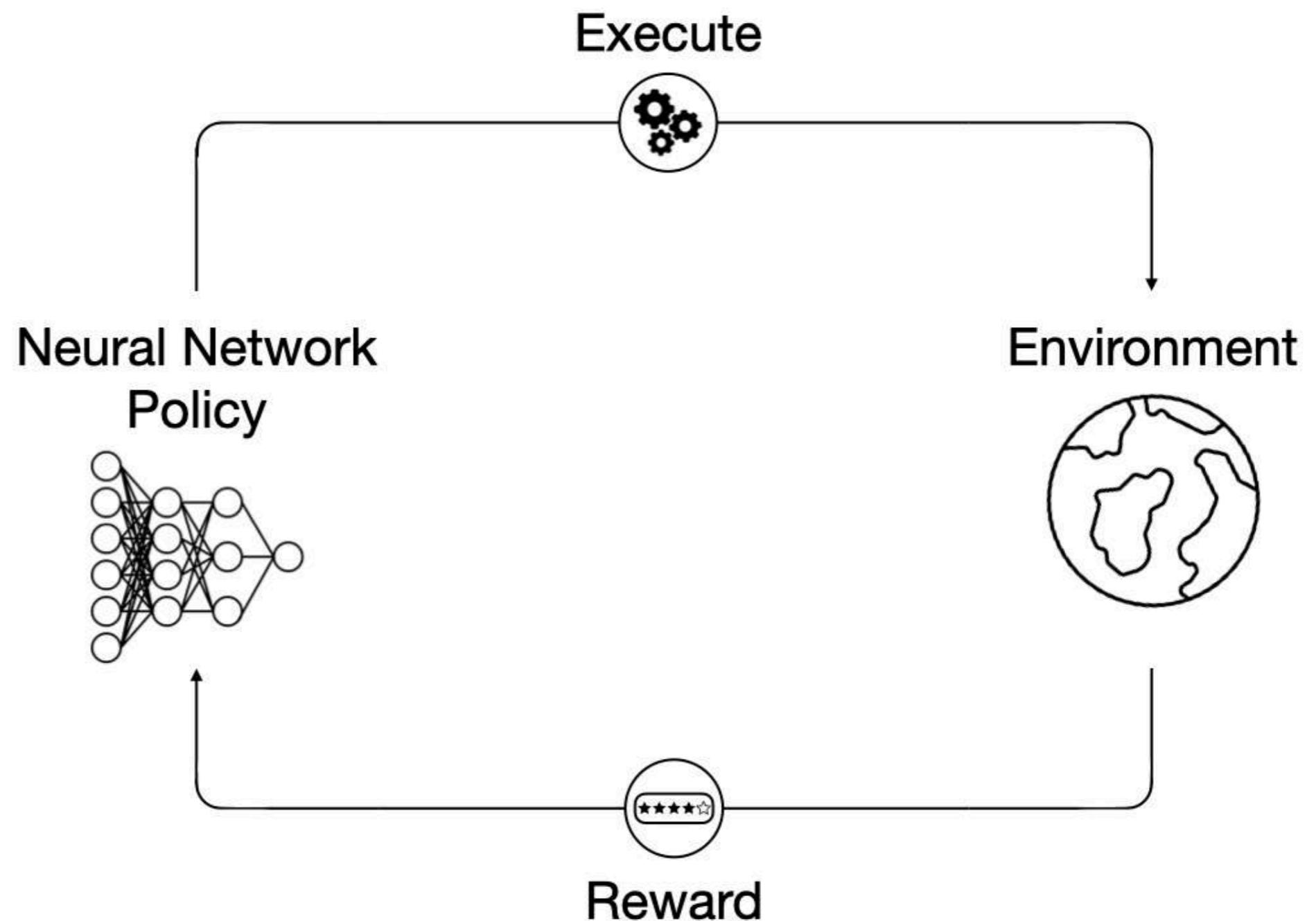


# Learning to Synthesize Programs as Interpretable and Generalizable Policies

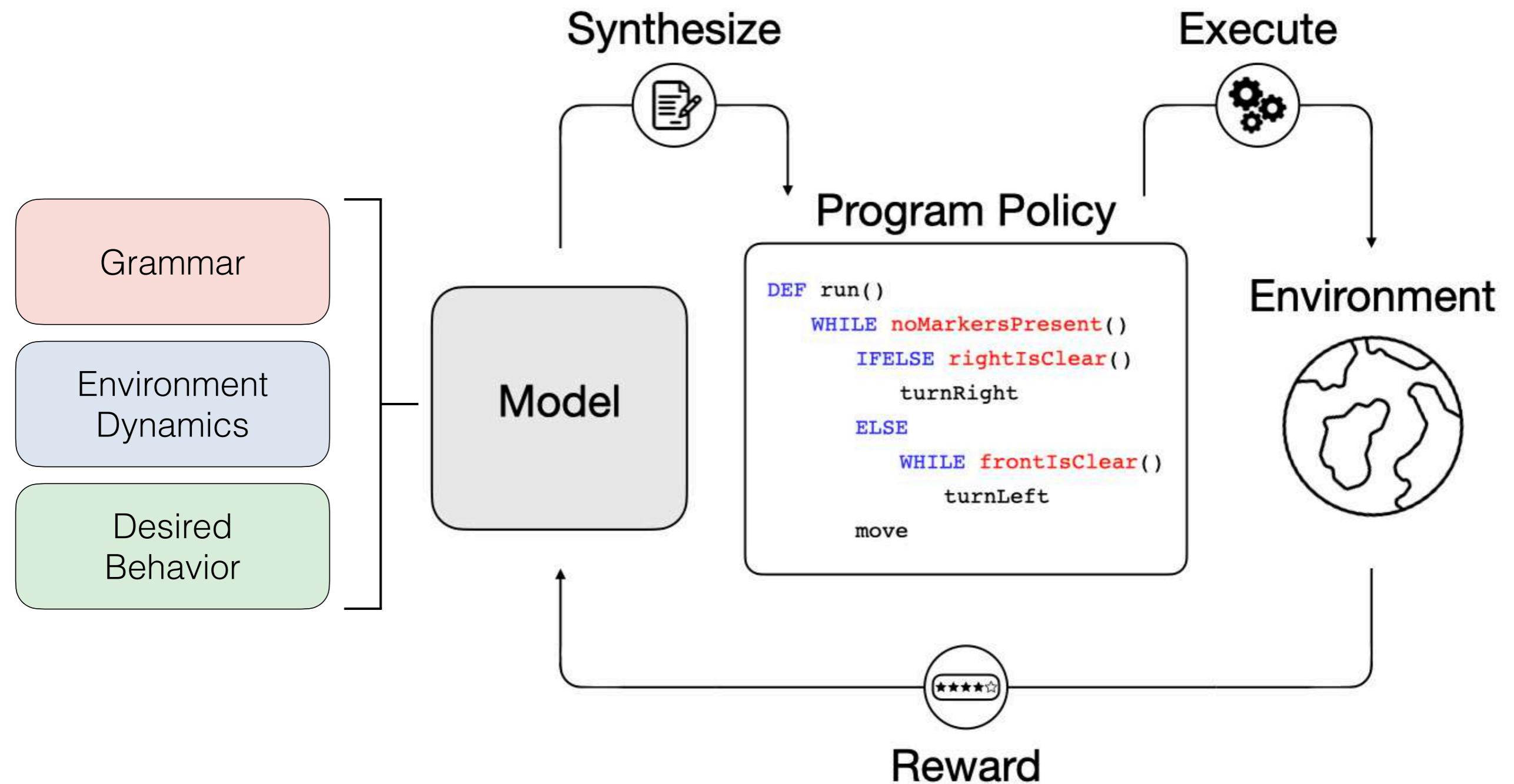
NeurIPS 2021



# Reinforcement Learning



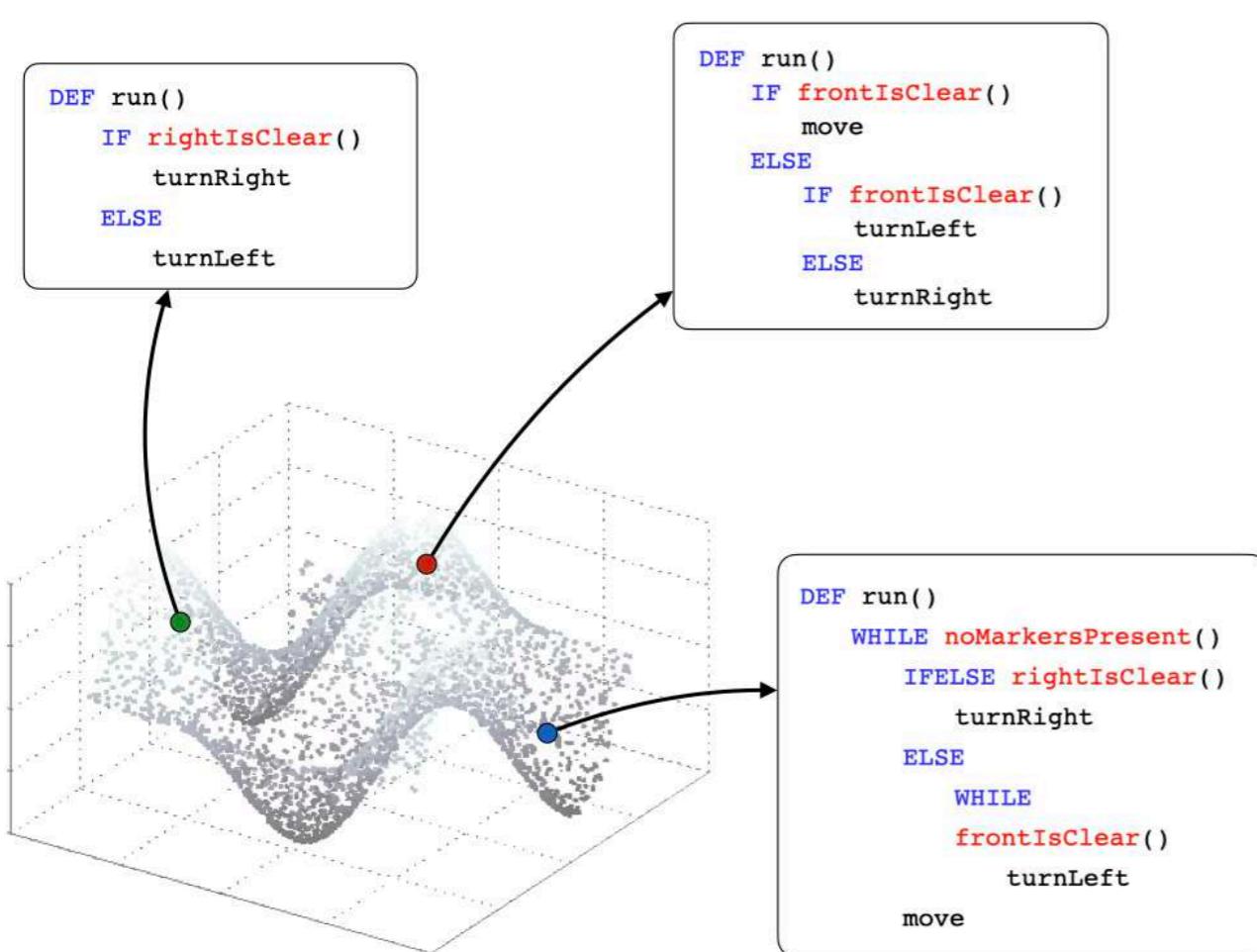
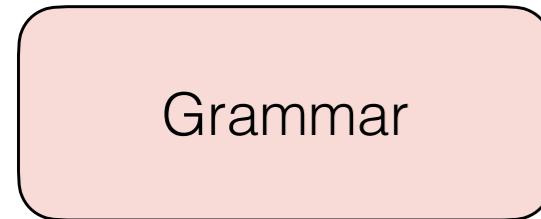
# Reinforcement Learning by Synthesizing Programs



# LEAPS: Learning Embeddings for Latent Program Synthesis

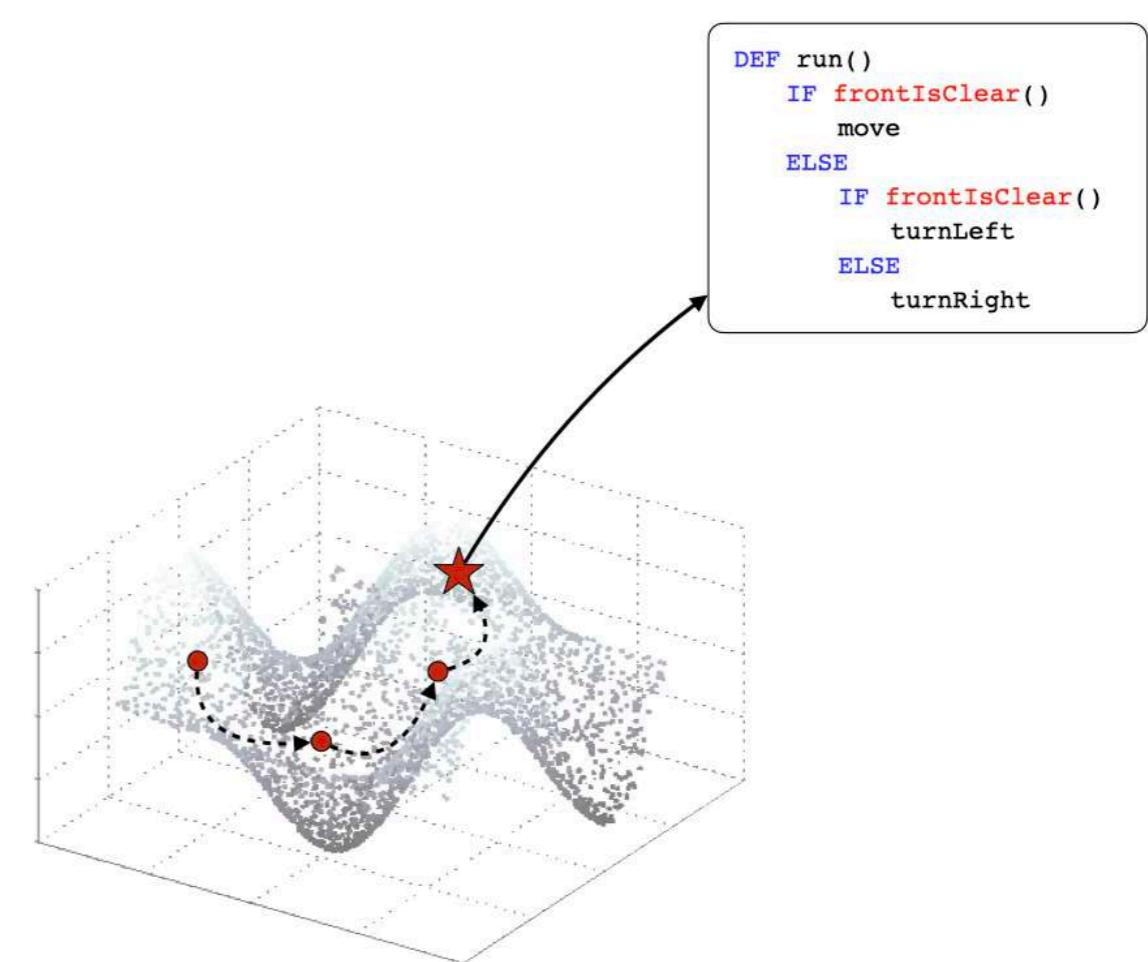
## Stage 1

Learn a program embedding space from randomly generated programs

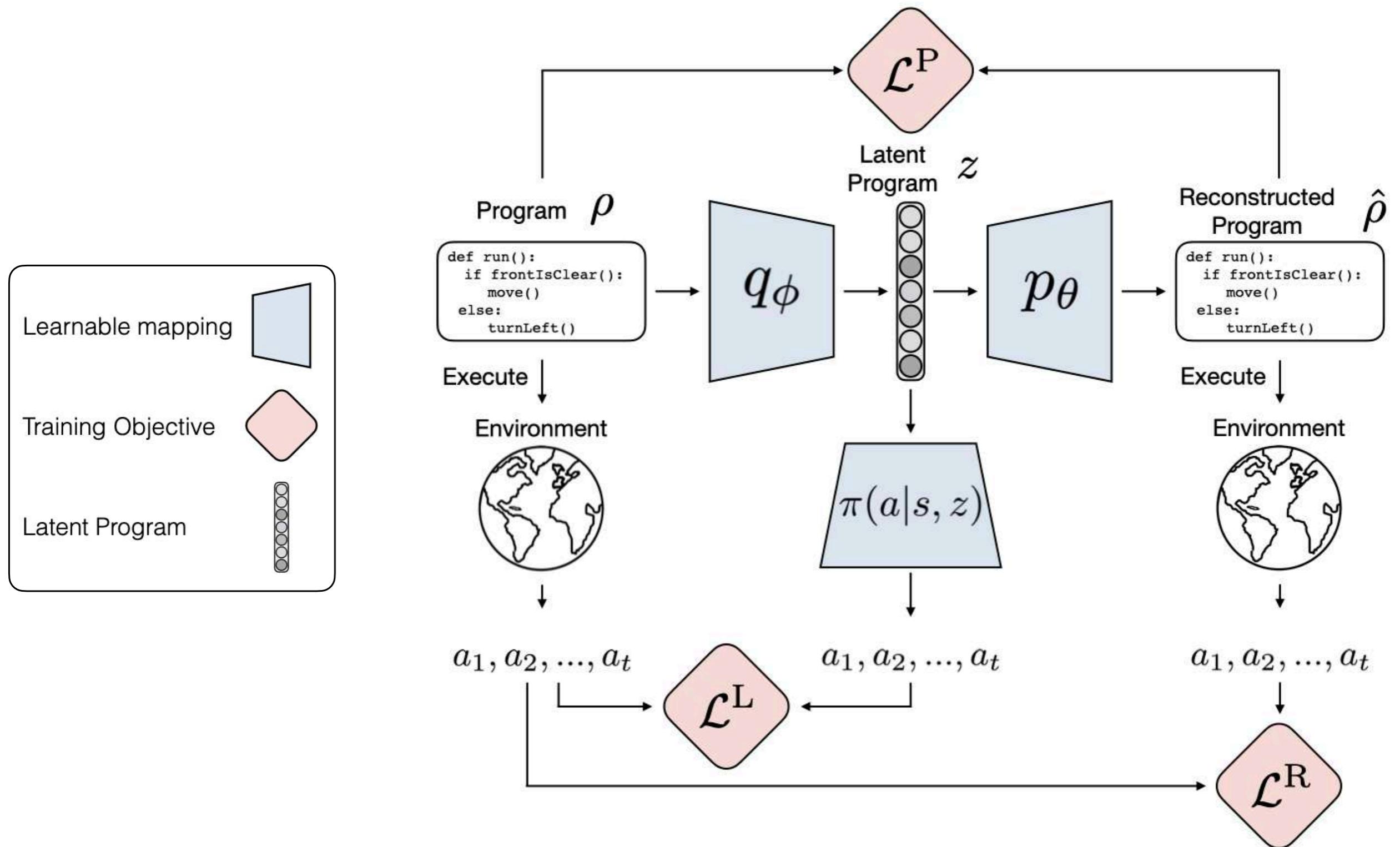


## Stage 2

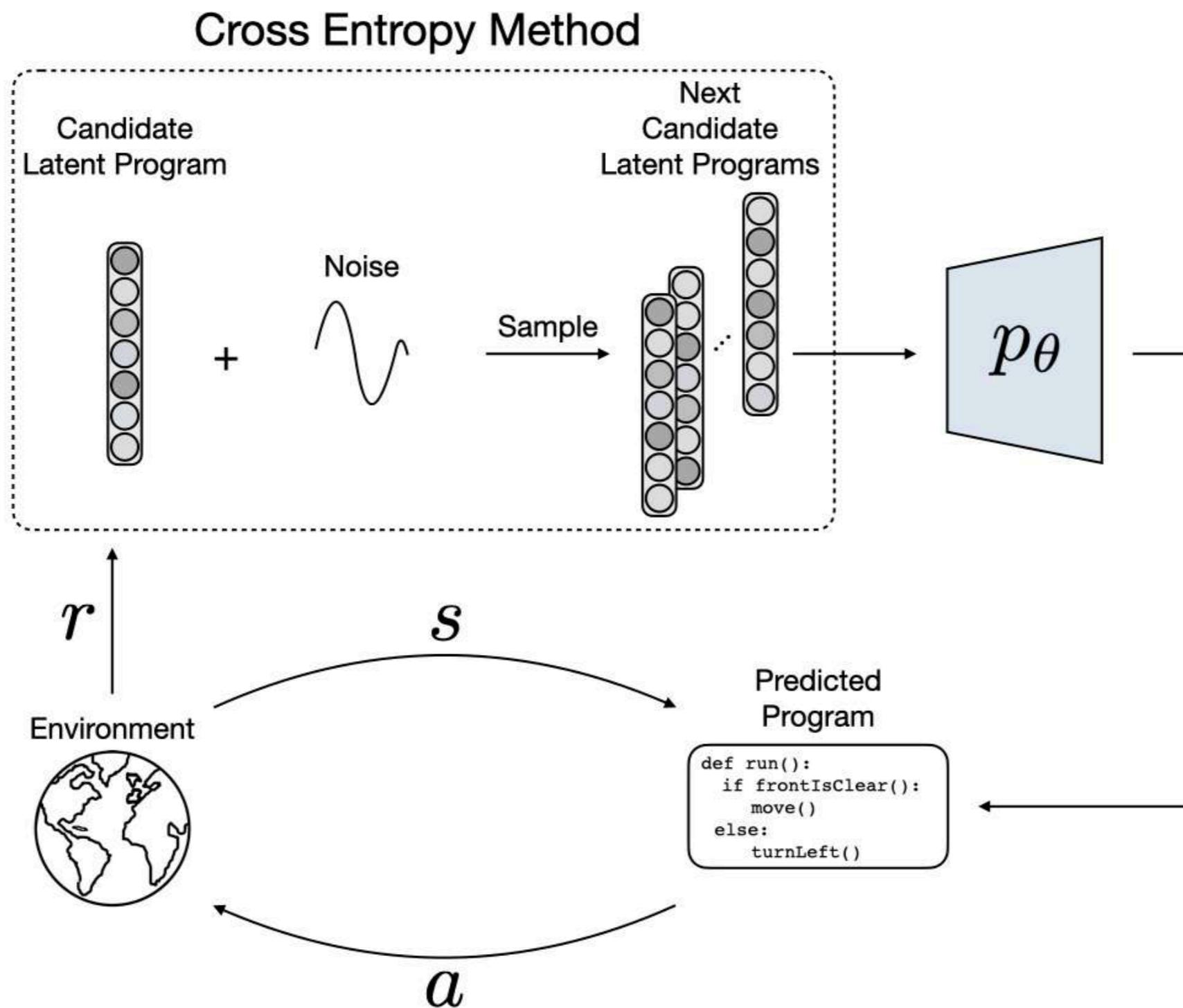
Search for a task-solving program



# Stage 1: Learning a Program Embedding Space

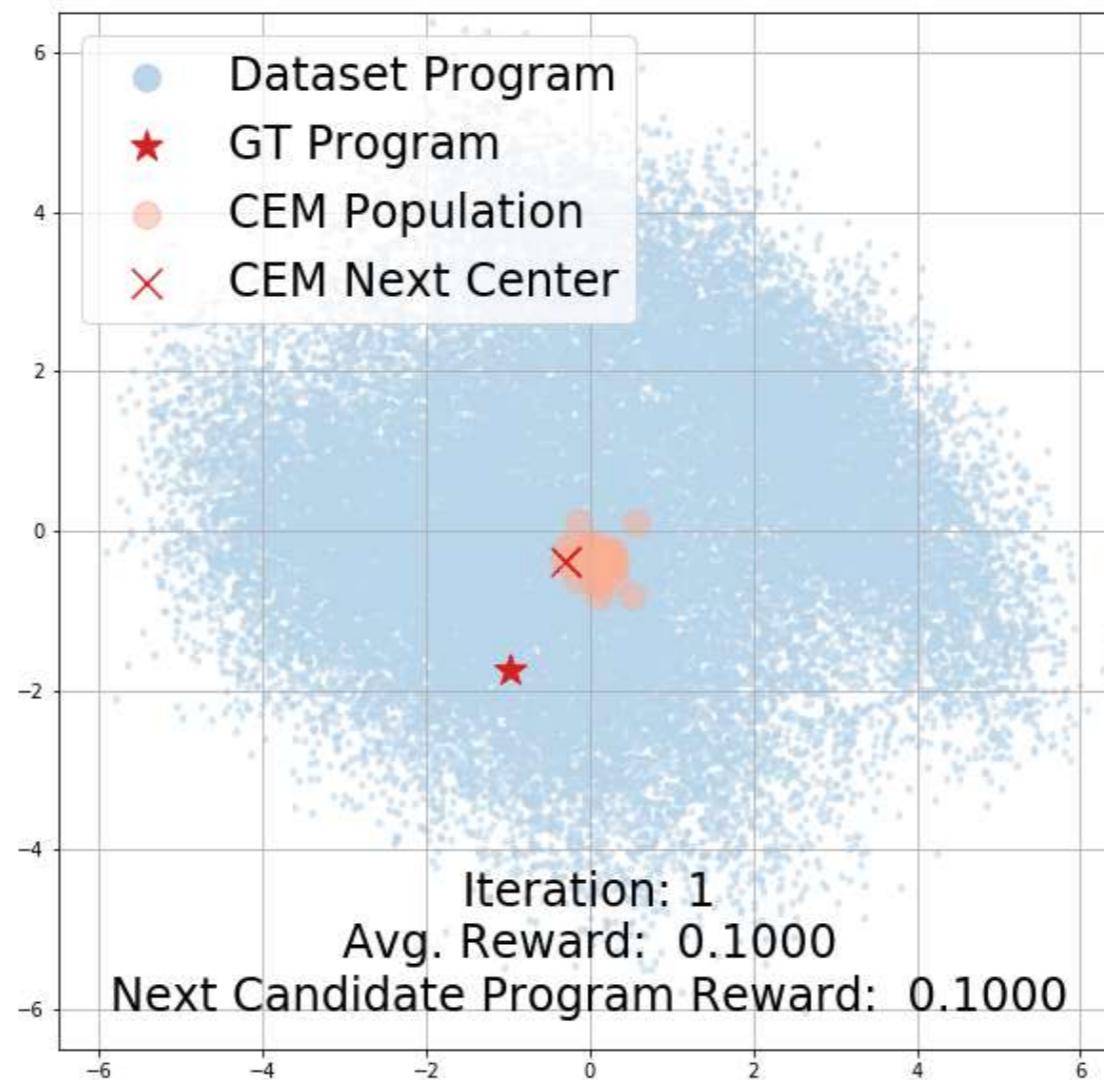


# Stage 2: Latent Program Search with Cross Entropy Method



# Cross Entropy Method Trajectory Visualization

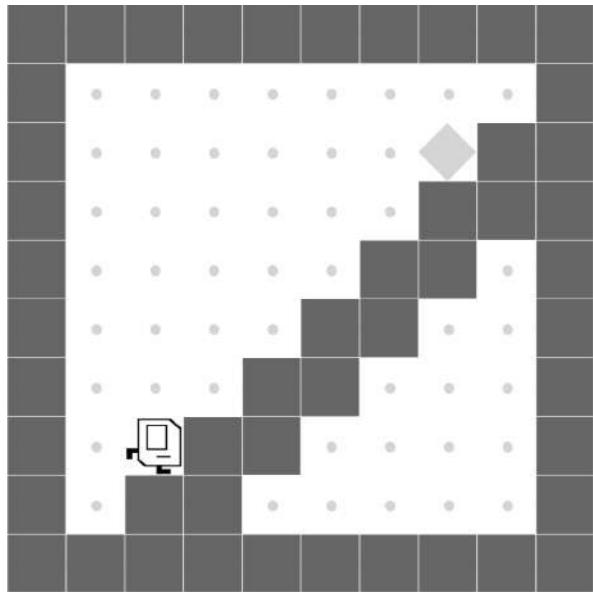
---



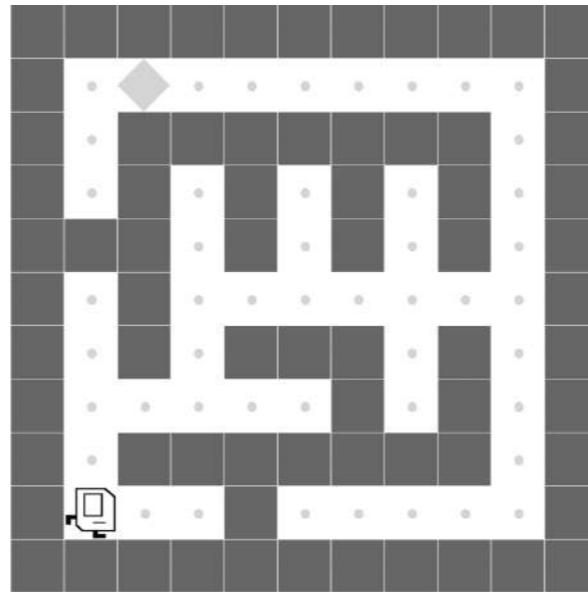
# Karel Tasks

---

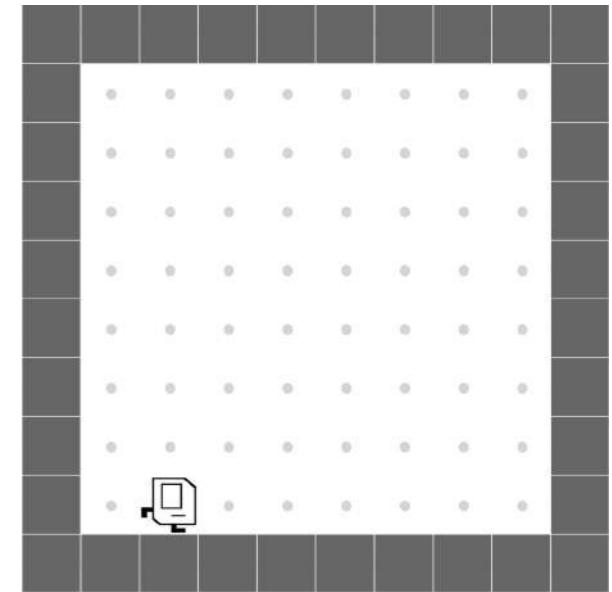
StairClimber



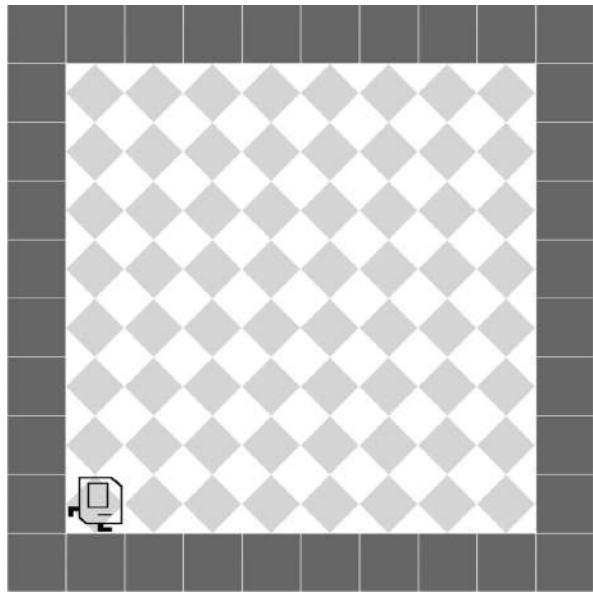
Maze



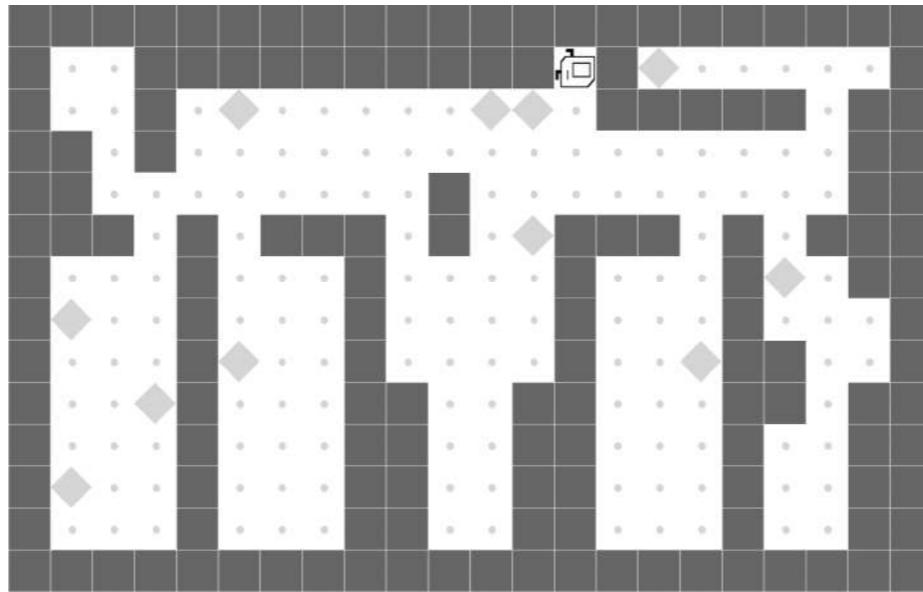
FourCorners



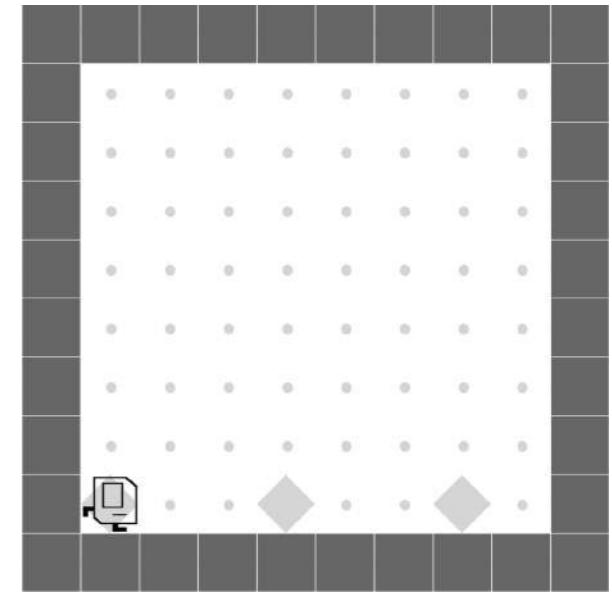
Harvester



CleanHouse



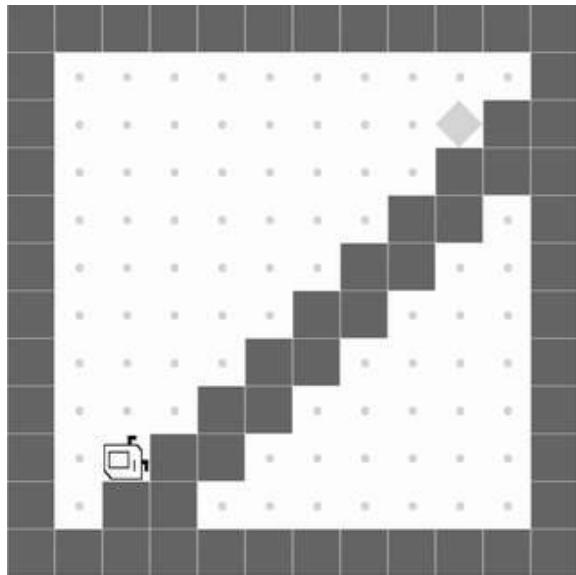
TopOff



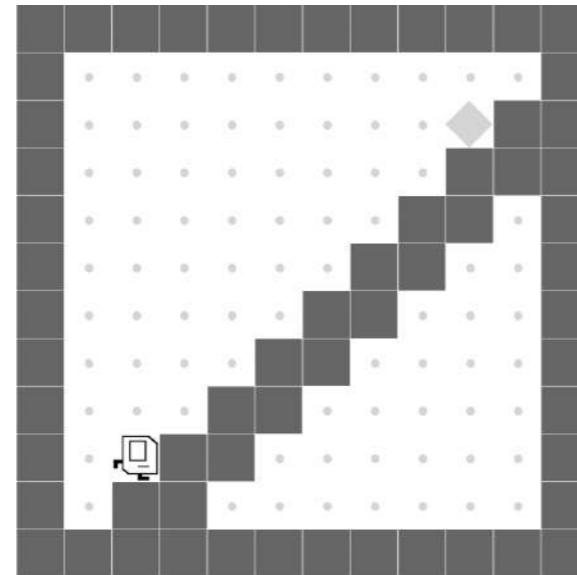
# Qualitative Results

---

StairClimber

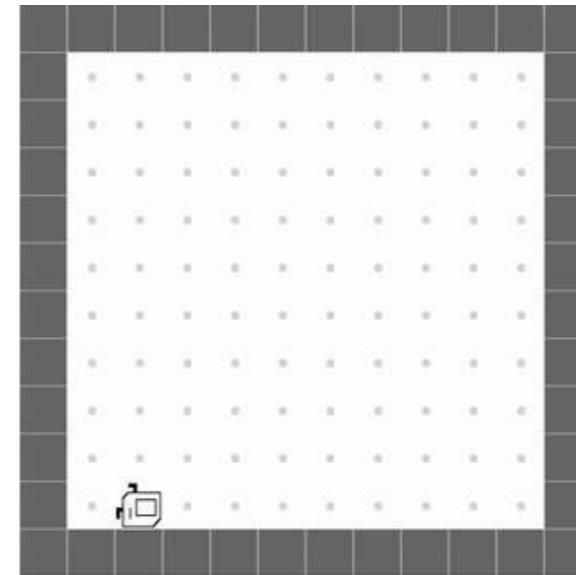


DRL

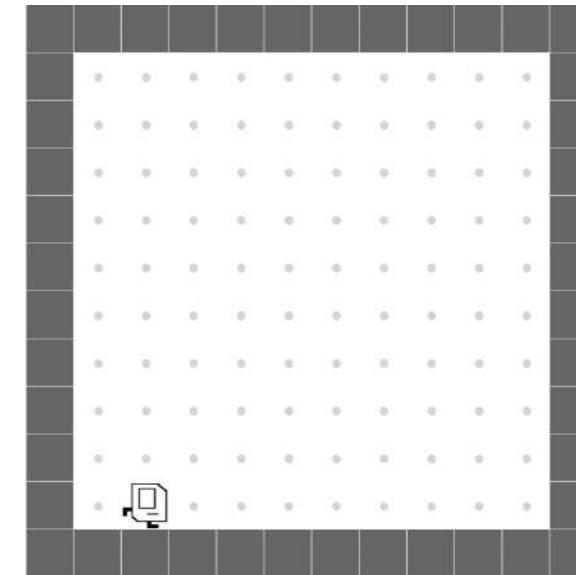


LEAPS

FourCorners

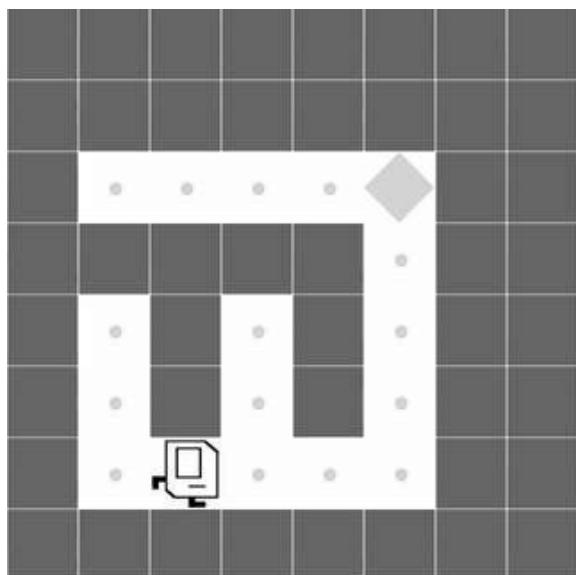


DRL

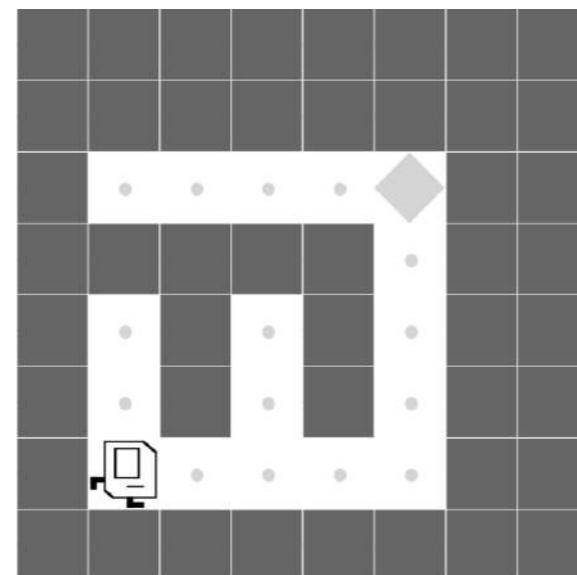


LEAPS

Maze

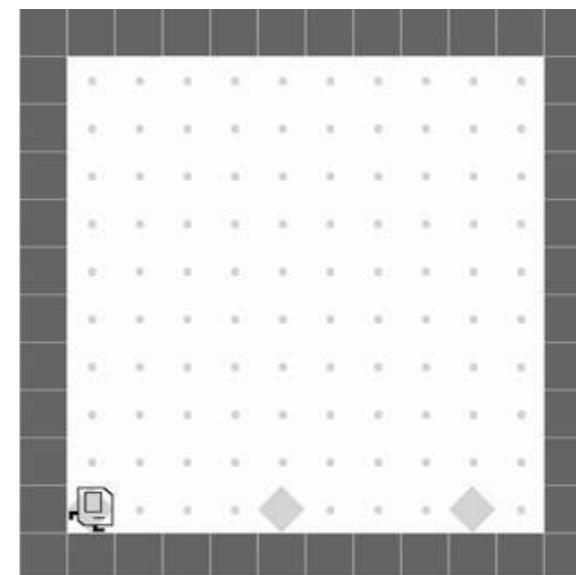


DRL

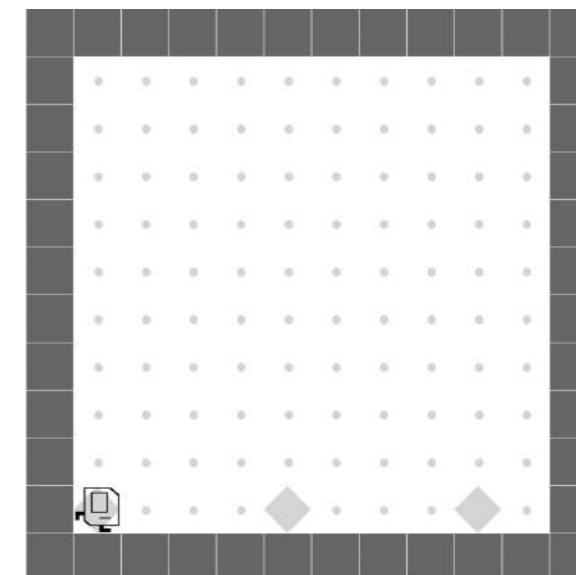


LEAPS

TopOff

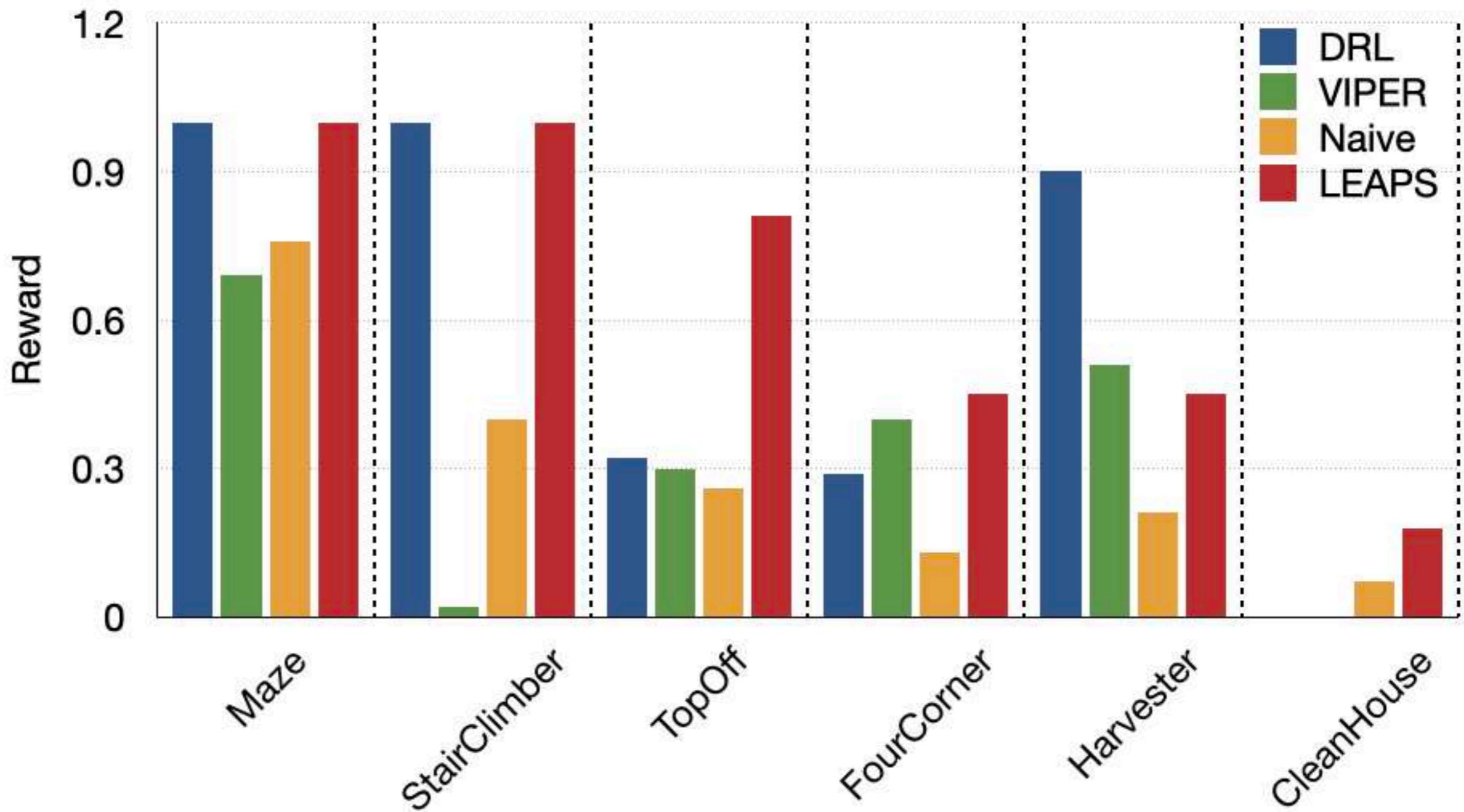


DRL



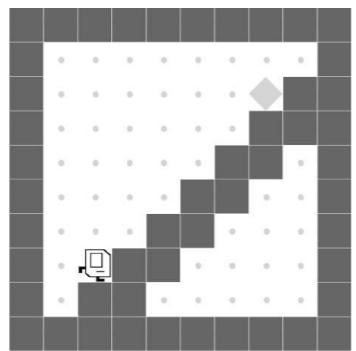
LEAPS

# Quantitative Results



# Zero-shot Generalization

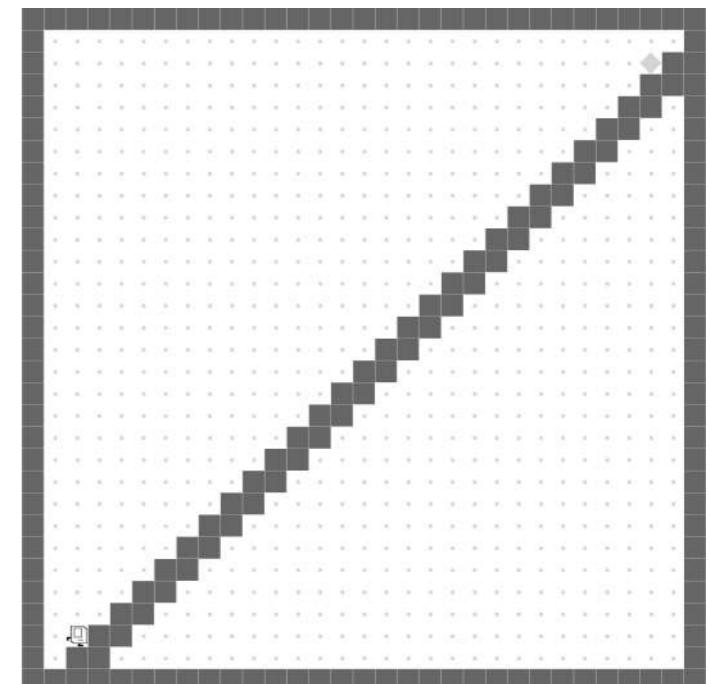
Learning on  
8x8 grids



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

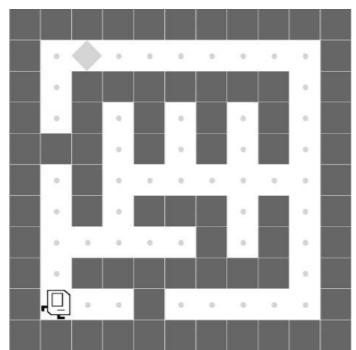


Evaluation on  
100x100 grids

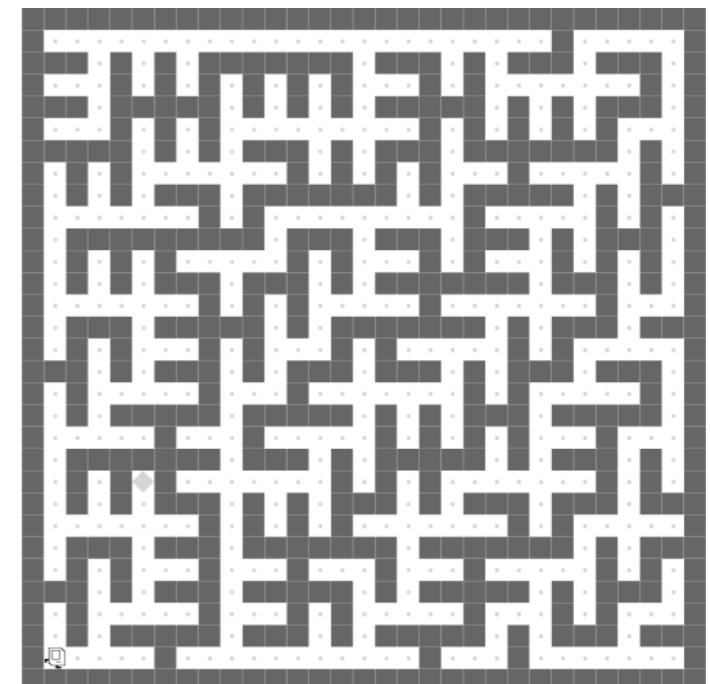


StairClimber

Maze

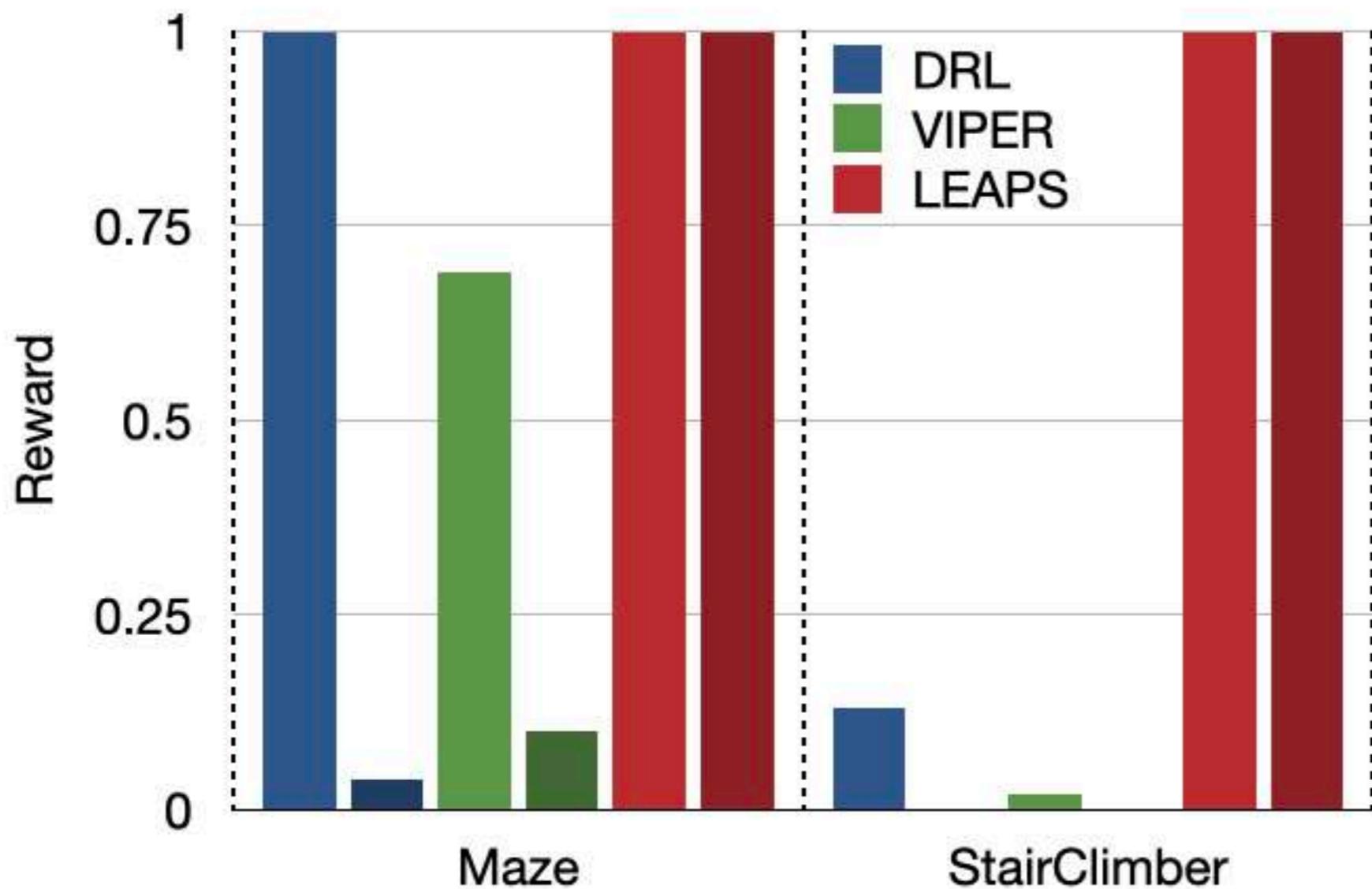


```
DEF run()
  IF frontIsClear()
    turnLeft
  WHILE noMarkersPresent()
    turnRight
    move
```



# Zero-shot Generalization

---

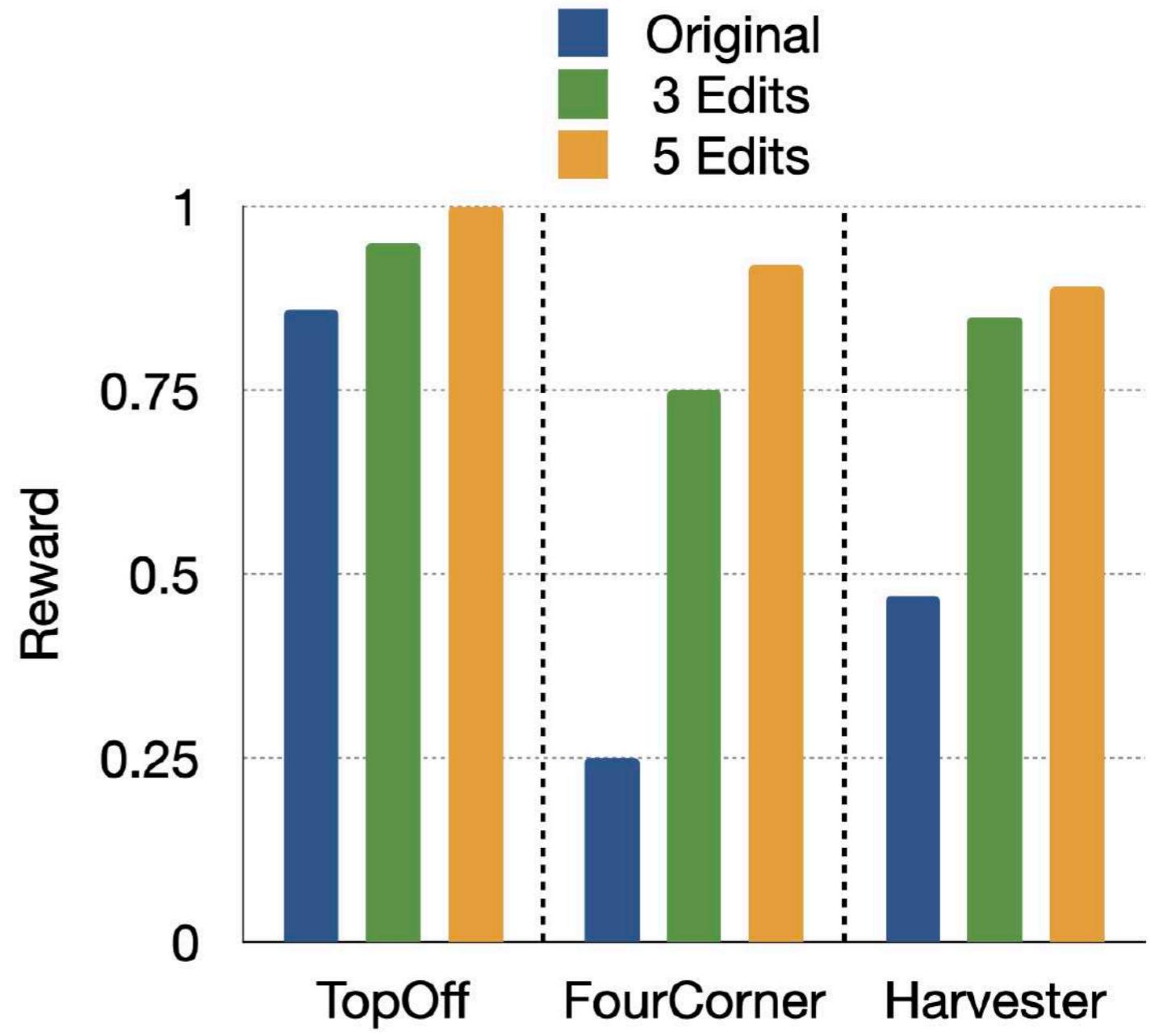


# Interpretability

## Human Debugging Interface



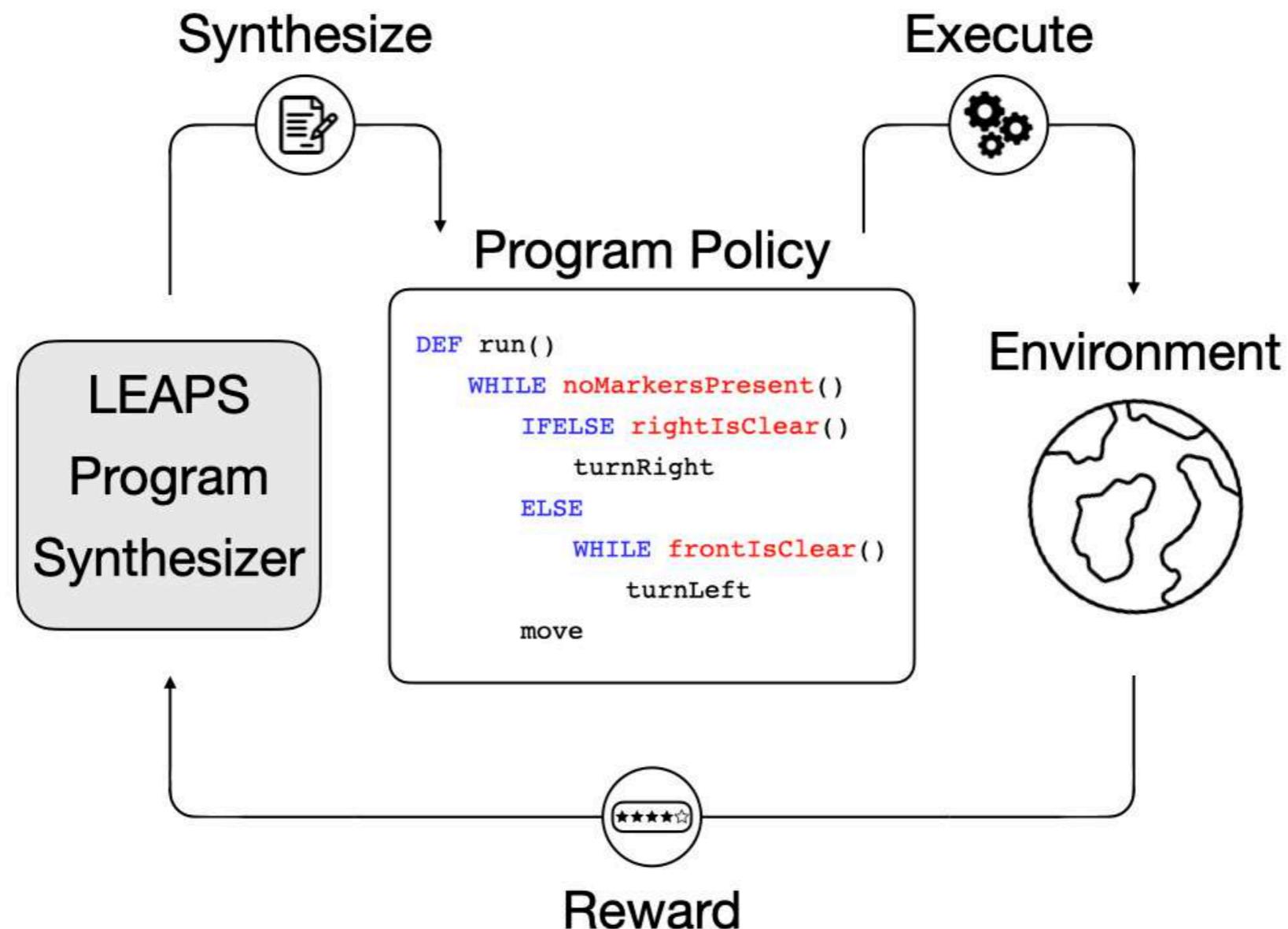
## Improved Performance



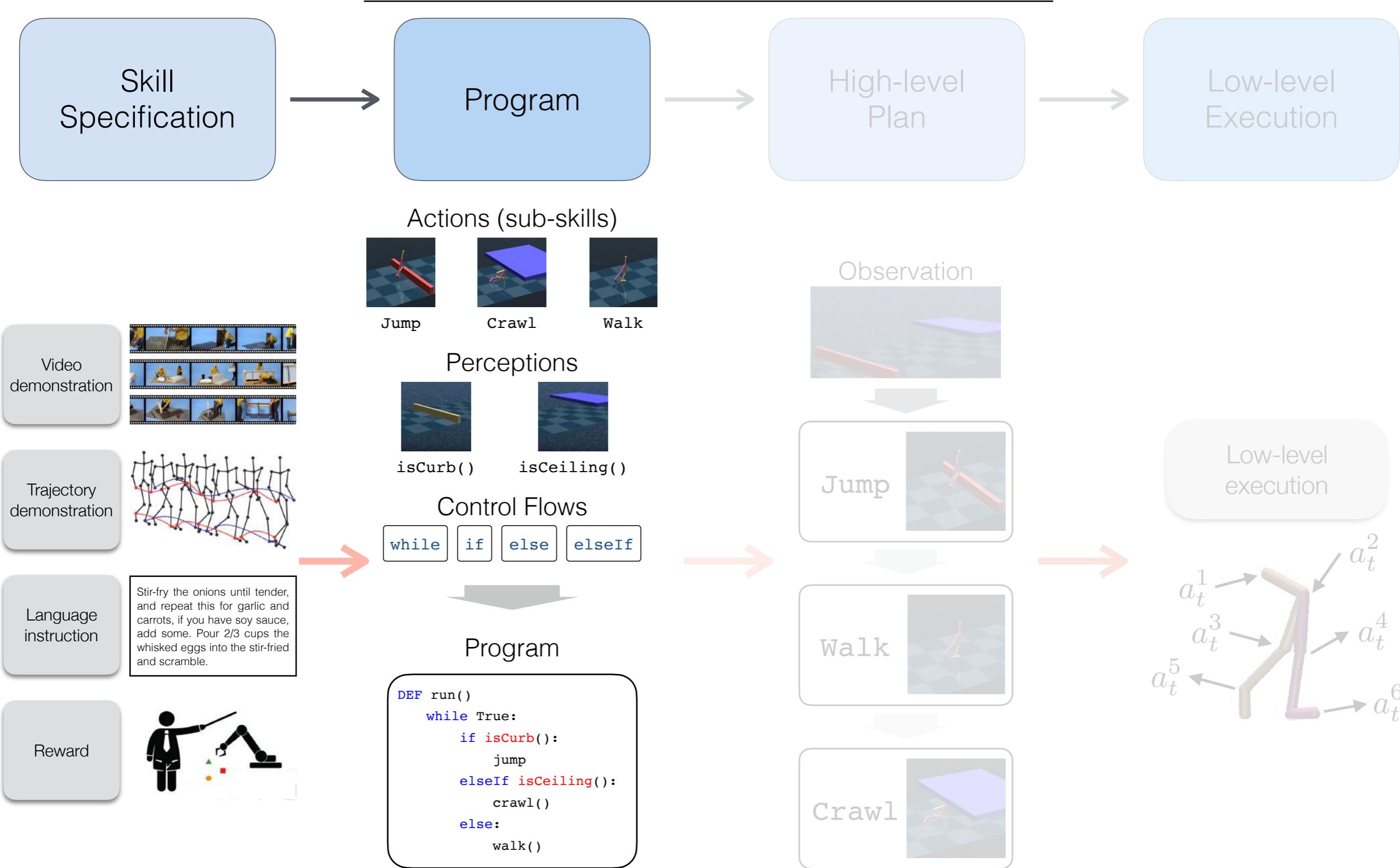
# Takeaway

---

- Synthesize generalizable and interpretable programs from rewards



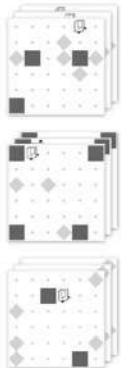
# Program Inference



# Program Inference

Imitation learning from demonstrations

Demonstrations



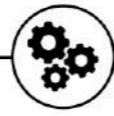
Synthesize



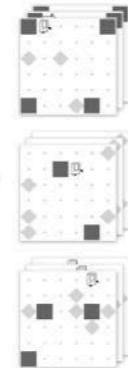
Program

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

Execute

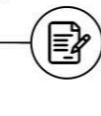


Execution



Reinforcement learning from rewards

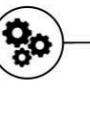
Synthesize



Program Policy

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

Execute



Environment

Reward

LEAPS  
Program  
Synthesizer

# Program Inference

Imitation learning from demonstrations

Demonstrations



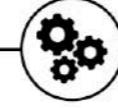
Synthesize



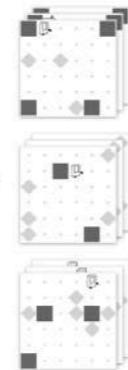
Program

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

Execute



Execution



Reinforcement learning from rewards

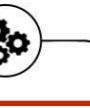
Synthesize



Program Policy

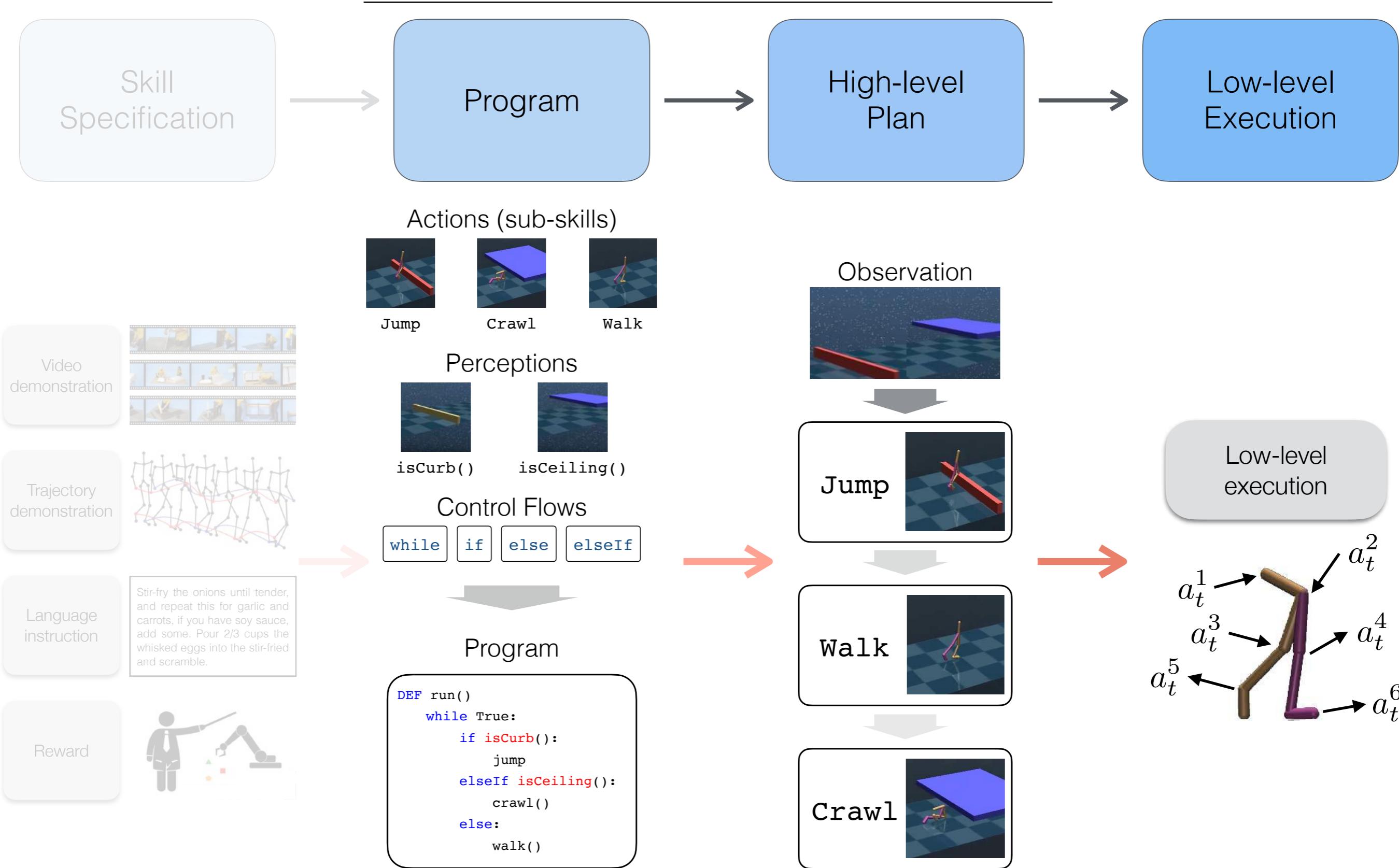
```
DEF run()
  WHILE noMarkersPresent()
    IFELSE rightIsClear()
      turnRight
    ELSE
      WHILE frontIsClear()
        turnLeft
      move
```

Execute



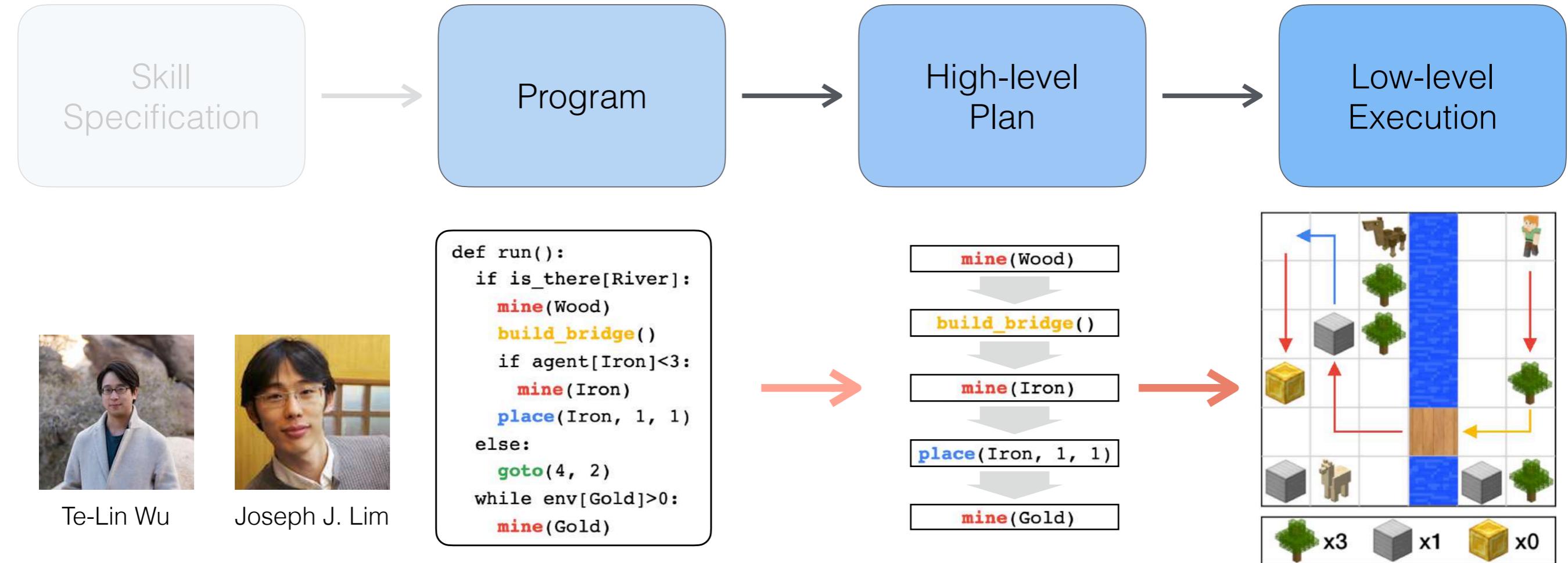
Reward

# Task Execution



# Program Guided Agent

ICLR 2020 (Spotlight)



Te-Lin Wu



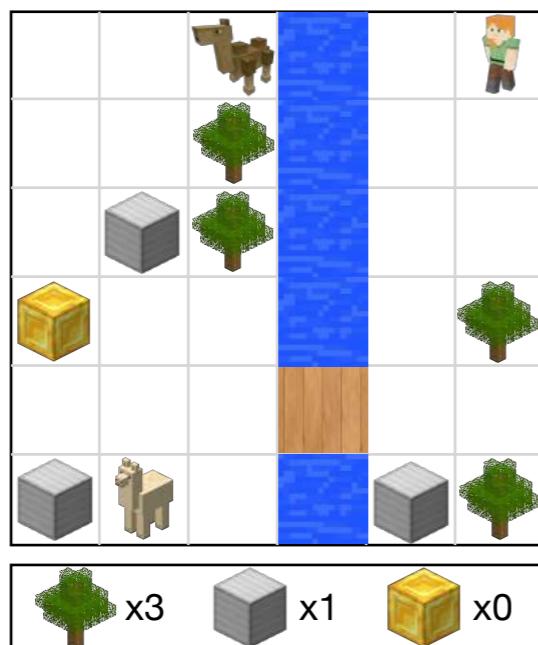
Joseph J. Lim

# Problem Formulation

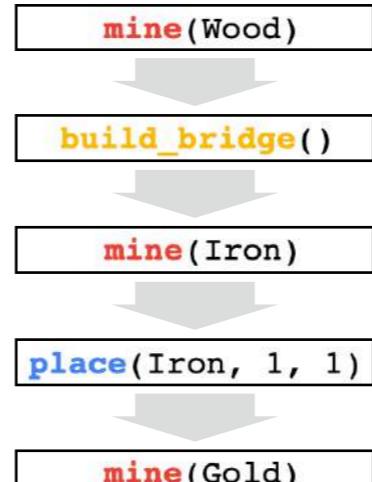
Program  
(task)

```
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)
```

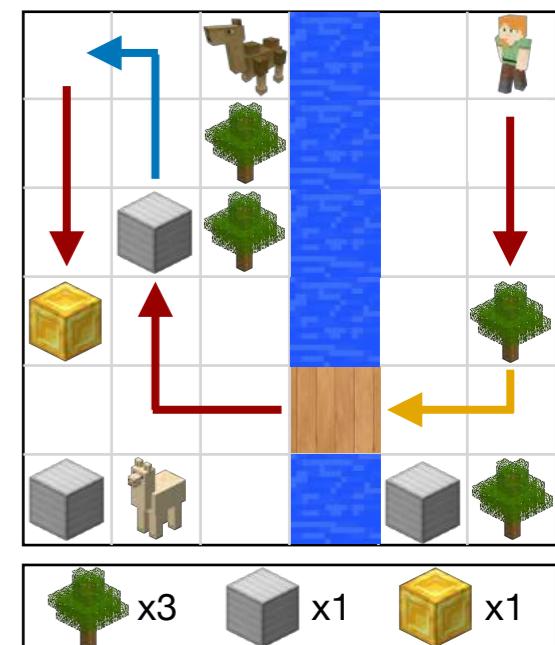
Observation



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)
```

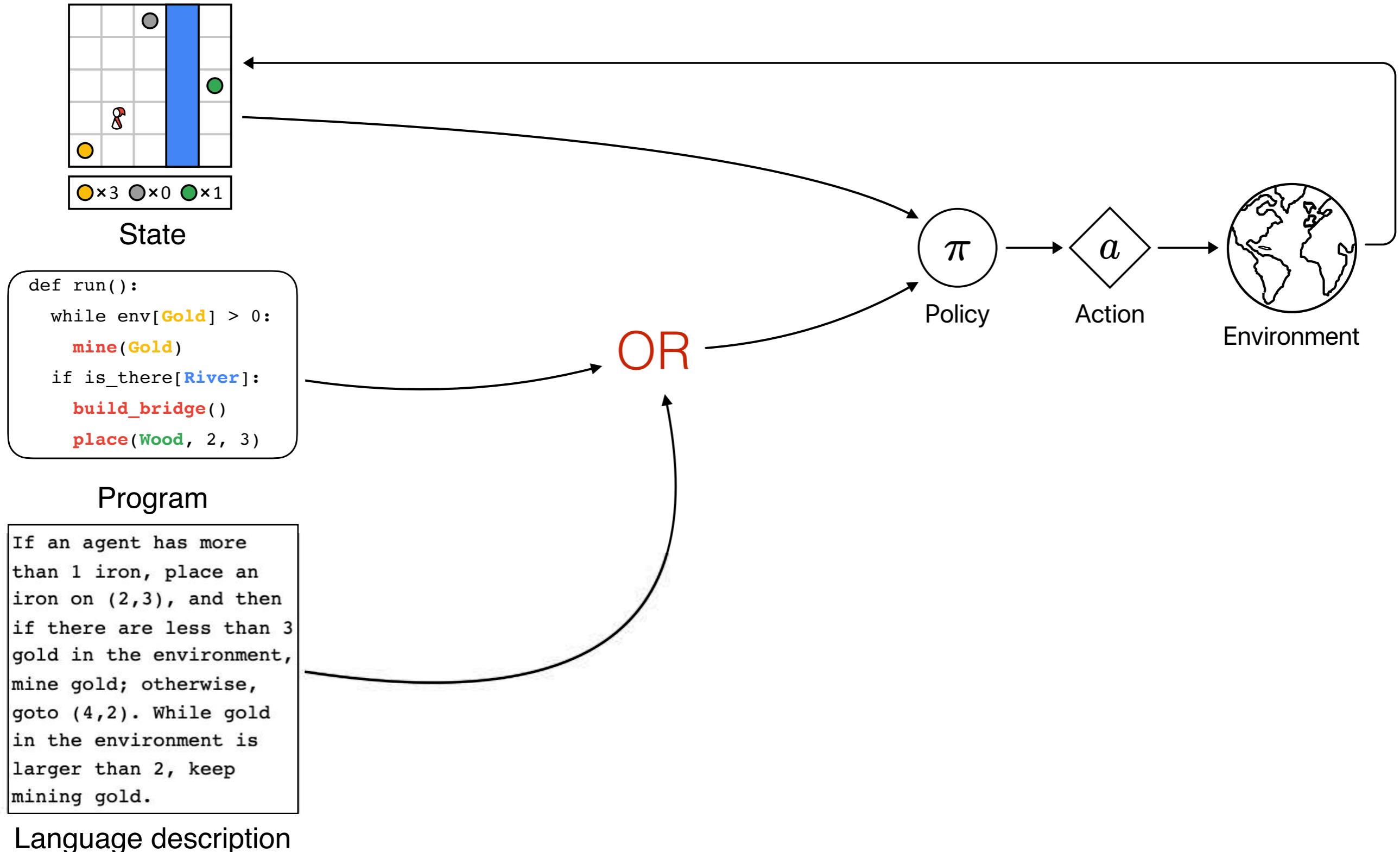
```
def run()
    while agent[Wood] <= 11:
        place(Wood, 2, 4)
        place(Iron, 1, 1)
        place(Iron, 8, 5)
        mine(Gold)
        mine(Gold)
        mine(Gold)
        repeat(4):
            sell(Gold)
            sell(Iron)
```

## Natural Language Descriptions

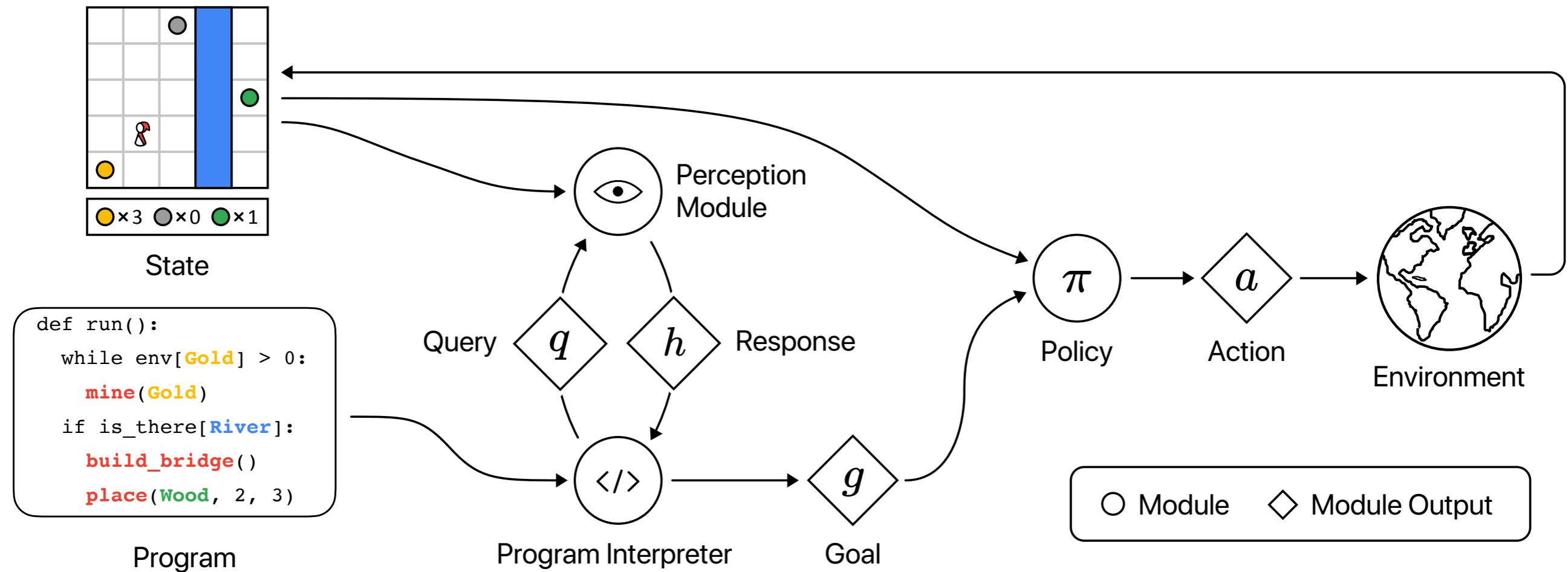
If a river is in the environment, mine a wood and then use it to build a bridge. And then if agent has less than there iron, place an iron at (2,3). Otherwise if no river, goto location (4,2). Finally, whenever there's still gold in the environment, mine a gold.

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



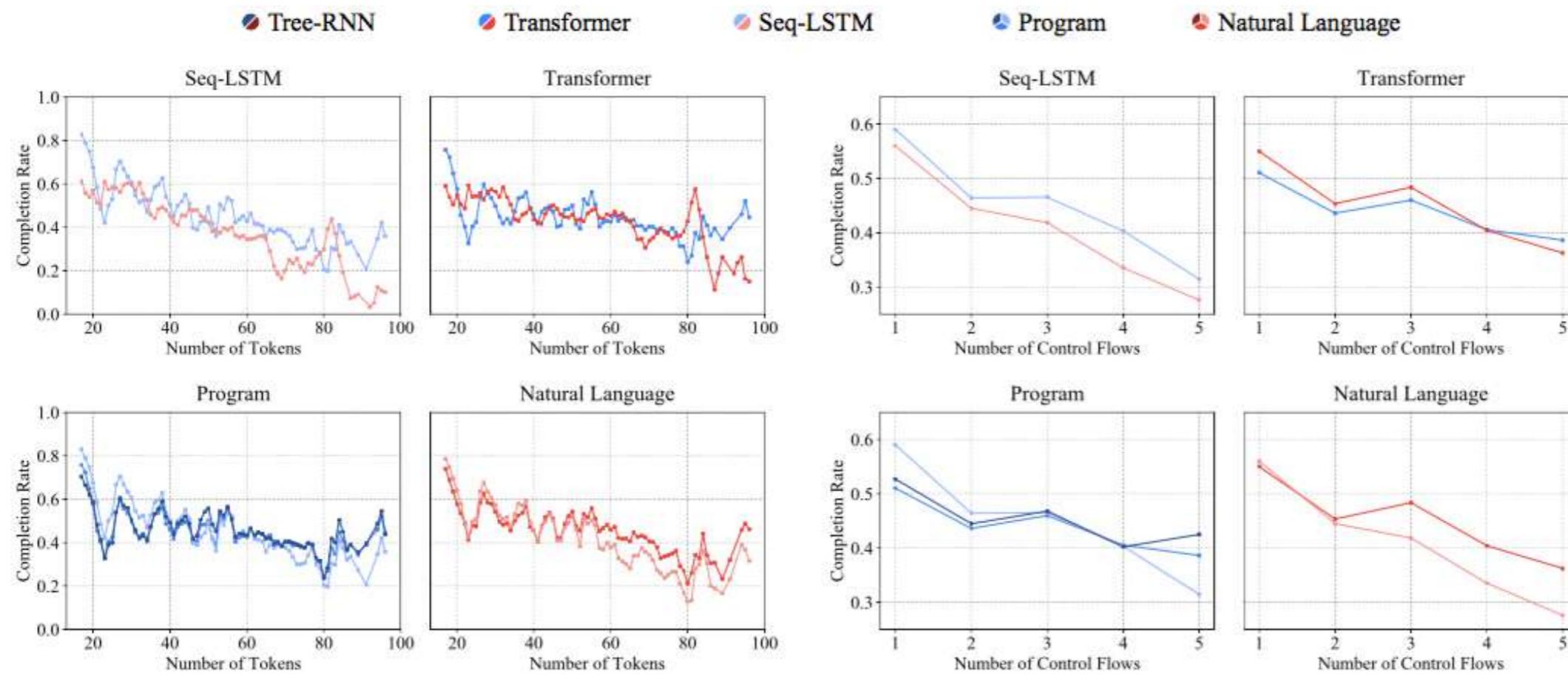
# Program Guided Agent



# 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%



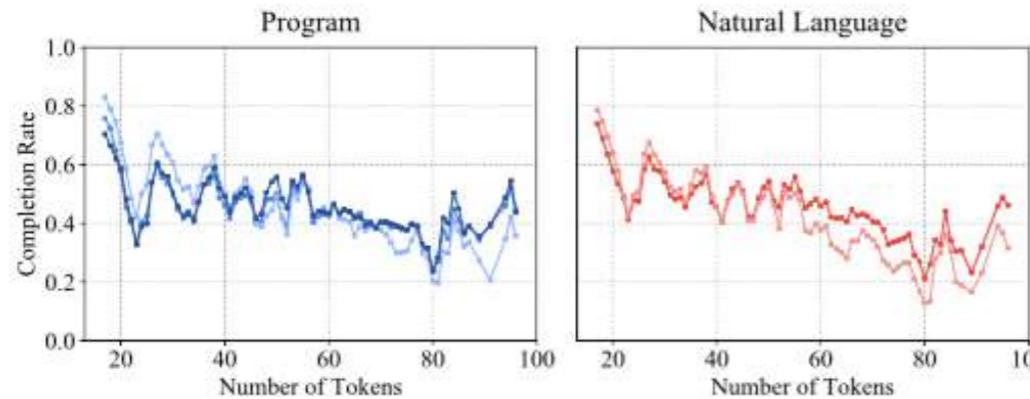
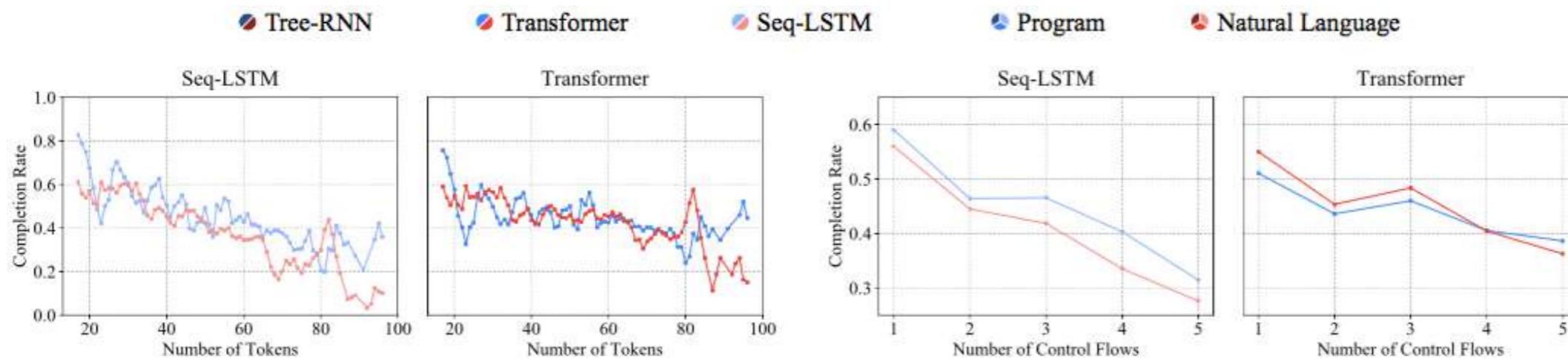
(a) Instruction Length

(b) Instruction Complexity

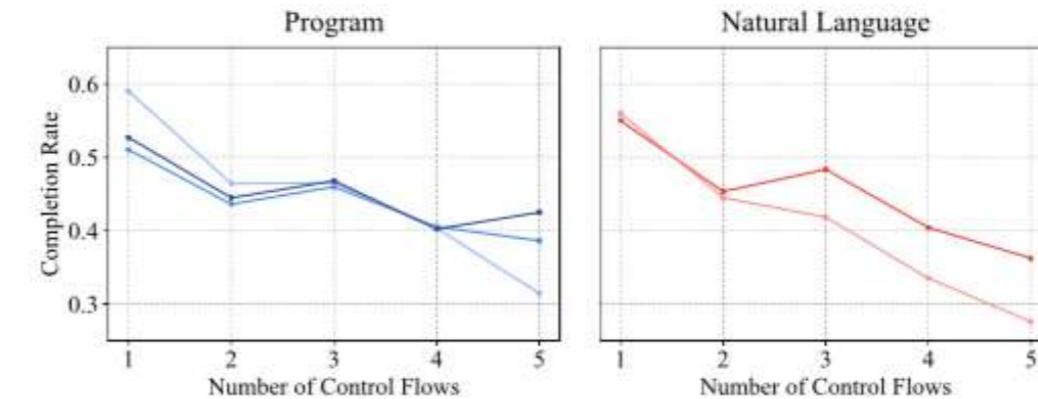
# Quantitative Results

## Natural Languages < Programs

Instruction Method		Natural language descriptions		Programs				
		Seq-LSTM	Transformer	Seq-LSTM	Tree-RNN	Transformer	Ours (concat)	Ours
<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

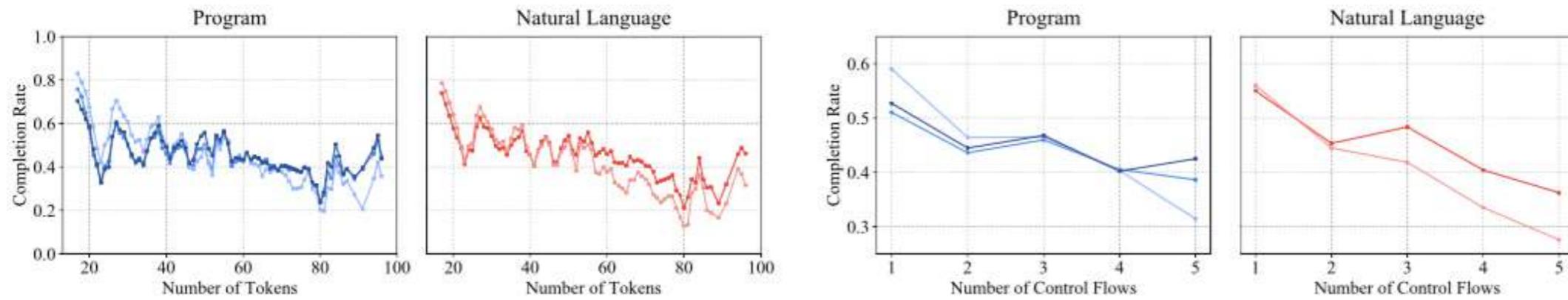
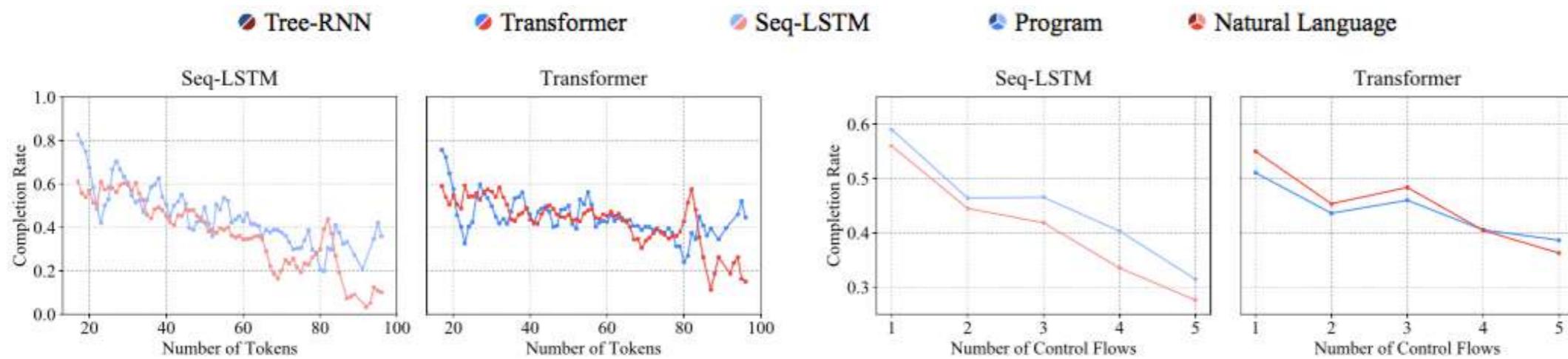


(b) Instruction Complexity

# Quantitative Results

End-to-end < Ours (modular)

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%
Dataset 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

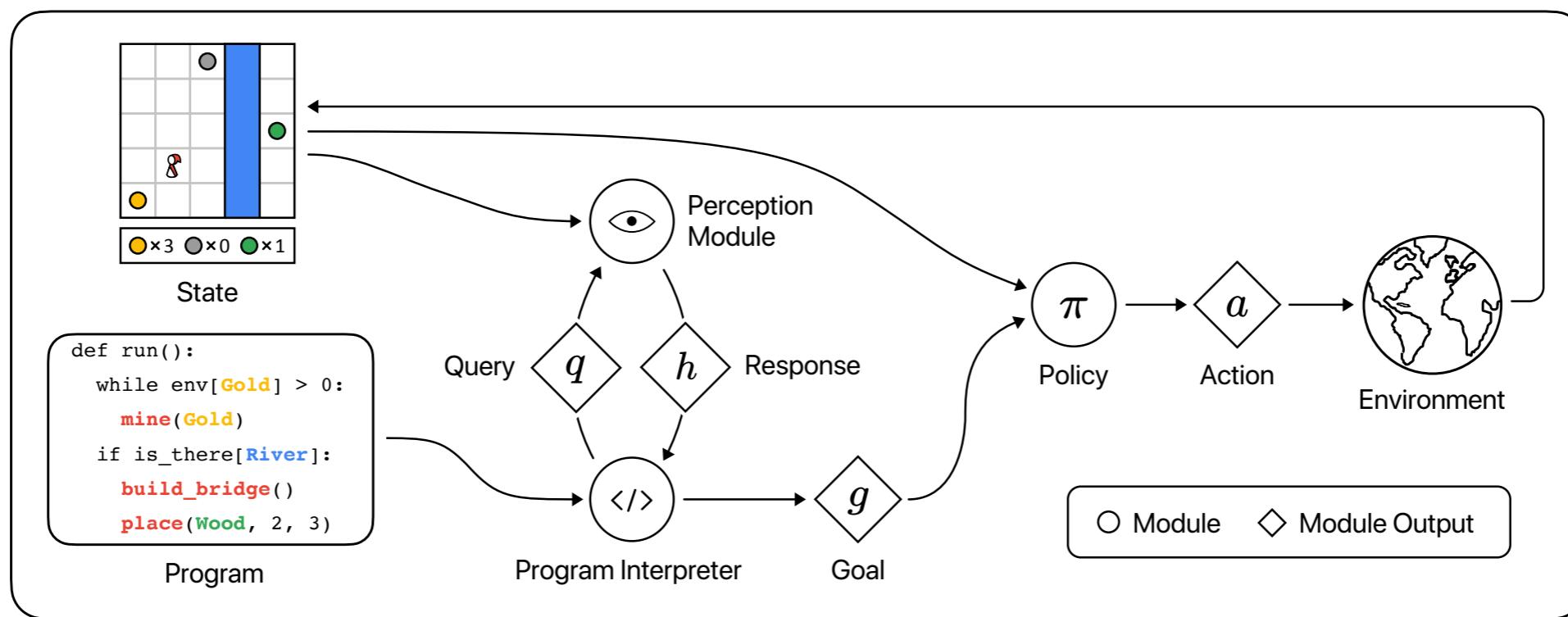
# Takeaway

- Specific tasks using programs

Program

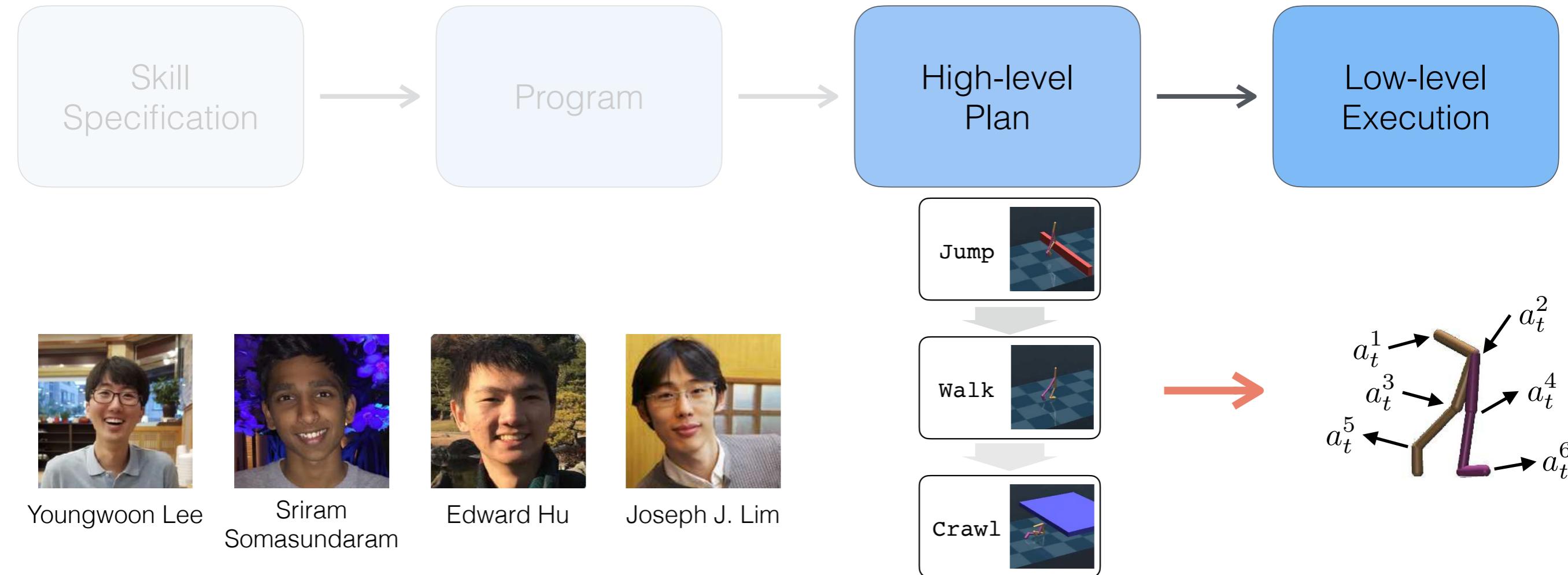
```
def run():
    if is_there[River]:
        mine(Wood)
        build_bridge()
    if agent[Iron]<3:
        mine(Iron)
        place(Iron, 1, 1)
    else:
        goto(4, 2)
    while env[Gold]>0:
        mine(Gold)
```

- Leverage the structure of programs with a modular framework



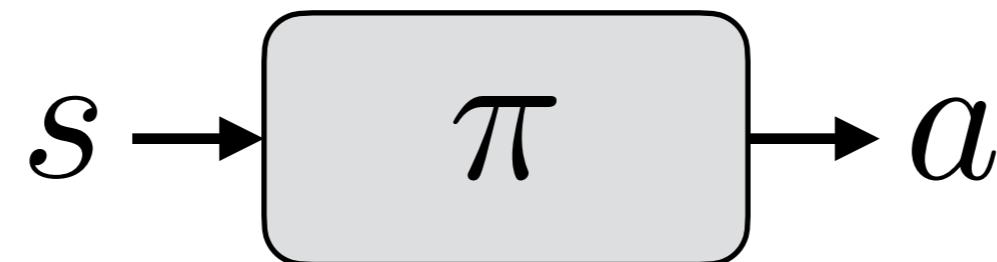
# Composing Complex Skills by Learning Transition Policies

ICLR 2019

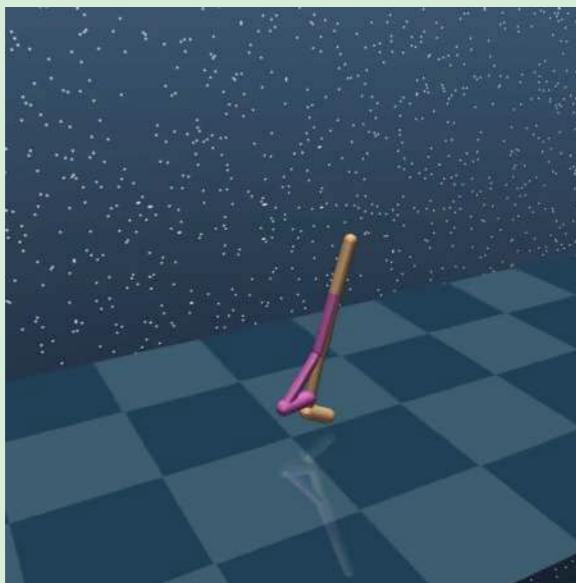


# Learned Skills

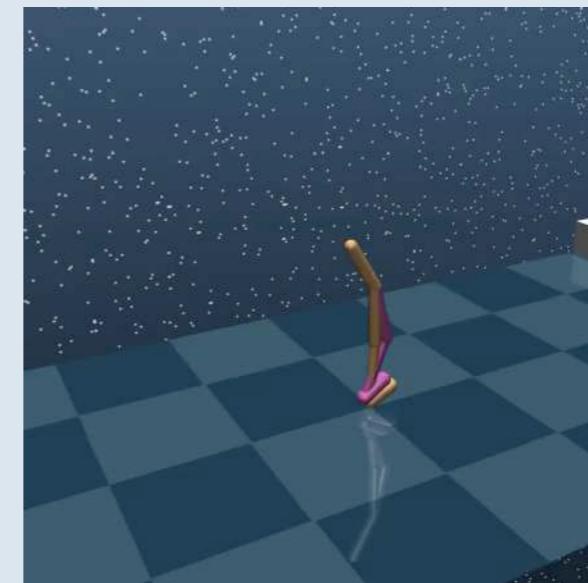
---



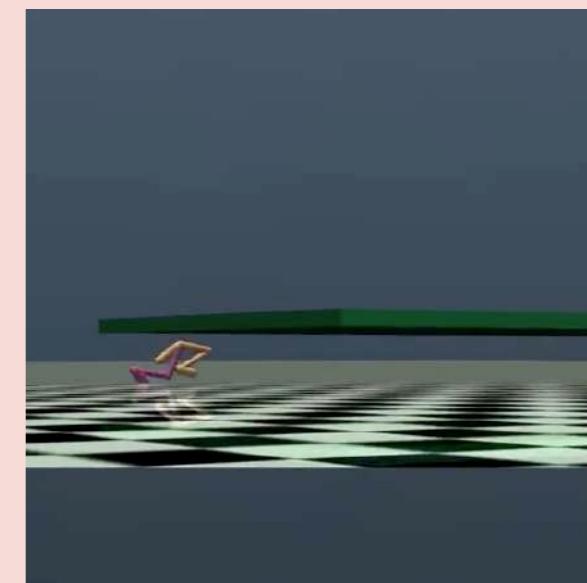
Walk



Jump



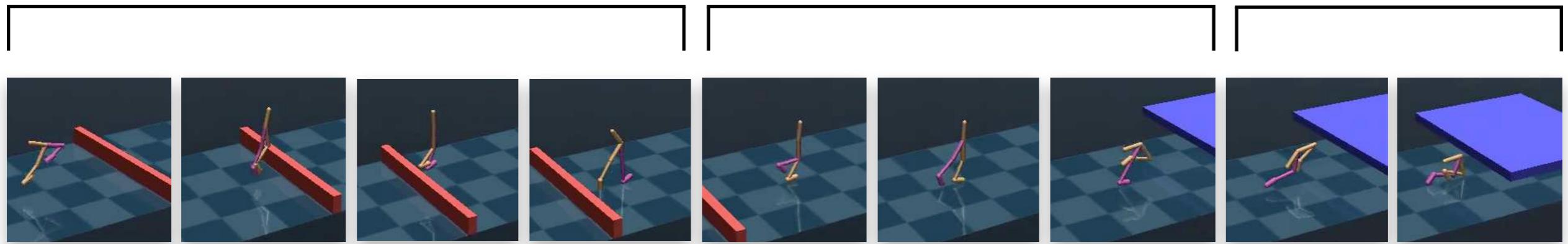
Crawl



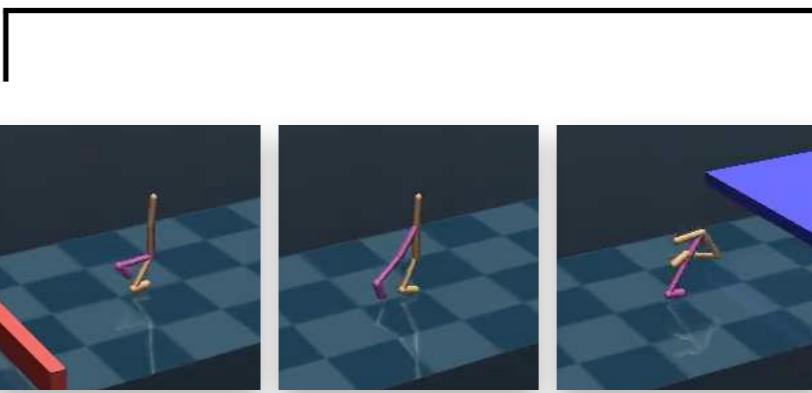
# Compose Complex Skills

High-level plan

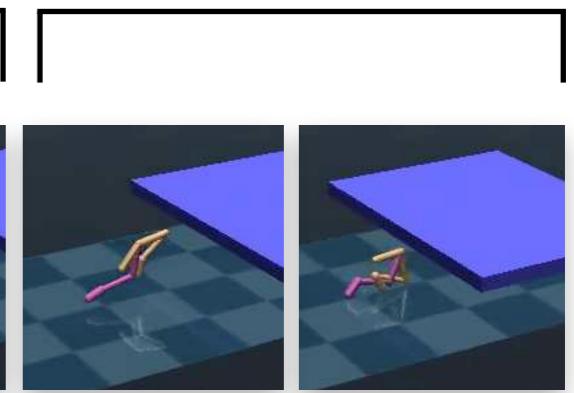
Jump



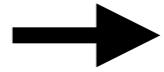
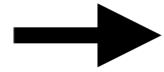
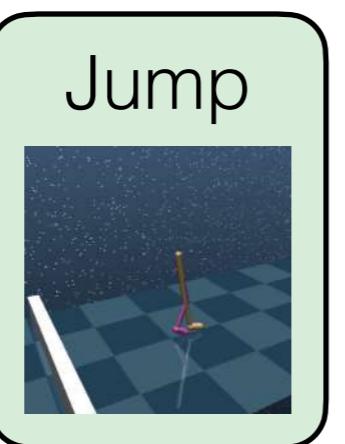
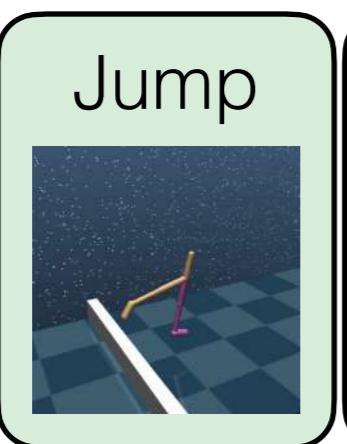
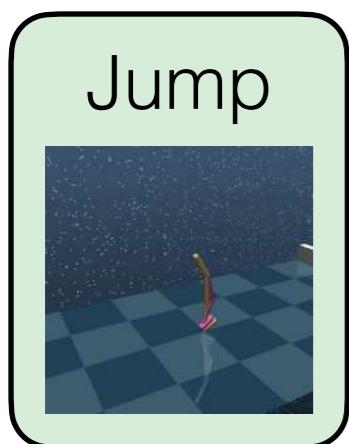
Walk



Crawl



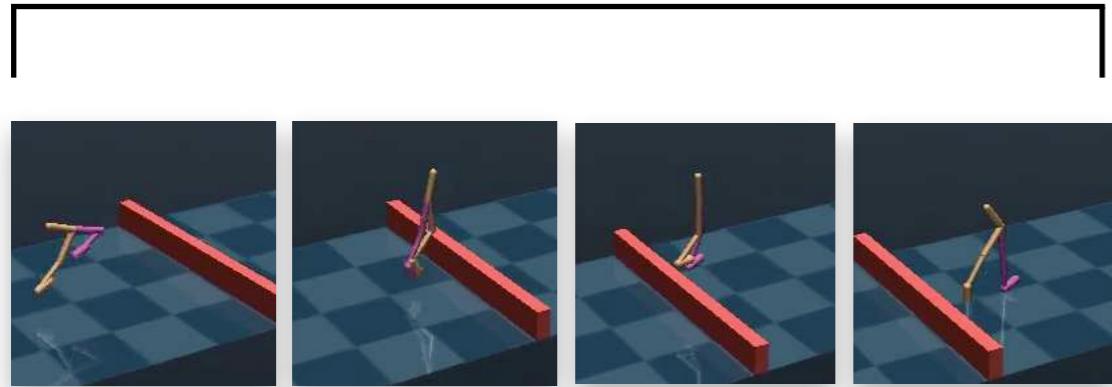
Sequentially execute corresponding policies



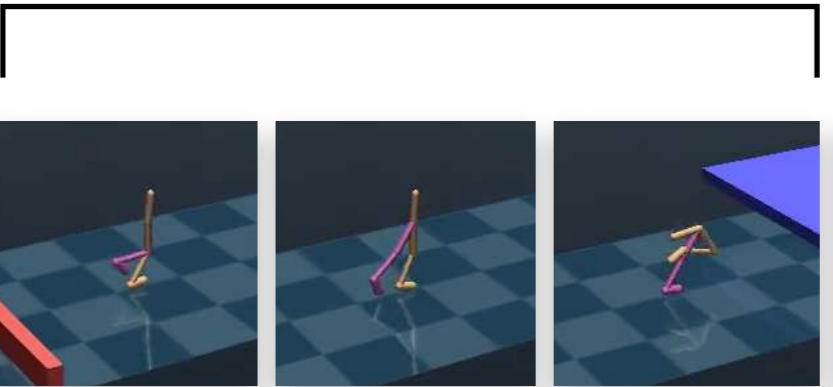
# Compose Complex Skills

High-level plan

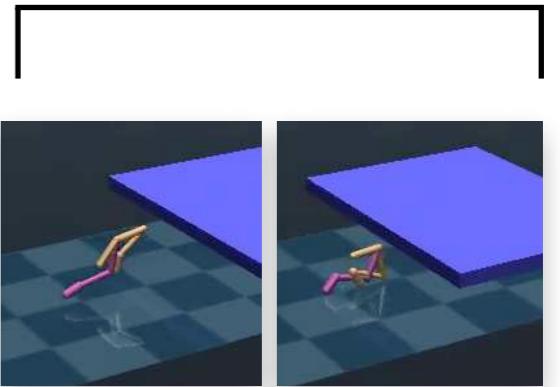
Jump



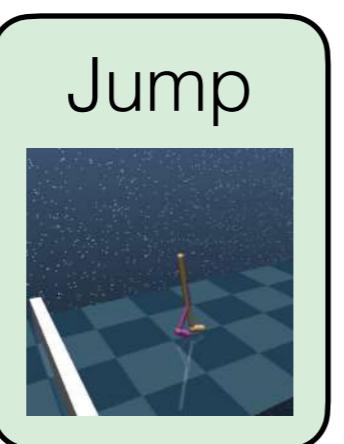
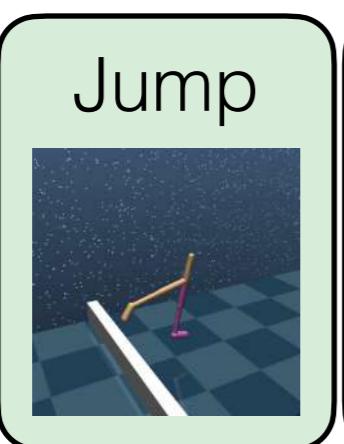
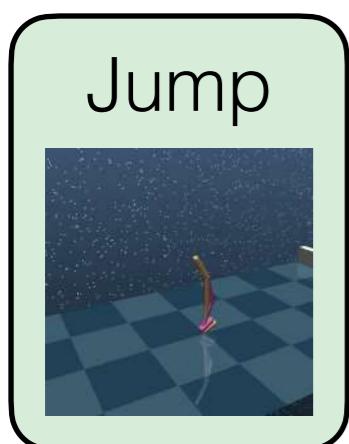
Walk



Crawl



Sequentially execute corresponding policies

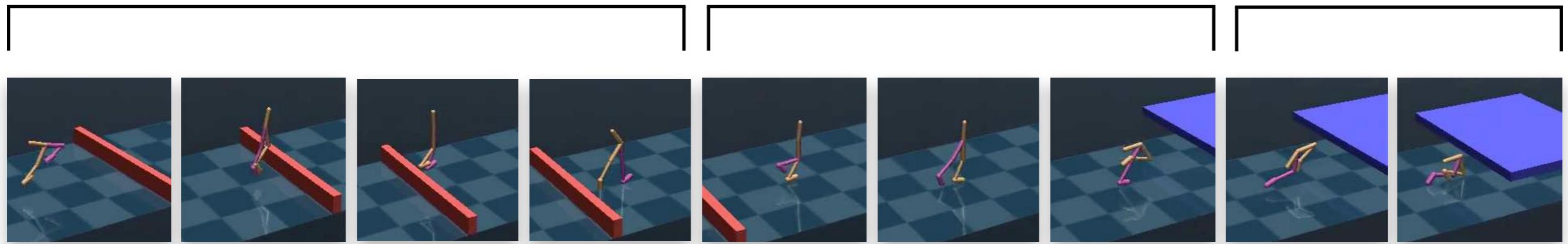


An **end state** of a previous policy might not be a good **initial state** of the following policy

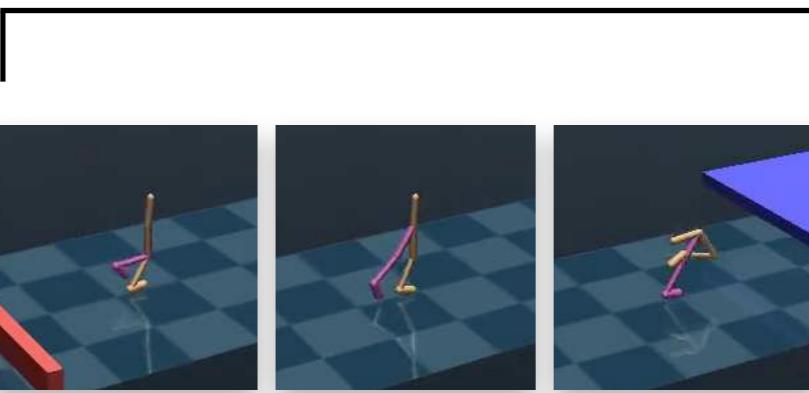
# Compose Complex Skills

High-level plan

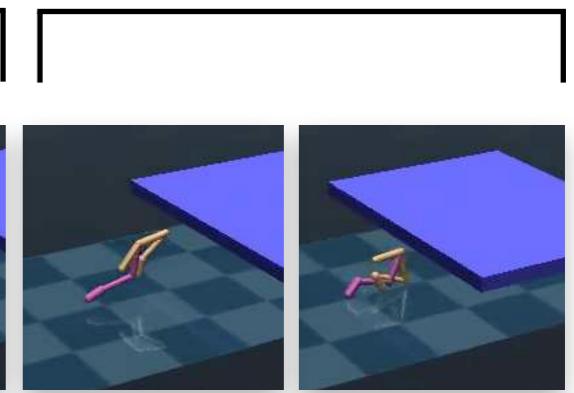
Jump



Walk



Crawl

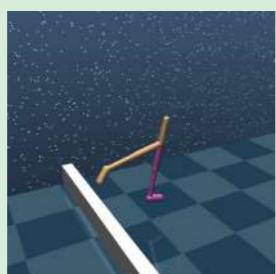


Sequentially execute corresponding policies

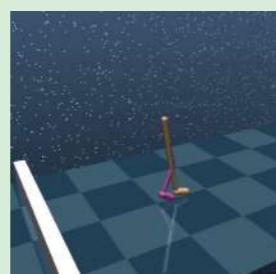
Jump



Jump



Jump



Trans  
 $\pi$

Walk



Walk



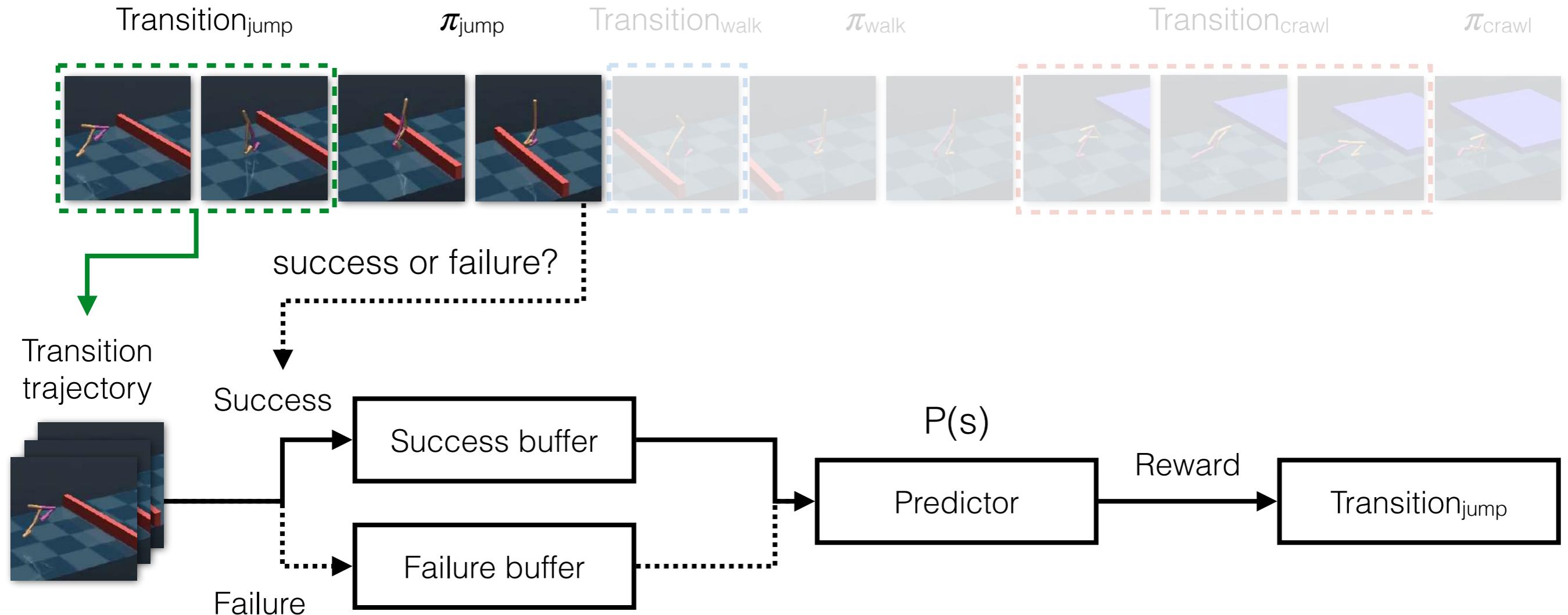
Trans  
 $\pi$

Crawl



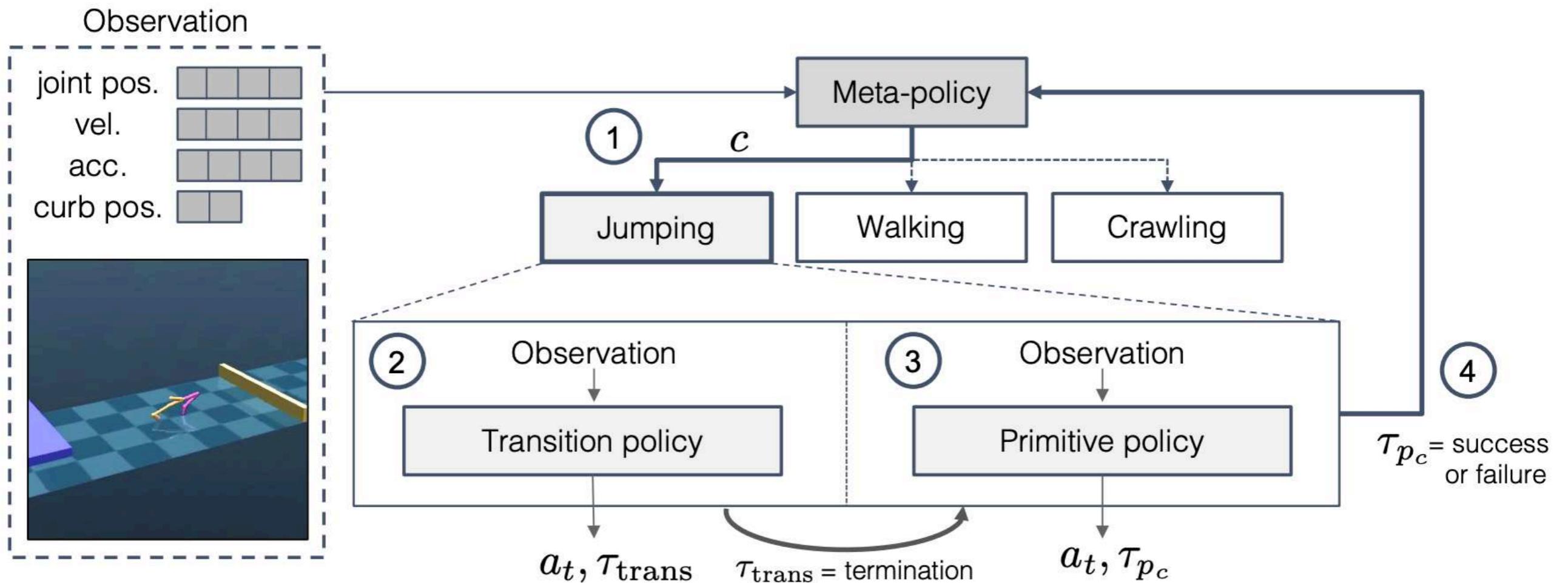
Transition policies

# 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

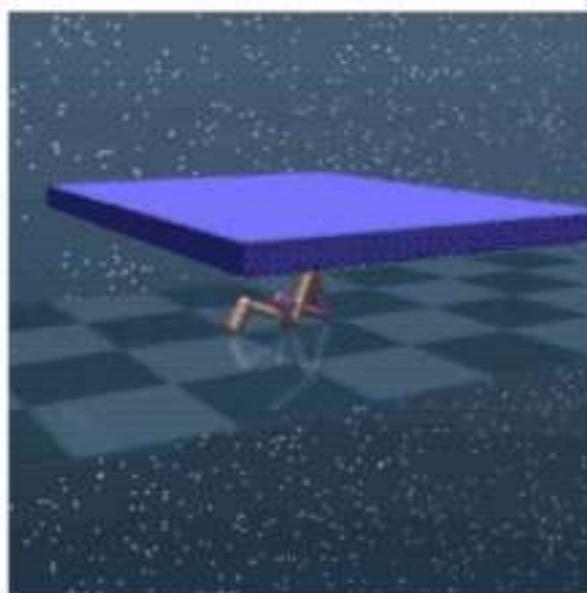


# Qualitative Results

---

Locomotion

Crawl



*Transition*

Walk

Walk Forward

*Transition*

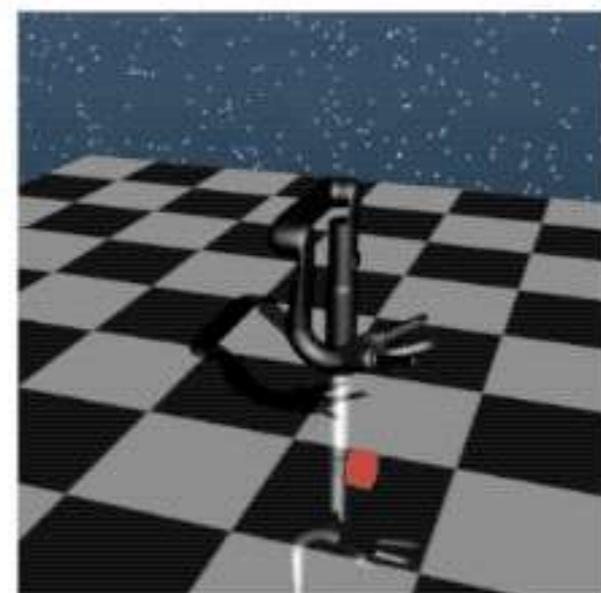
Walk Backward

Manipulation

Pick

*Transition*

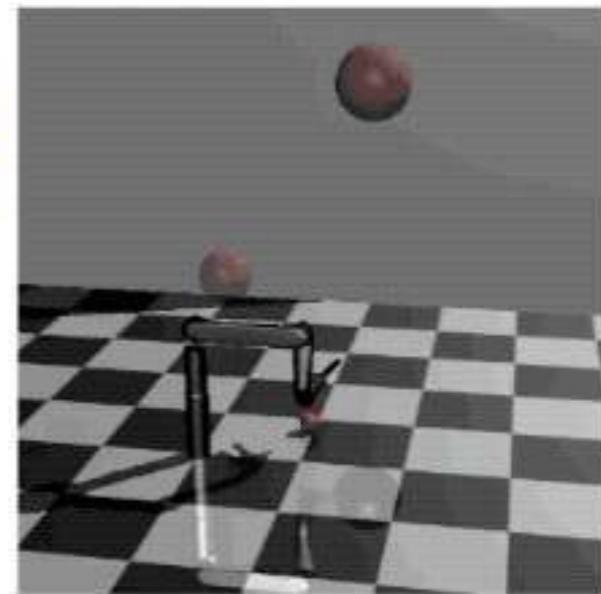
Pick



Toss

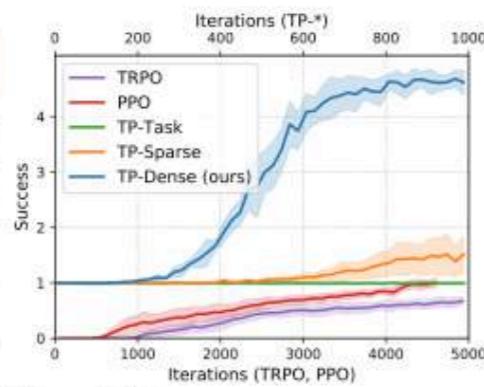
*Transition*

Hit

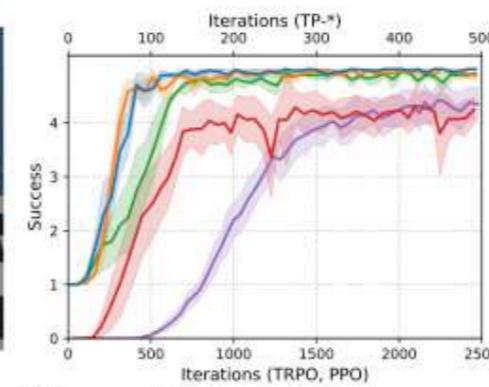
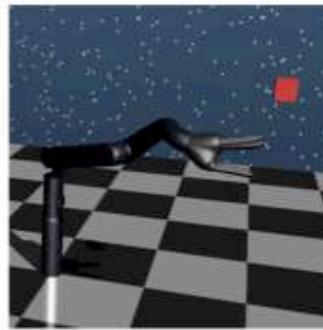


# Quantitative Results

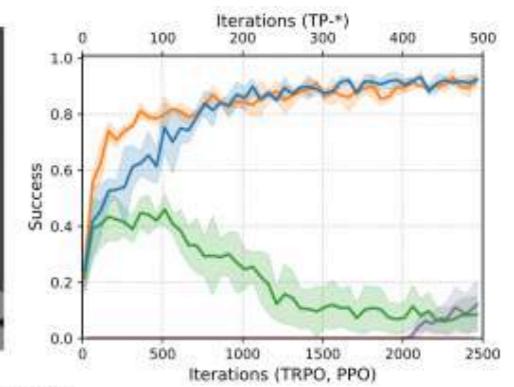
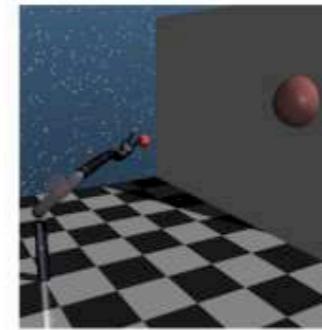
## Sample Efficiency



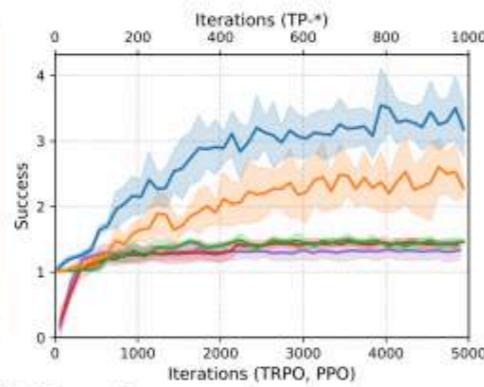
(a) Repetitive picking up



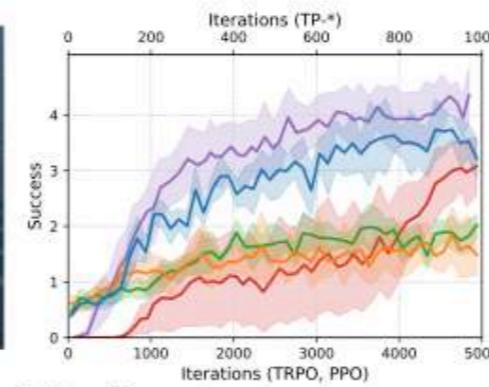
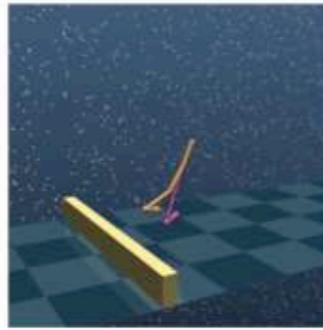
(b) Repetitive catching



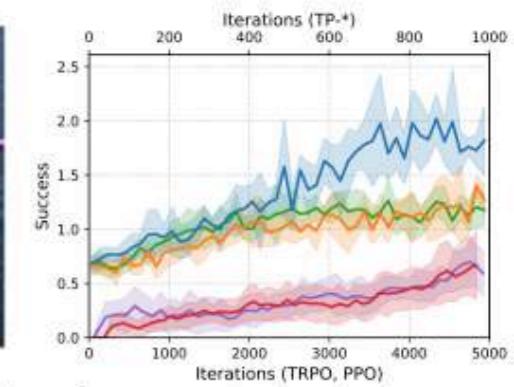
(c) Serve



(d) Patrol



(e) Hurdle

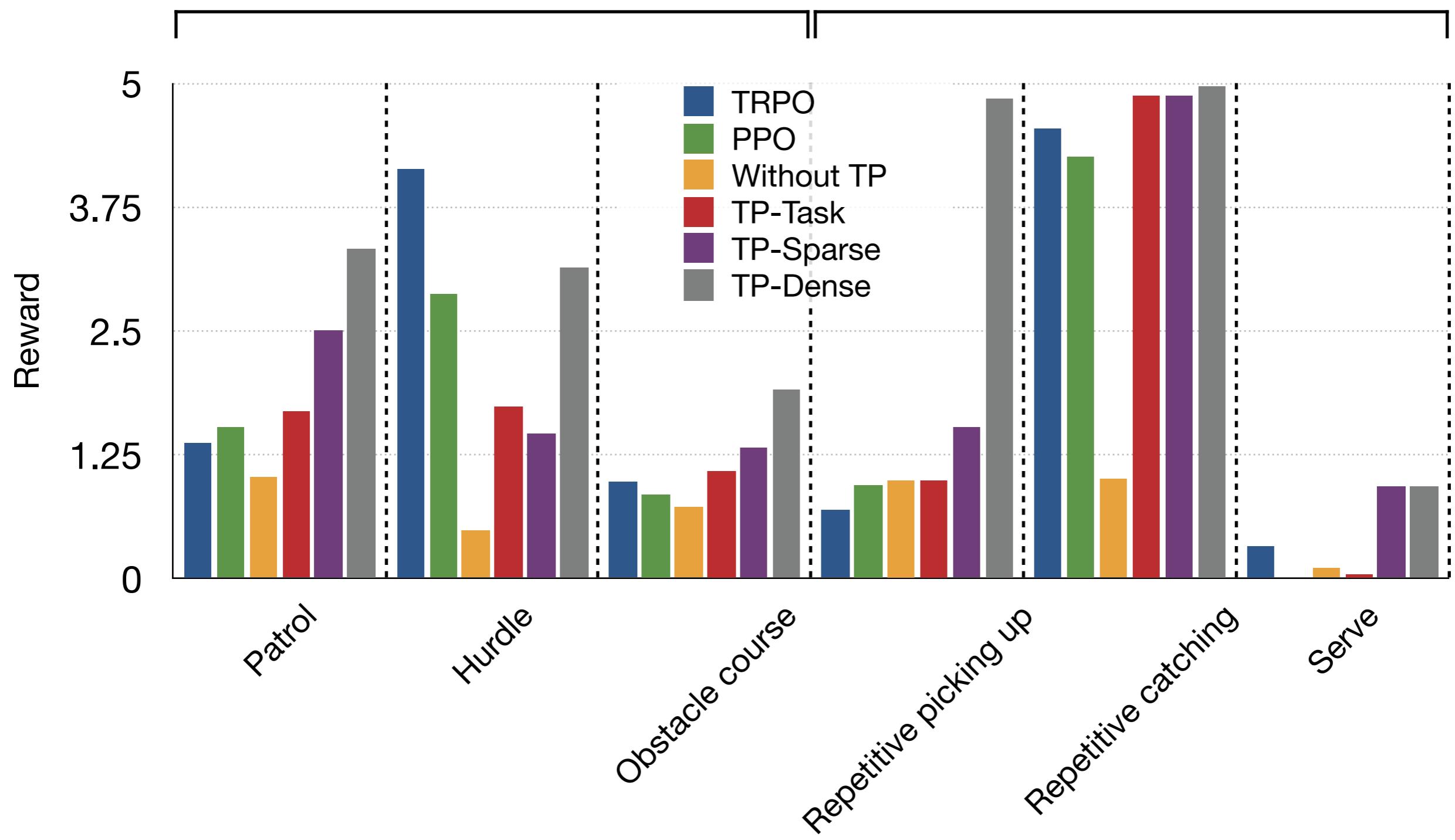


(f) Obstacle course

# Quantitative Results

## Locomotion

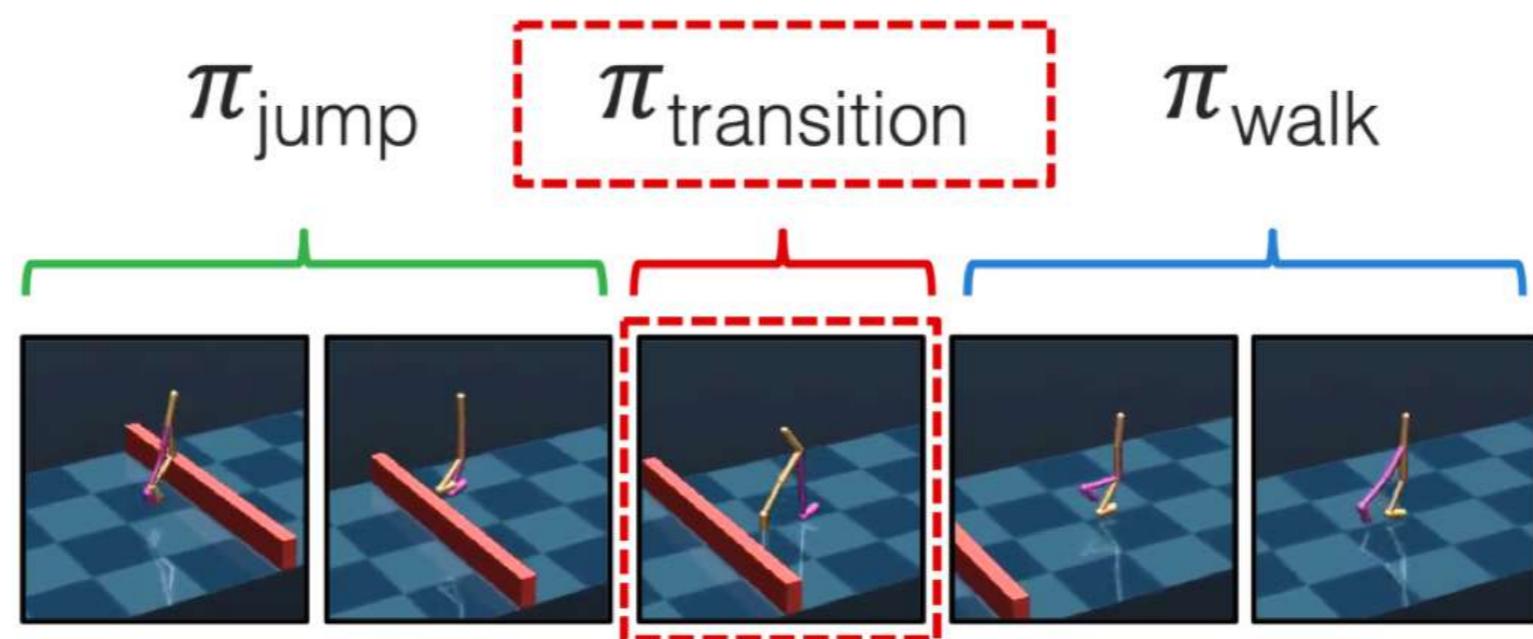
## Manipulation



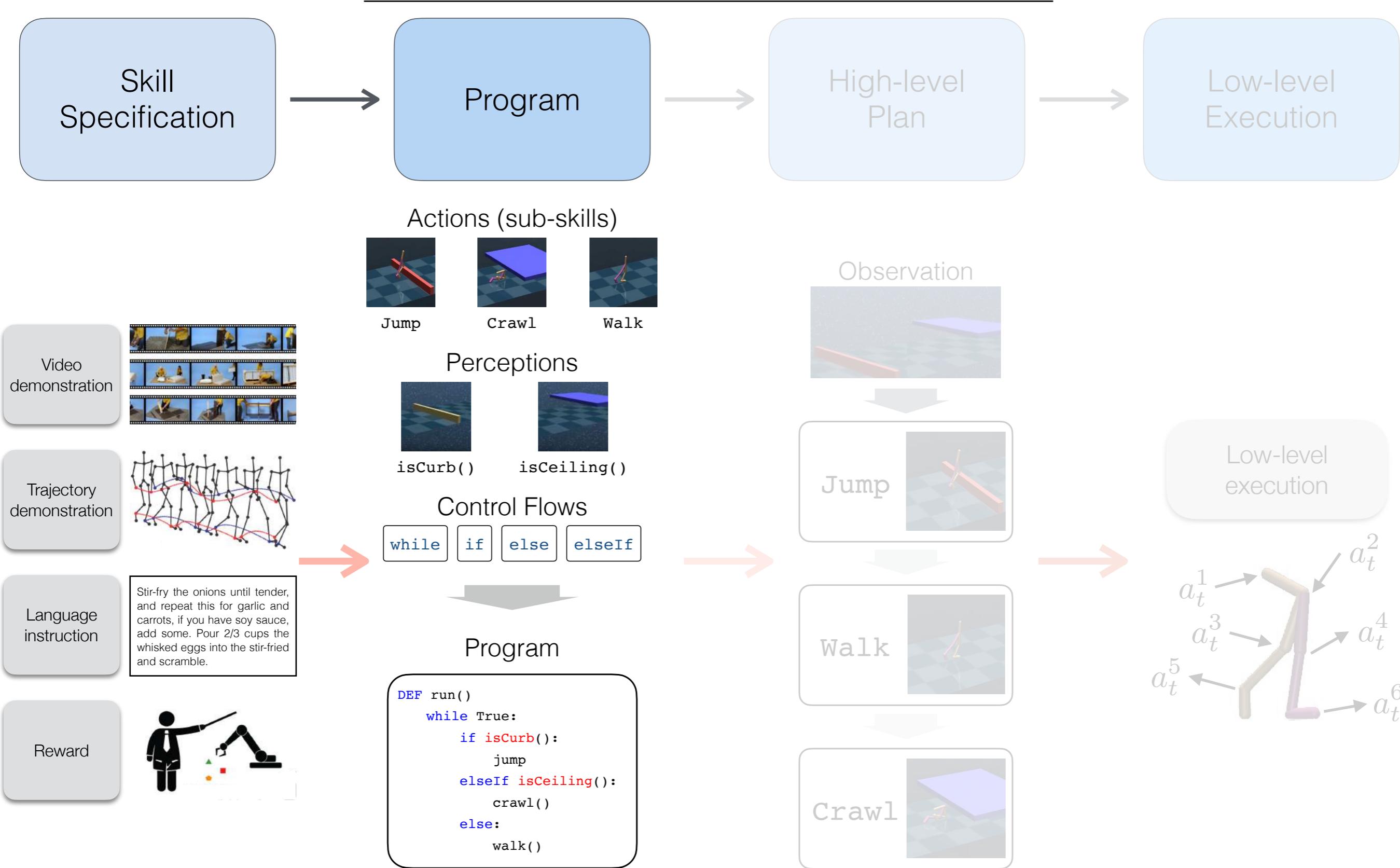
# Takeaway

---

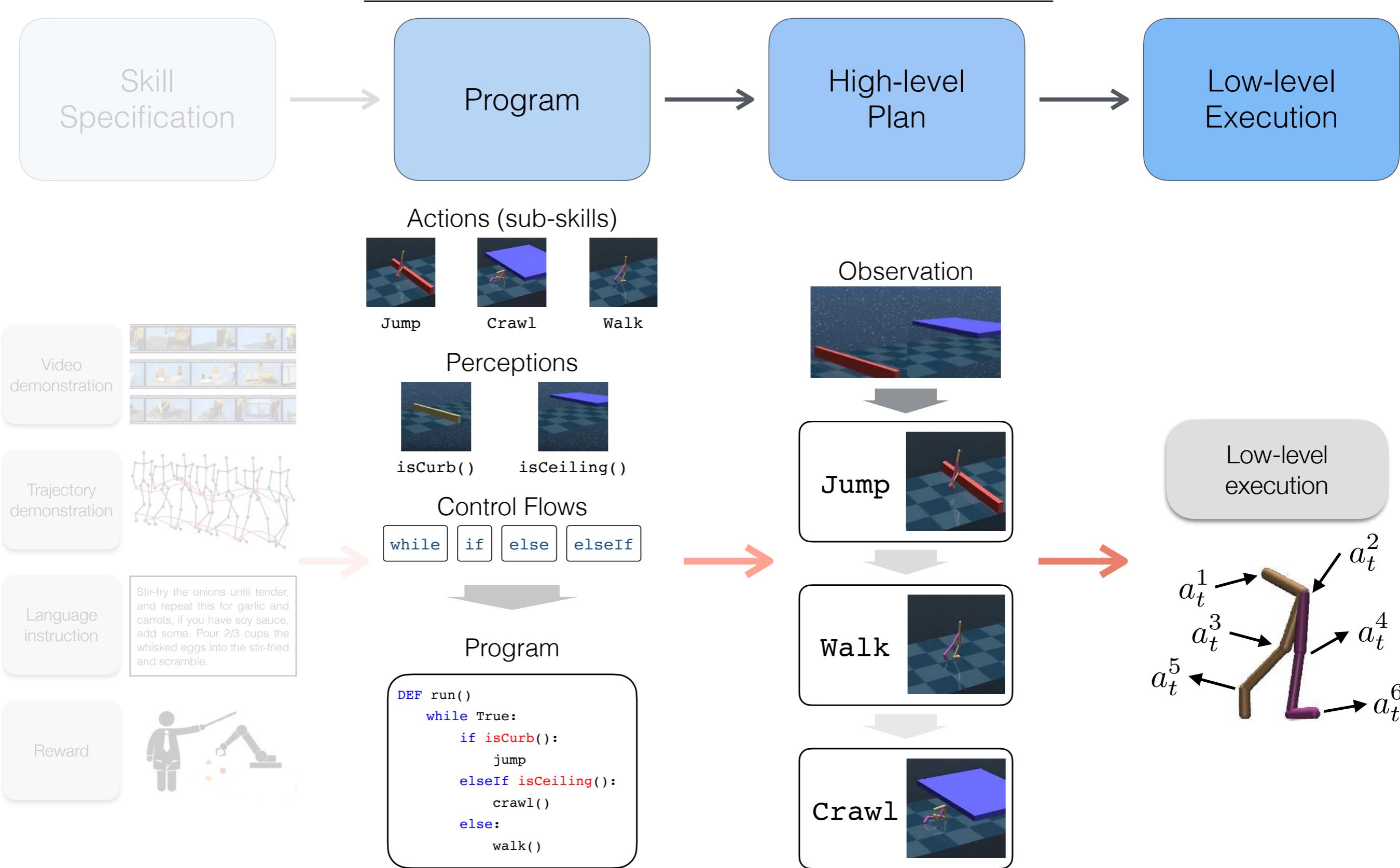
- Learning transition policies to smoothly compose learned skills



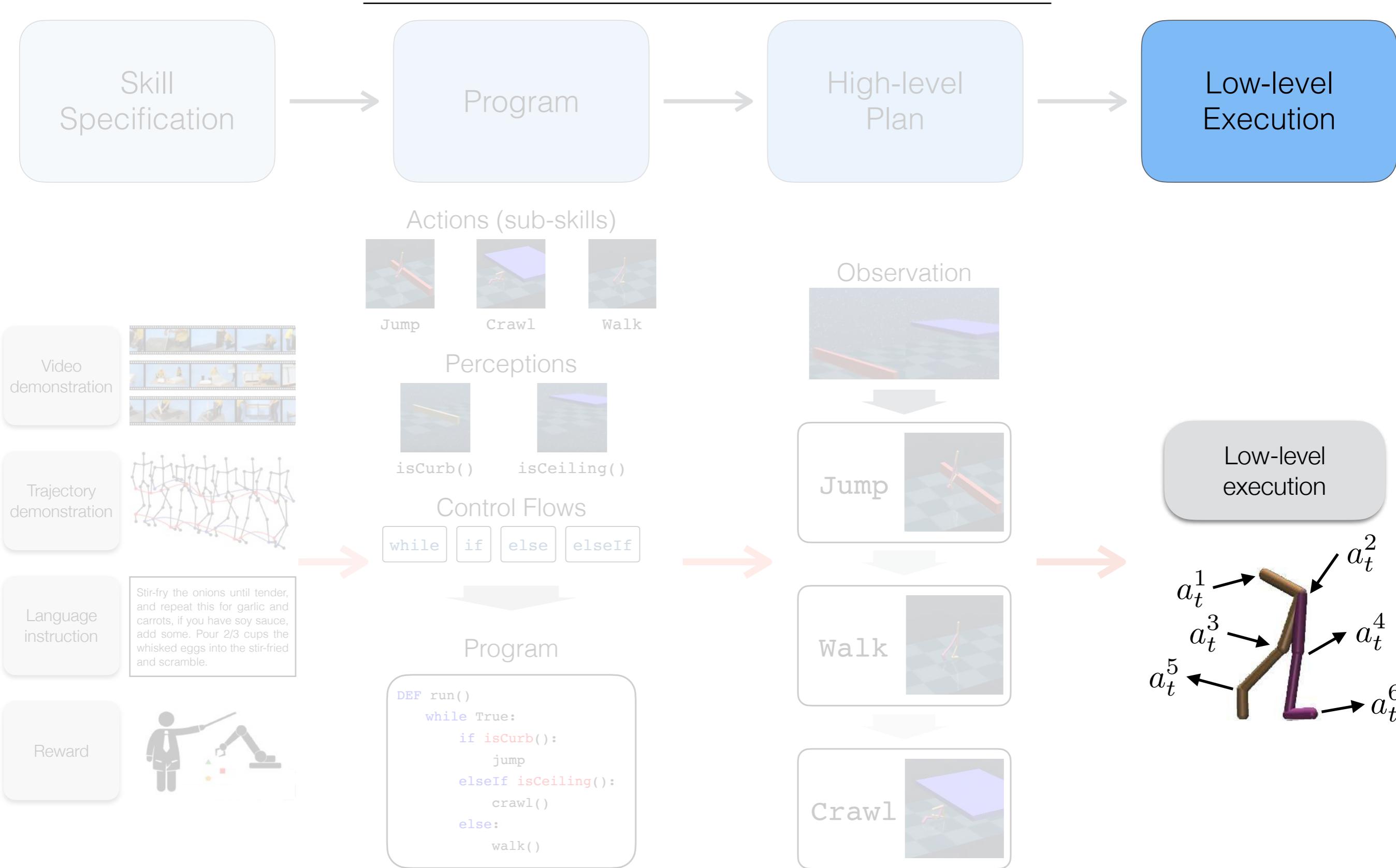
# Program Inference



# Task Execution



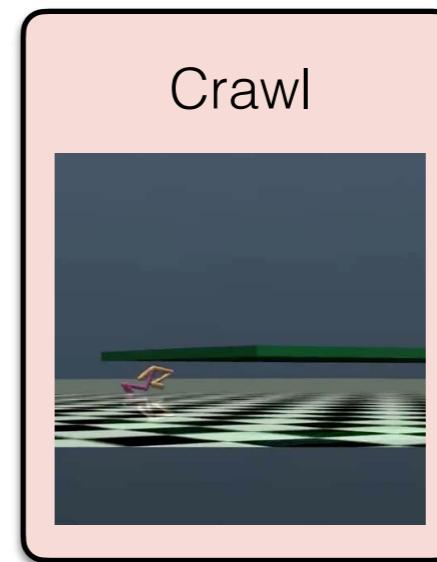
# Primitive Skill Acquisition



# Primitive Skill Acquisition

---

- Goal: acquire a diverse set of primitive skills efficiently



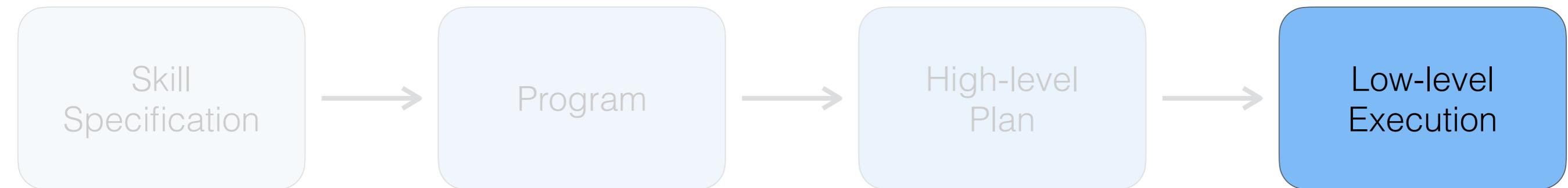
- Key directions

Meta-learning  
Meta-RL

Learning from  
experts

# Toward Multimodal Model-Agnostic Meta-Learning

Meta-learning Workshop at NeurIPS 2018



Risto Vuorio

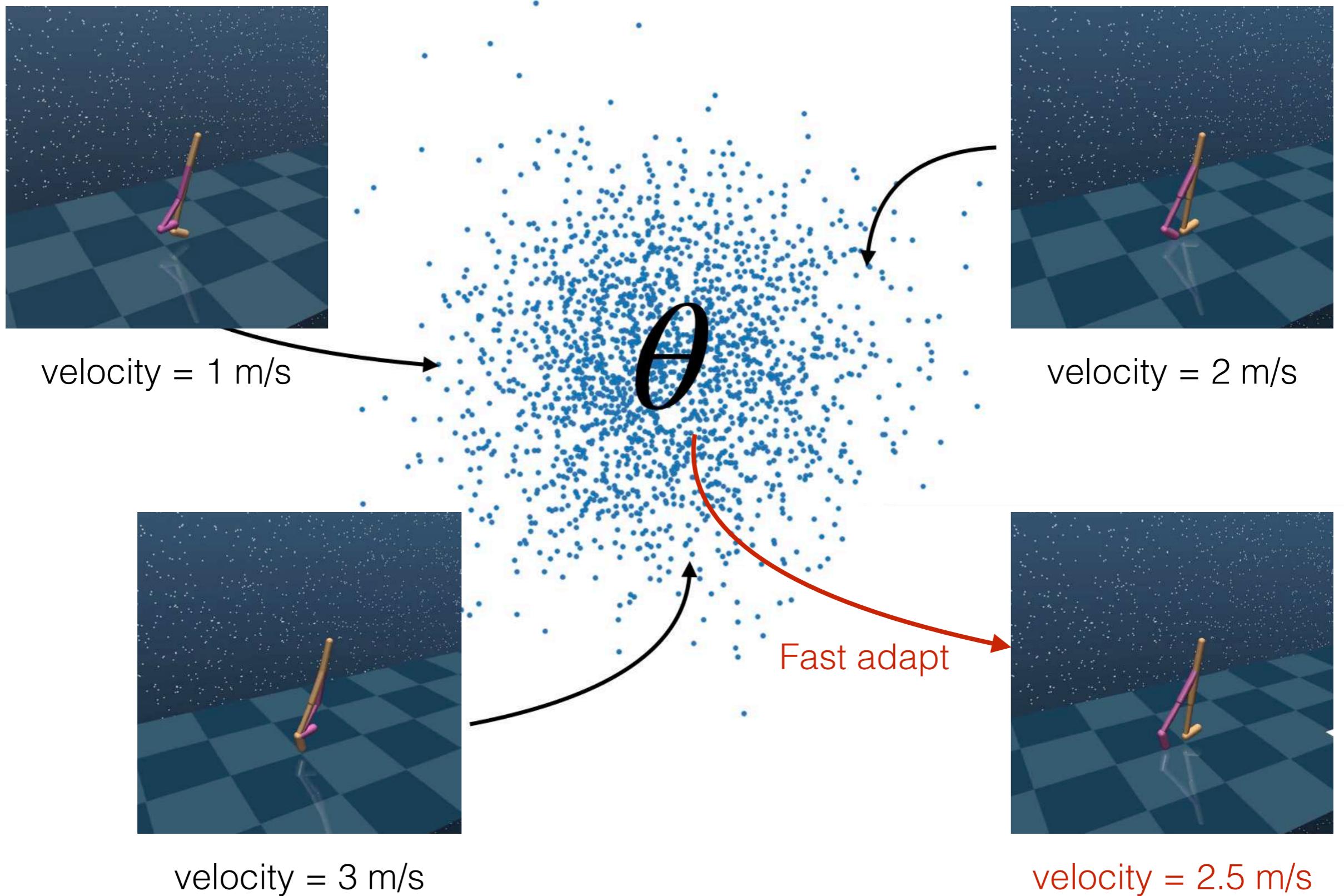


Hexiang Hu

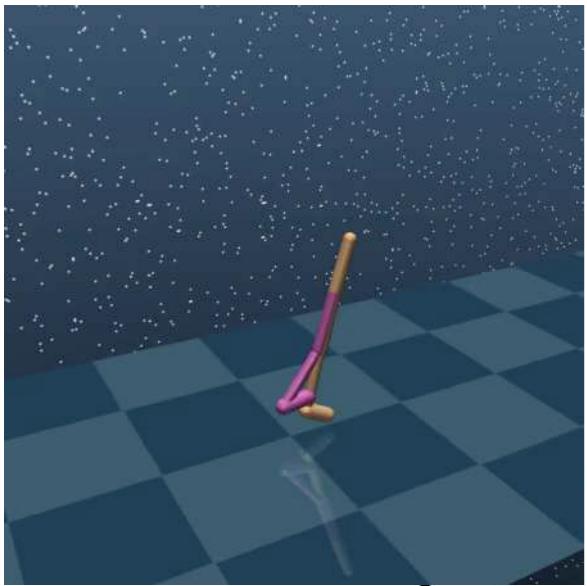


Joseph J. Lim

# Model-Agnostic Meta-Learning (MAML)



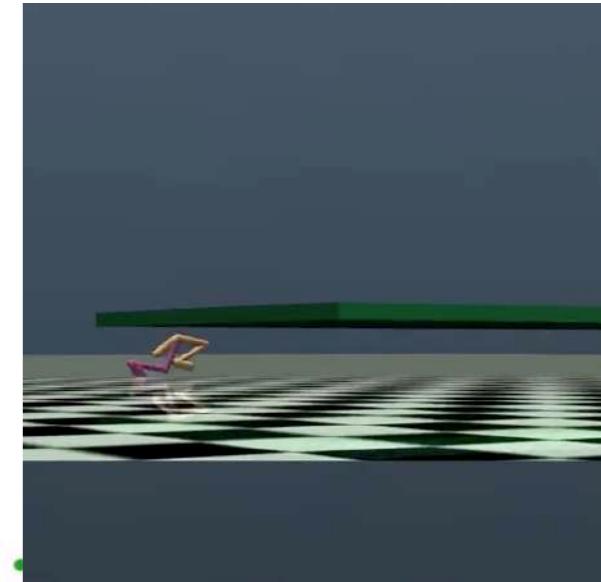
Walk



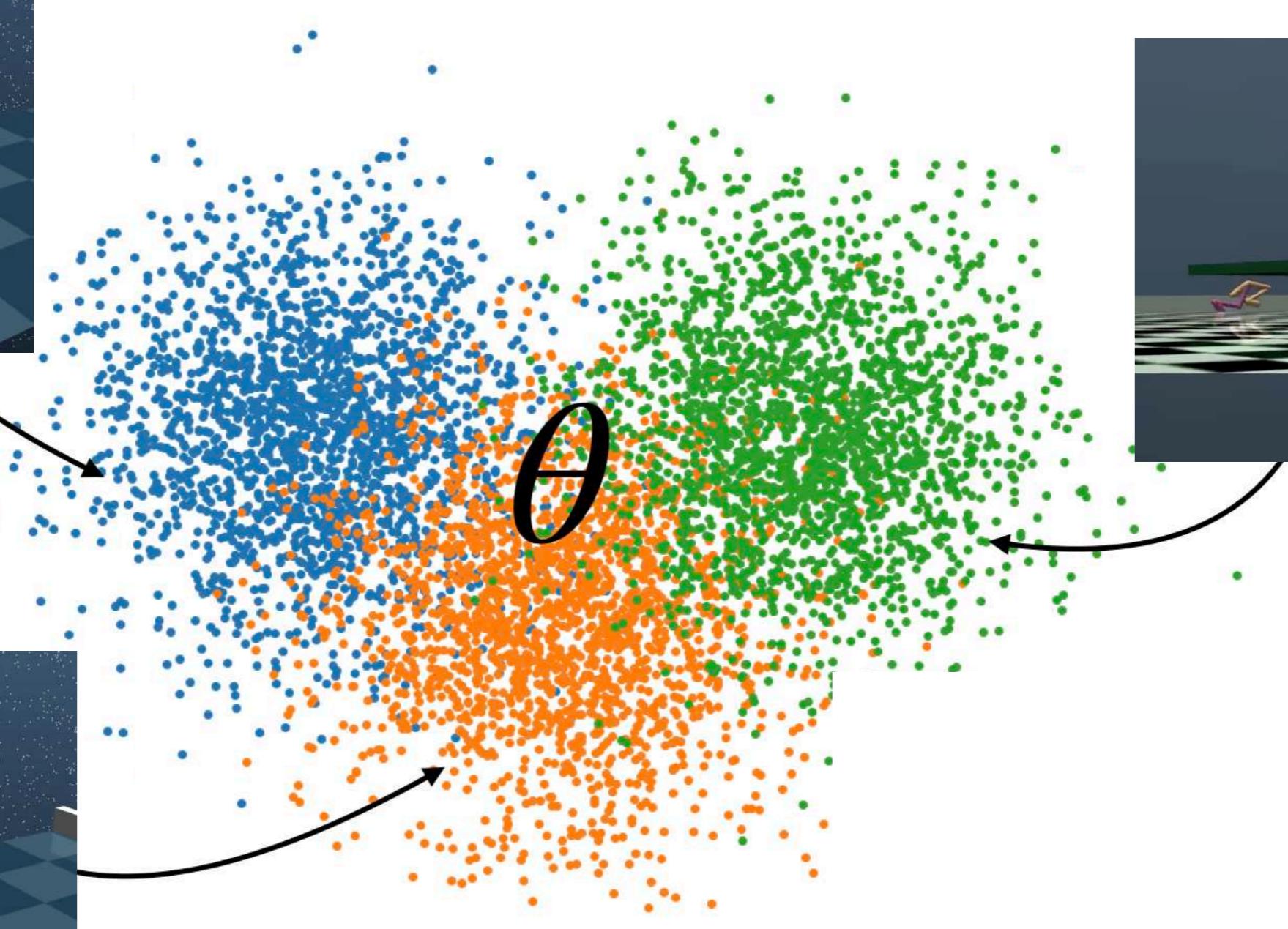
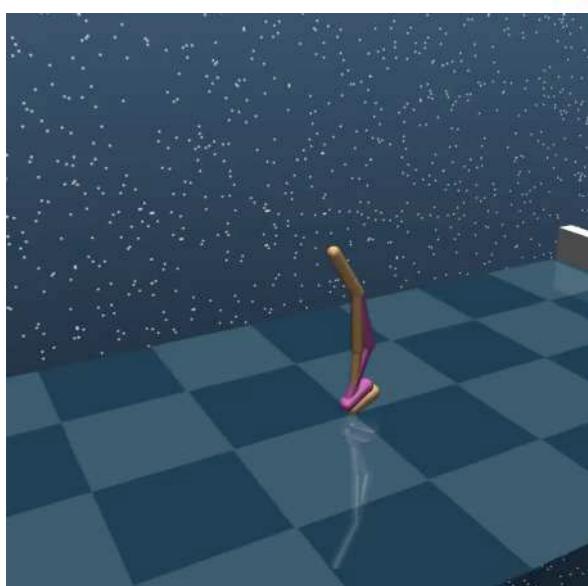
# Multimodal Task Distribution

---

Crawl

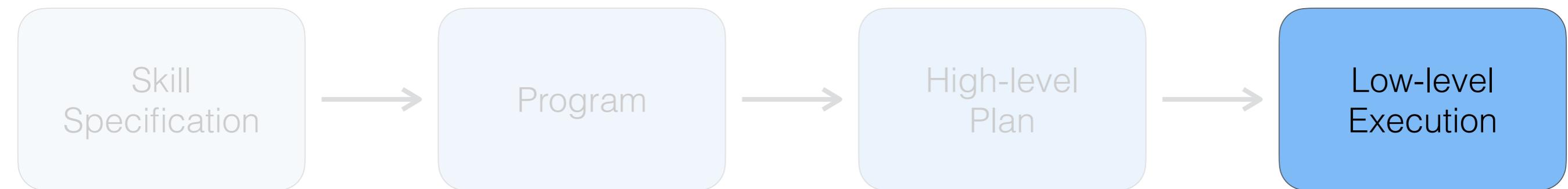


Jump



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

NeurIPS 2019 (Spotlight)



Risto Vuorio

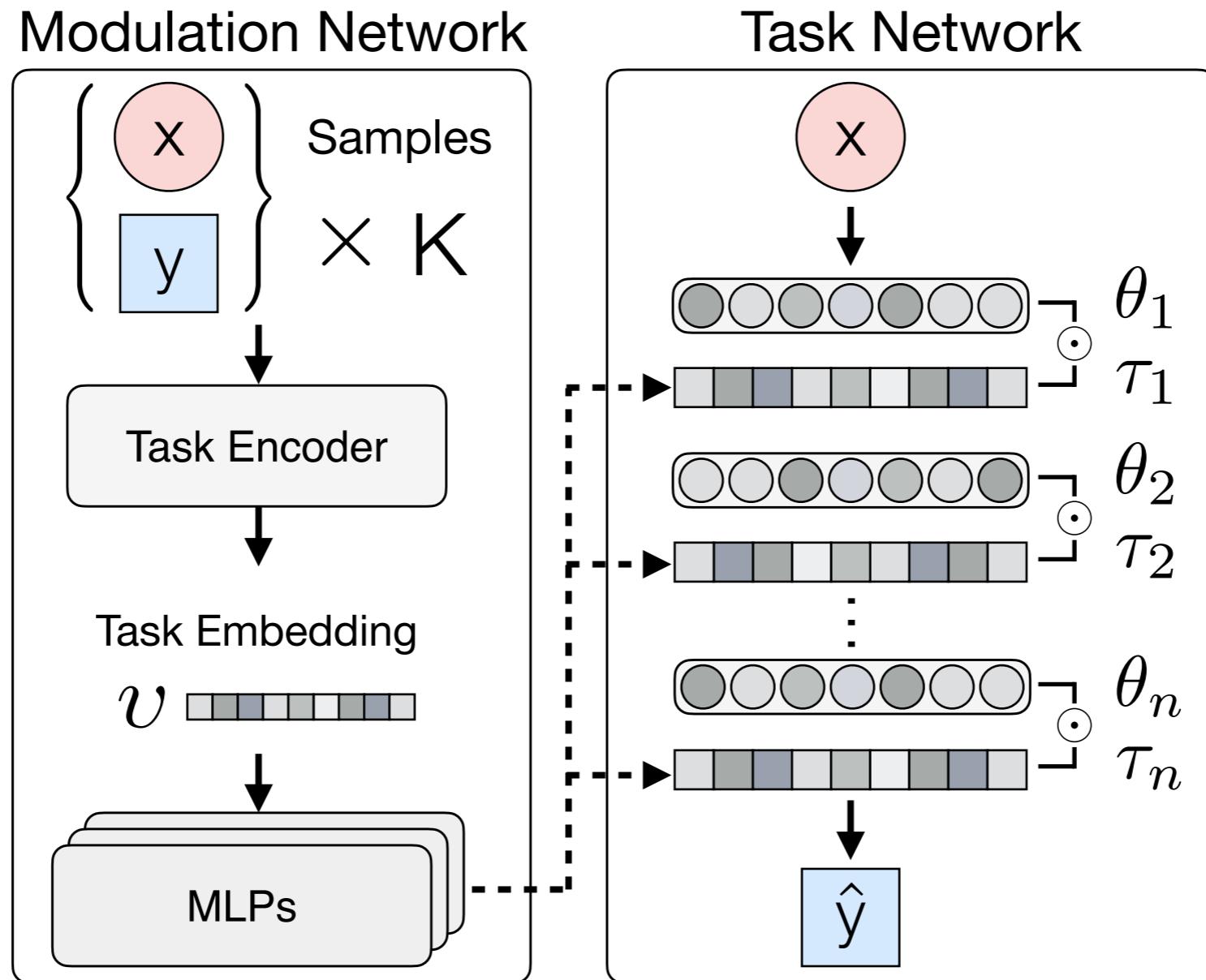


Hexiang Hu



Joseph J. Lim

# Multimodal Model-Agnostic Meta-Learning (MMAML)



# Training Algorithm

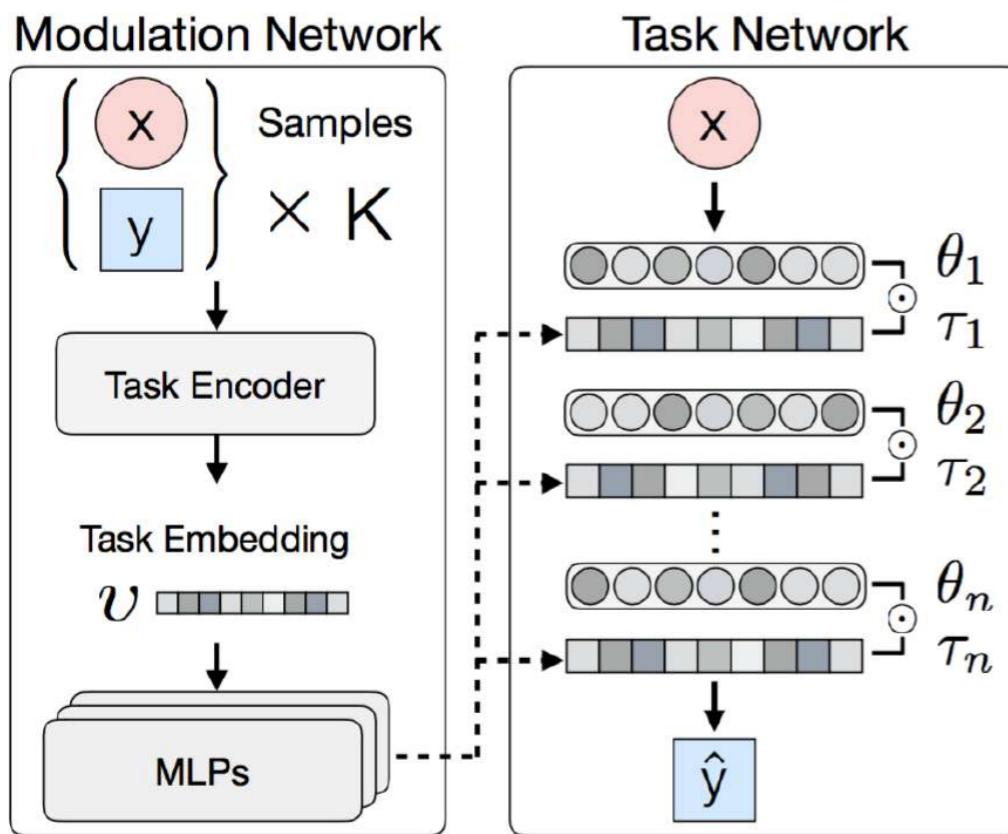
---

## 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

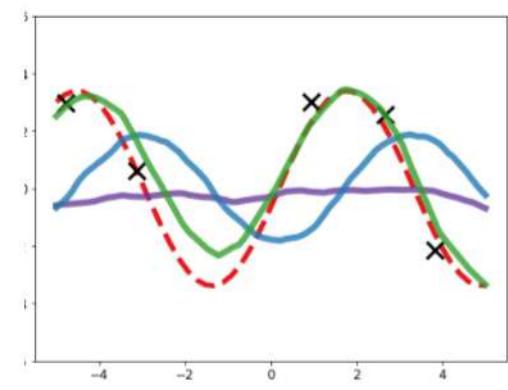
```

---

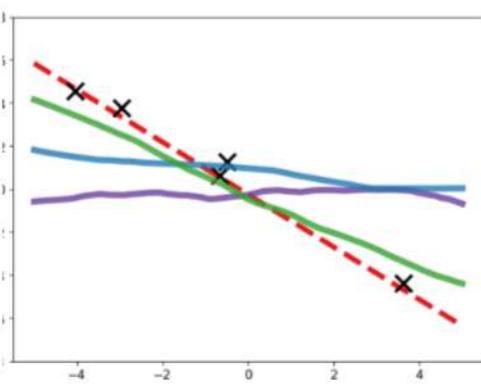
# Regression



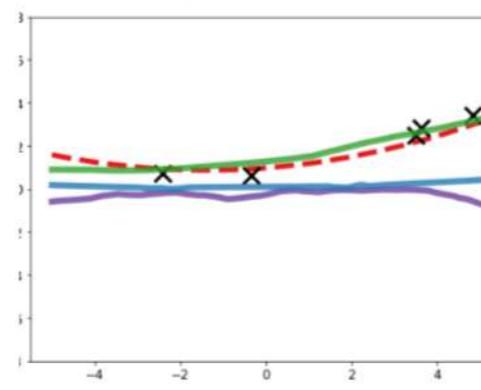
**Sinusoidal**



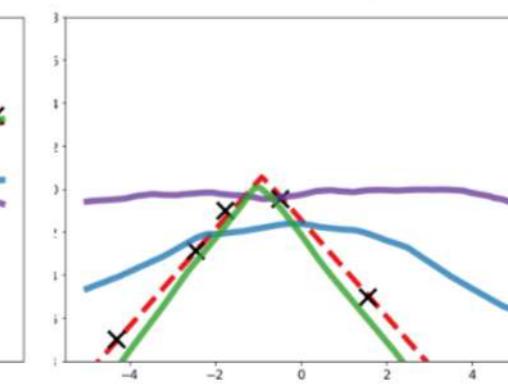
**Linear**



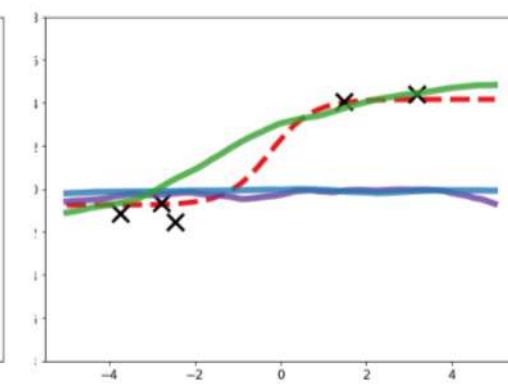
**Quadratic**



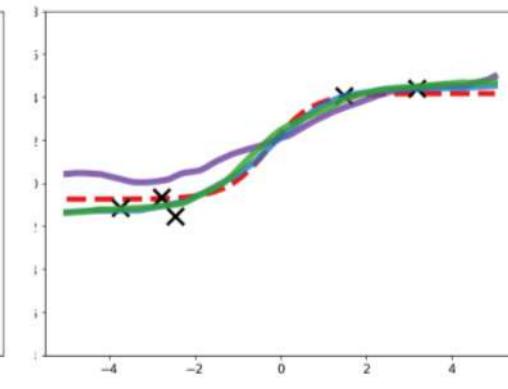
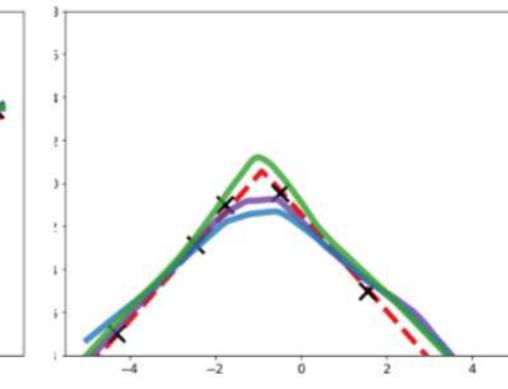
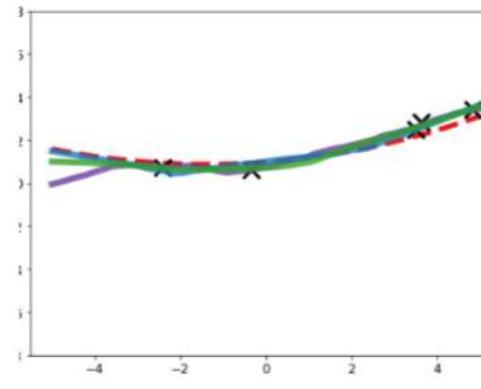
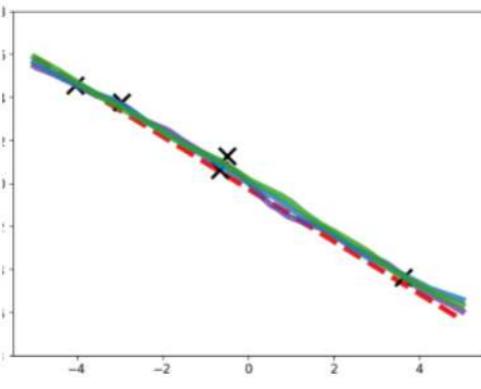
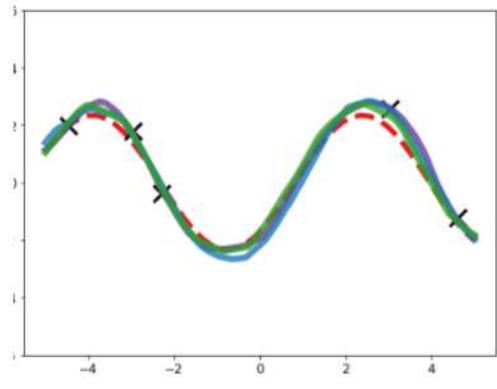
**Transformed  $\ell_1$  Norm**



**Tanh**



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

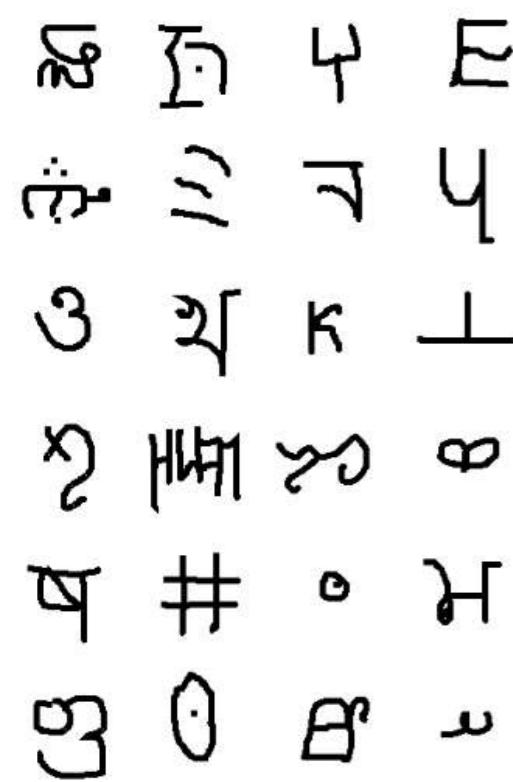


**(b) MMAML post adaptation vs. other posterior models**

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

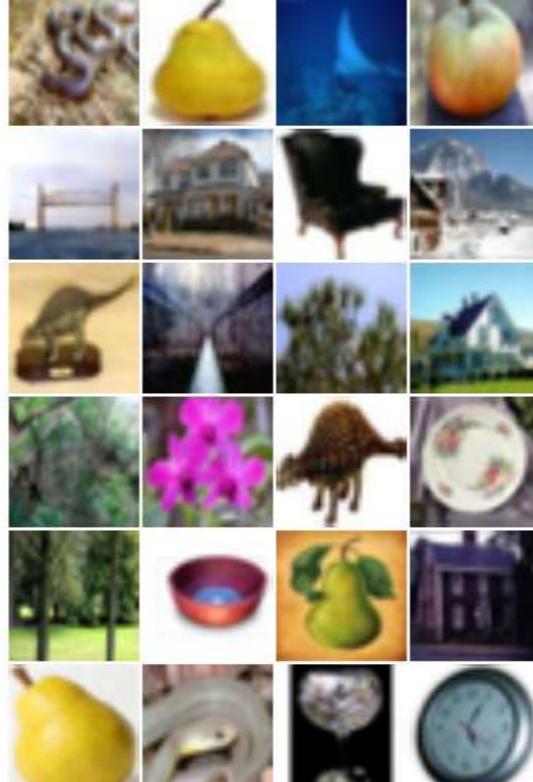
(a) Omniglot



(b) Mini-ImageNet



(c) FC100



(d) CUB



(e) Aircraft



**Method & Setup**

**2 Modes**

**3 Modes**

**5 Modes**

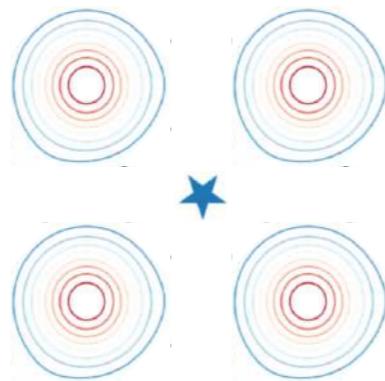
Way Shot	5-way 1-shot		20-way 1-shot		5-way 1-shot		20-way 1-shot		5-way 1-shot		20-way 3-shot	
MAML [1]	66.80%	77.79%	44.69%	54.55%	67.97%	28.22%	66.00%	70.87%	44.69%	56.57%		
Multi-MAML	66.85%	73.07%	<b>53.15%</b>	55.90%	62.20%	<b>39.77%</b>	71.94%	76.94%	58.66%	61.11%		
MMAML (ours)	<b>69.93%</b>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

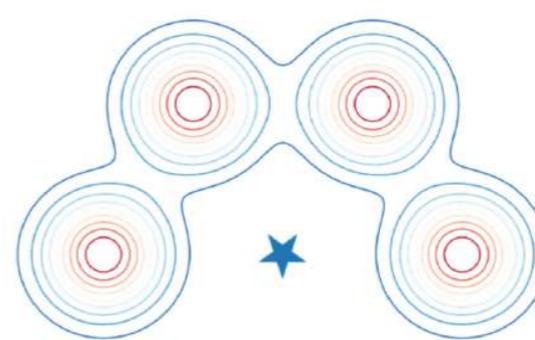
---

Goal modes

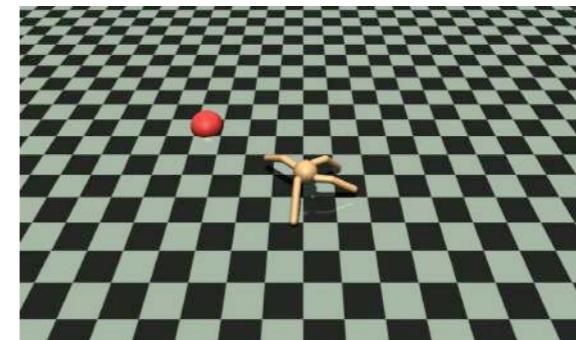
Reacher



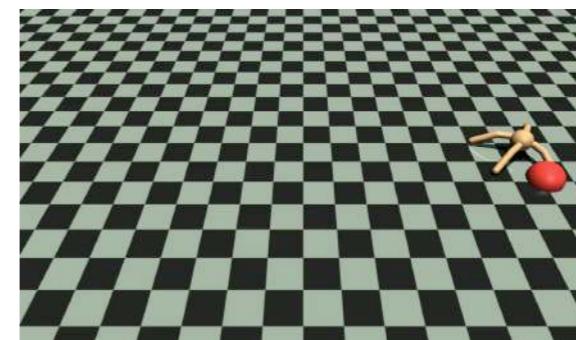
Ant



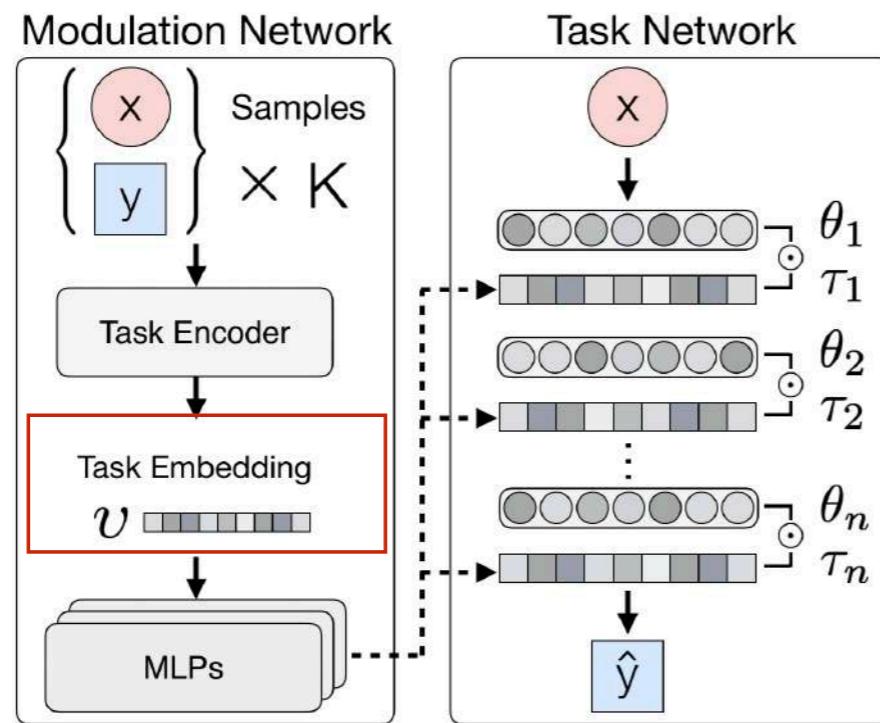
ProMP



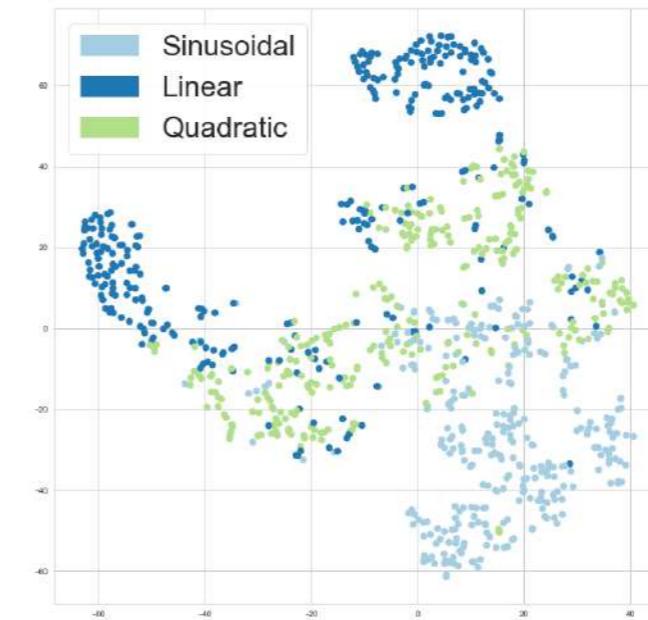
Ours



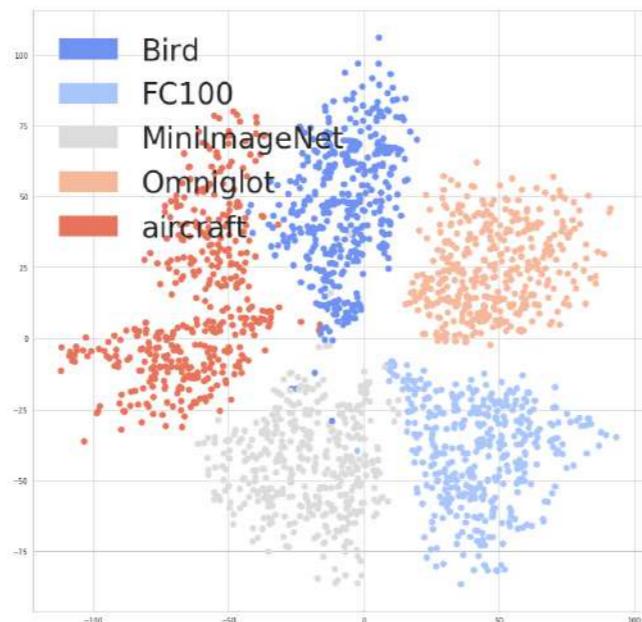
# Learned Task Embedding (tSNE plot)



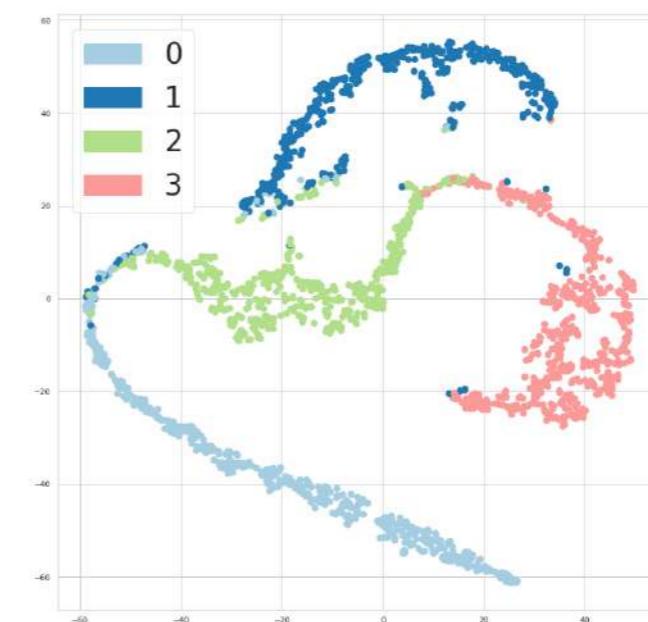
## 3-mode Regression



## 5-mode Classification



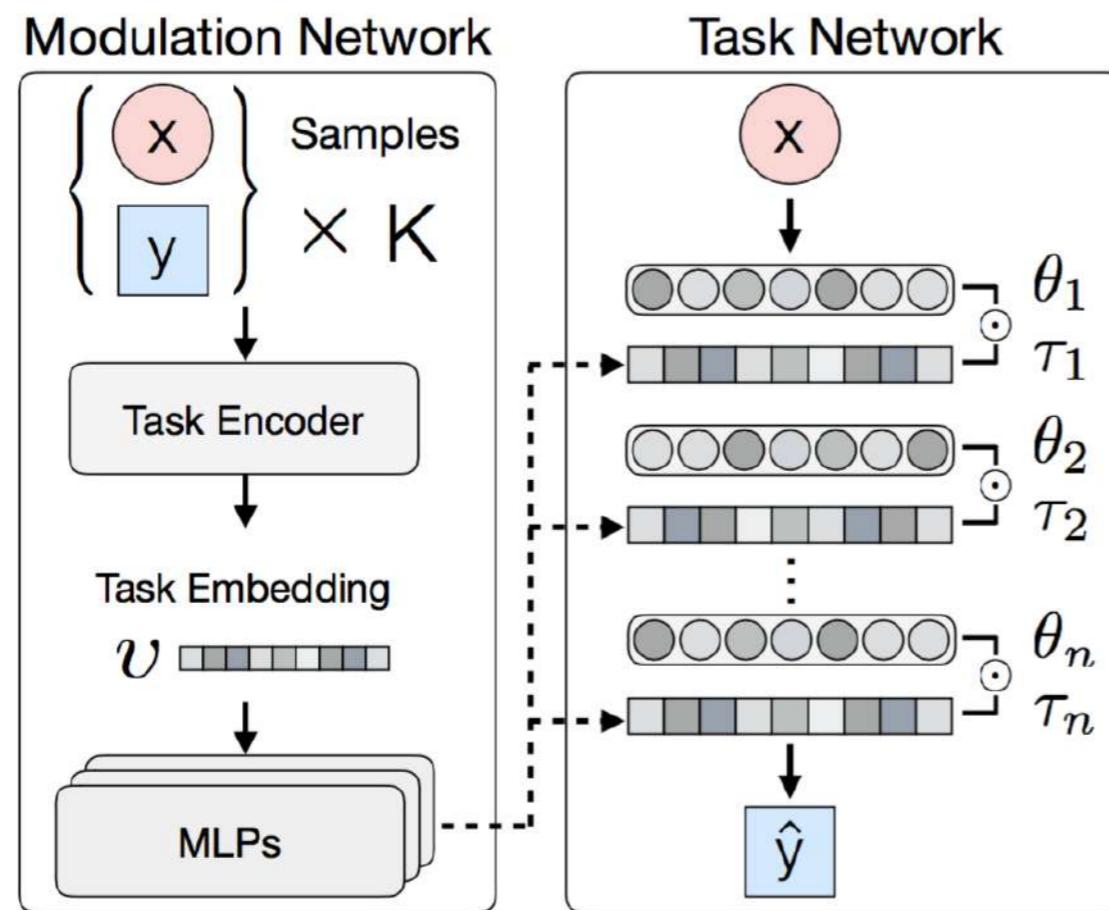
## 4-mode Reacher



# Takeaway

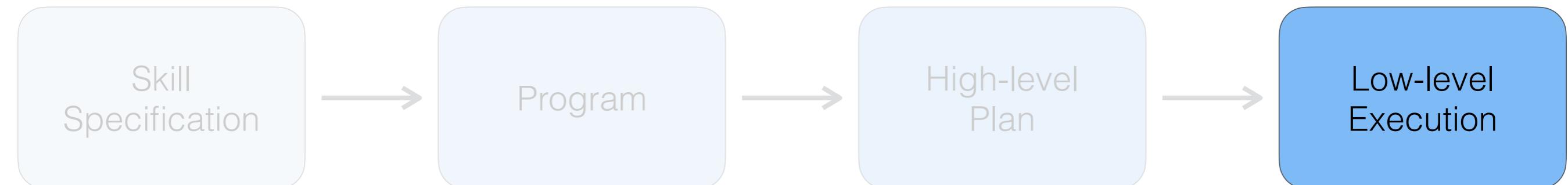
---

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



# Skill-based Meta-Reinforcement Learning

Deep RL workshop @ NeurIPS 2021  
Meta-learning workshop @ NeurIPS 2021  
submitted to ICLR 2022



Taewook Nam



Karl Pertsch



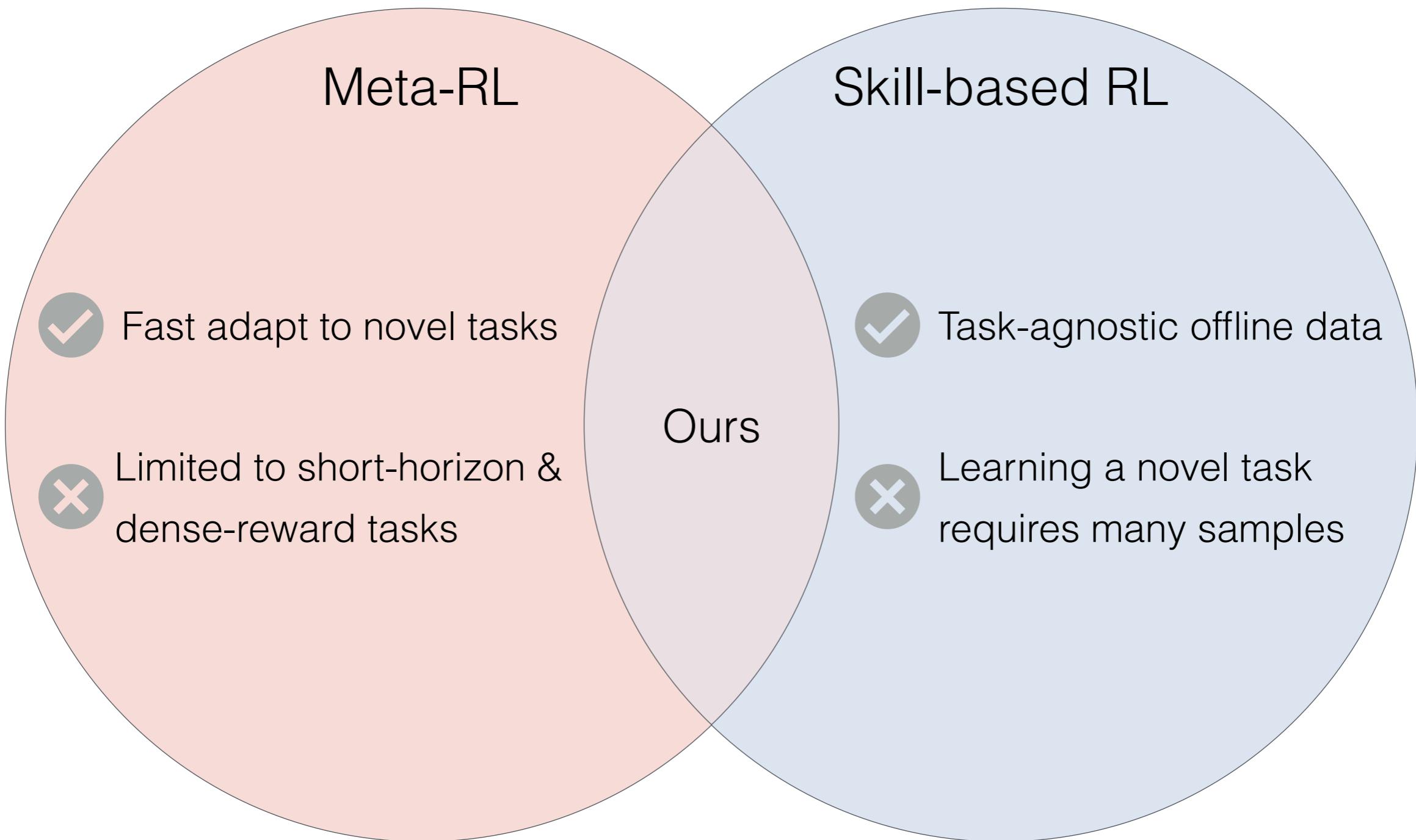
Sung Ju Hwang



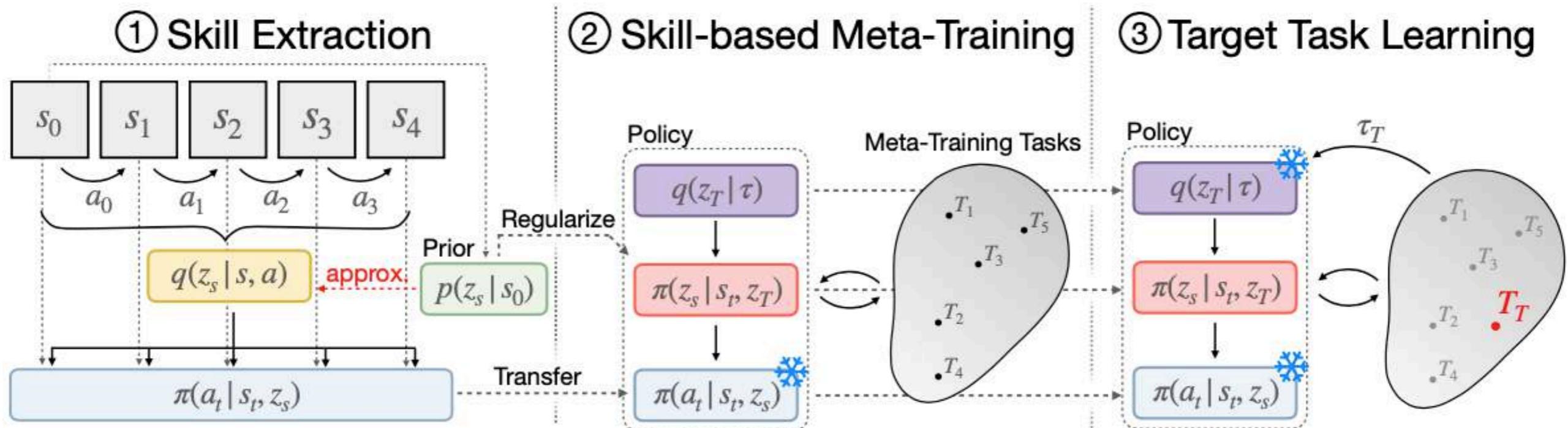
Joseph J. Lim

# Meta-RL with Skills

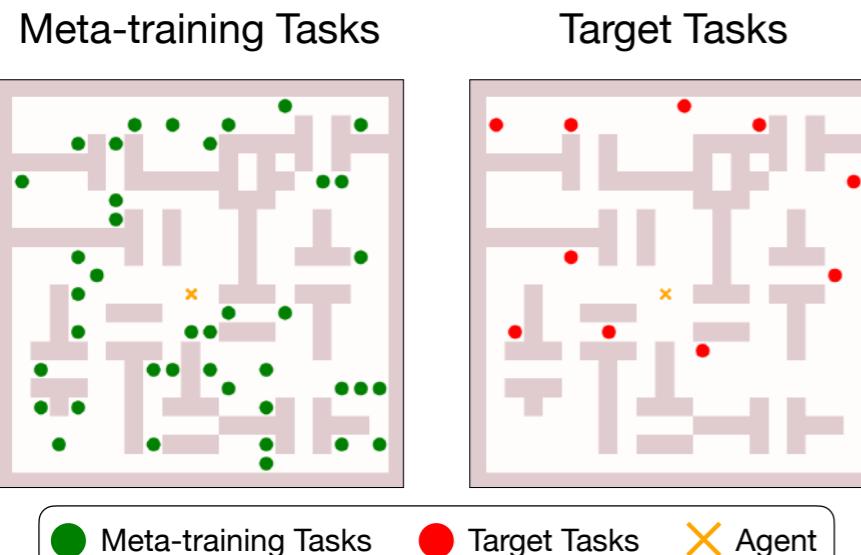
---



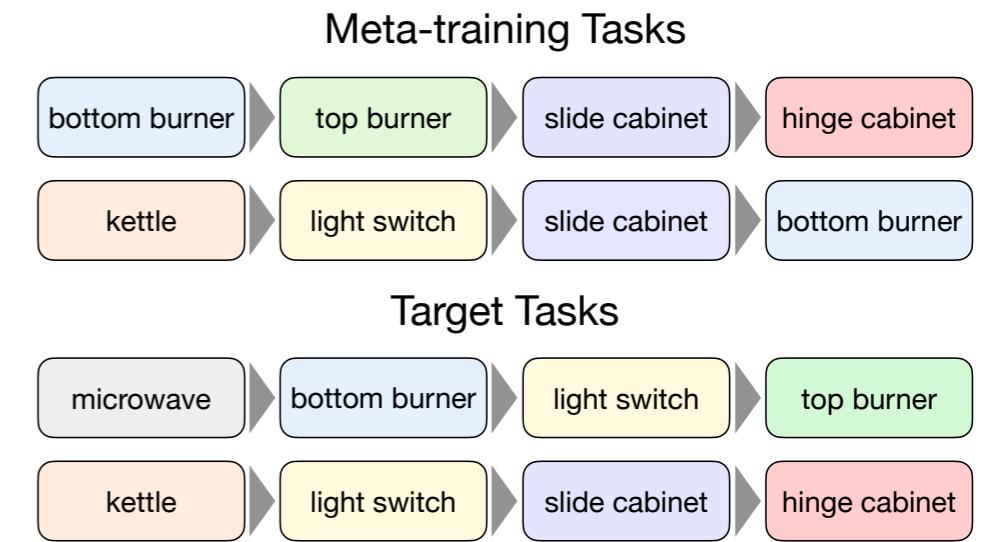
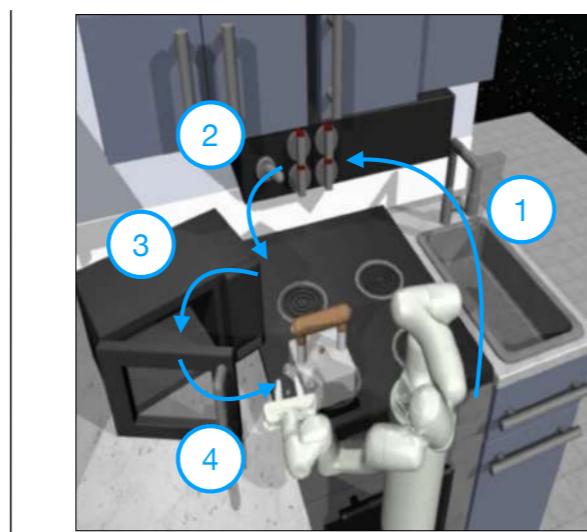
# SiMPL: Skill-based Meta Policy Learning



# Environments

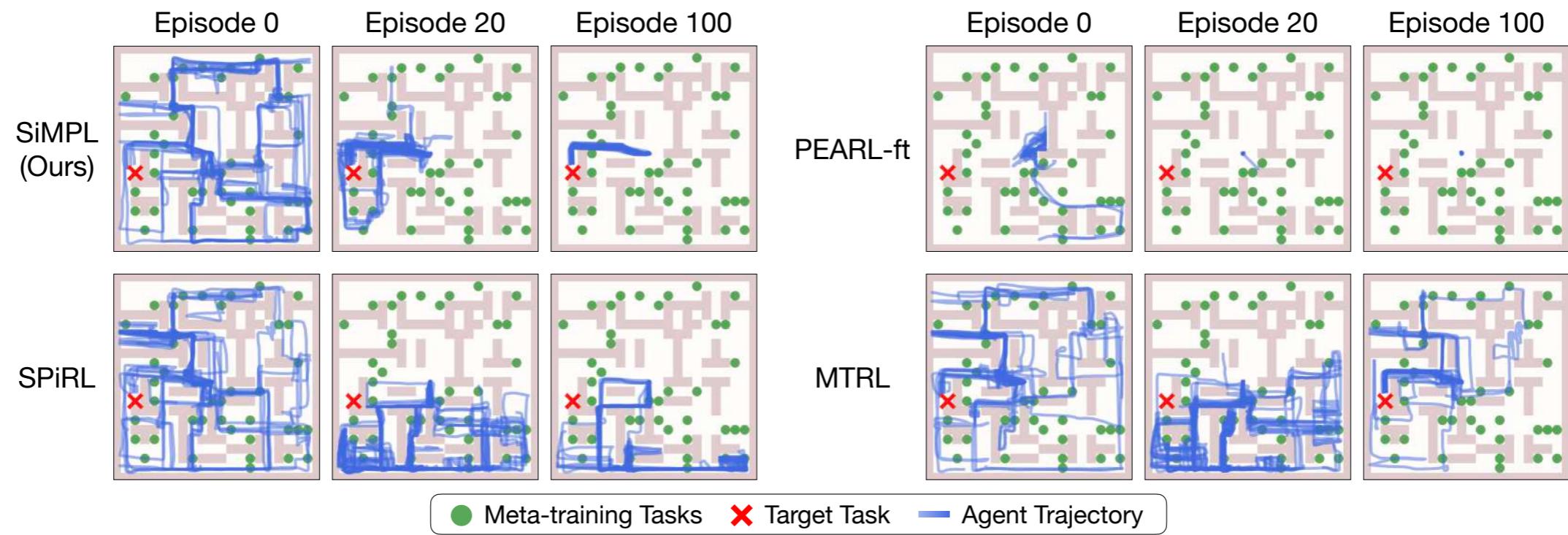
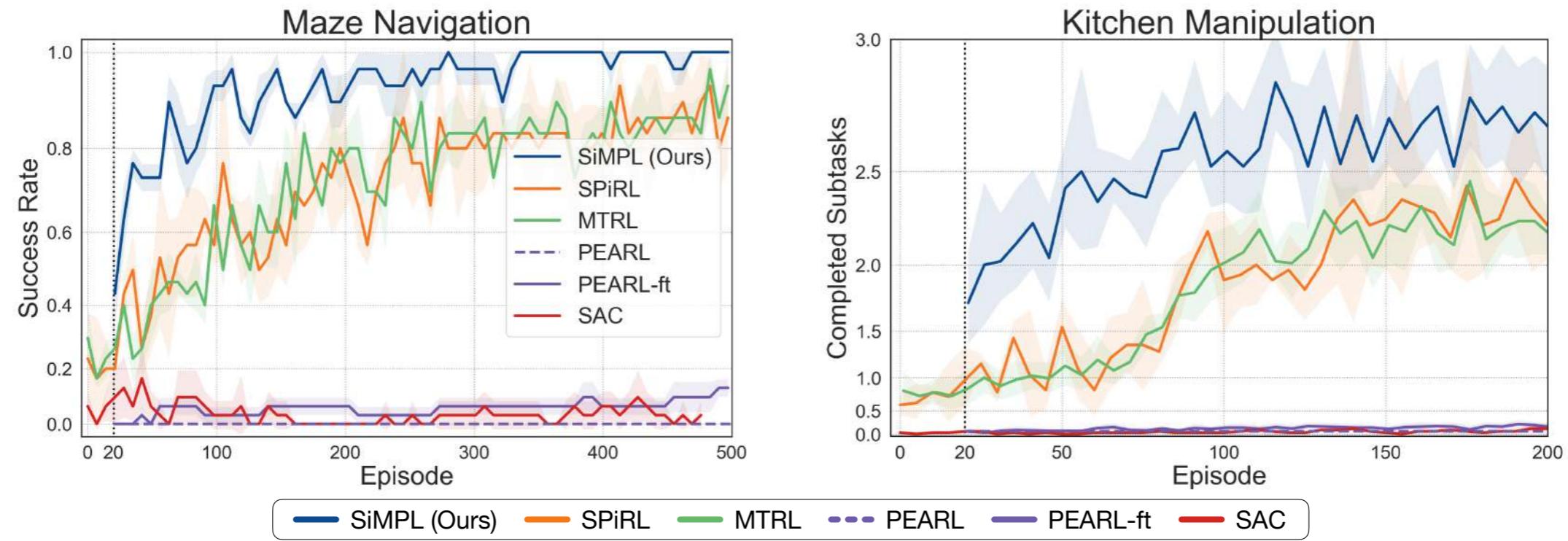


(a) Maze Navigation



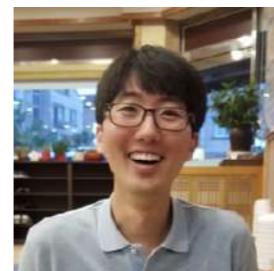
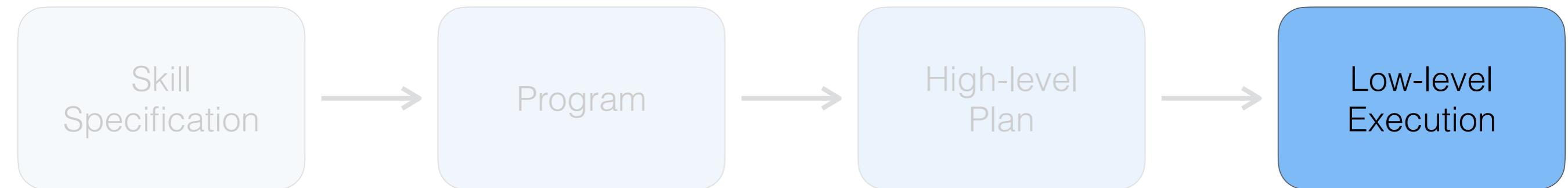
(b) Kitchen Manipulation

# Results



# Generalizable Imitation Learning from Observation via Inferring Goal Proximity

NeurIPS 2021



Youngwoon Lee

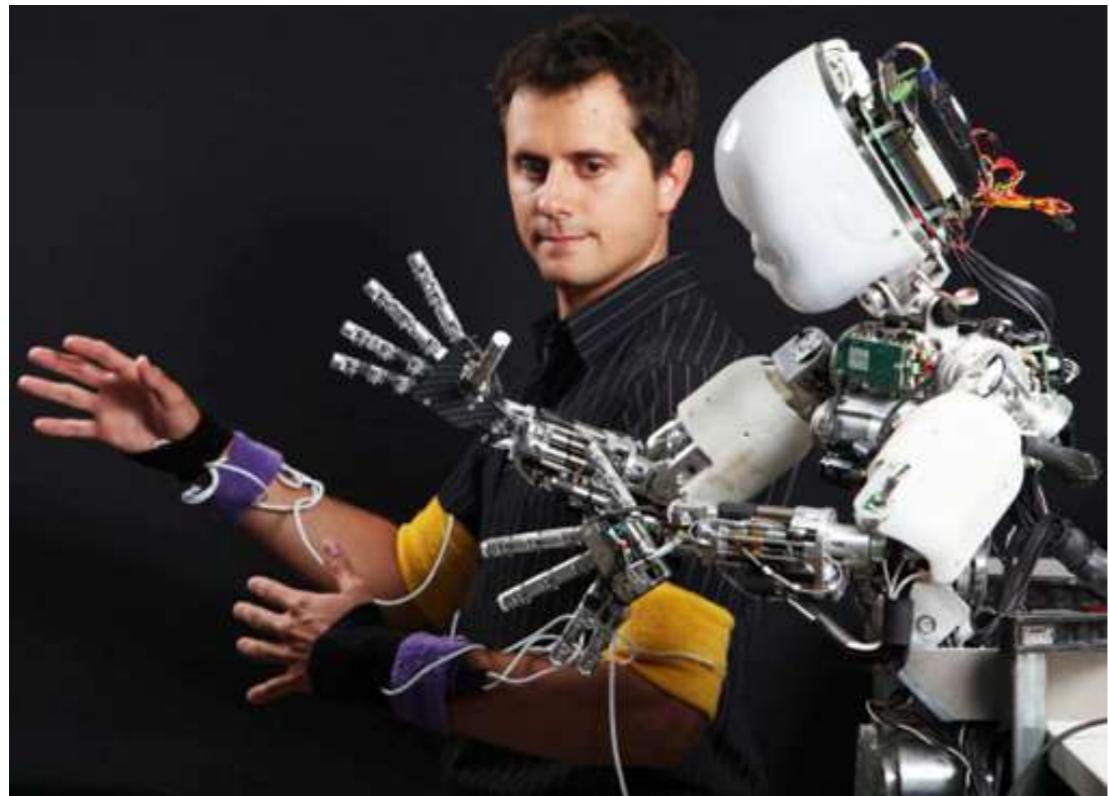


Andrew Szot



Joseph J. Lim

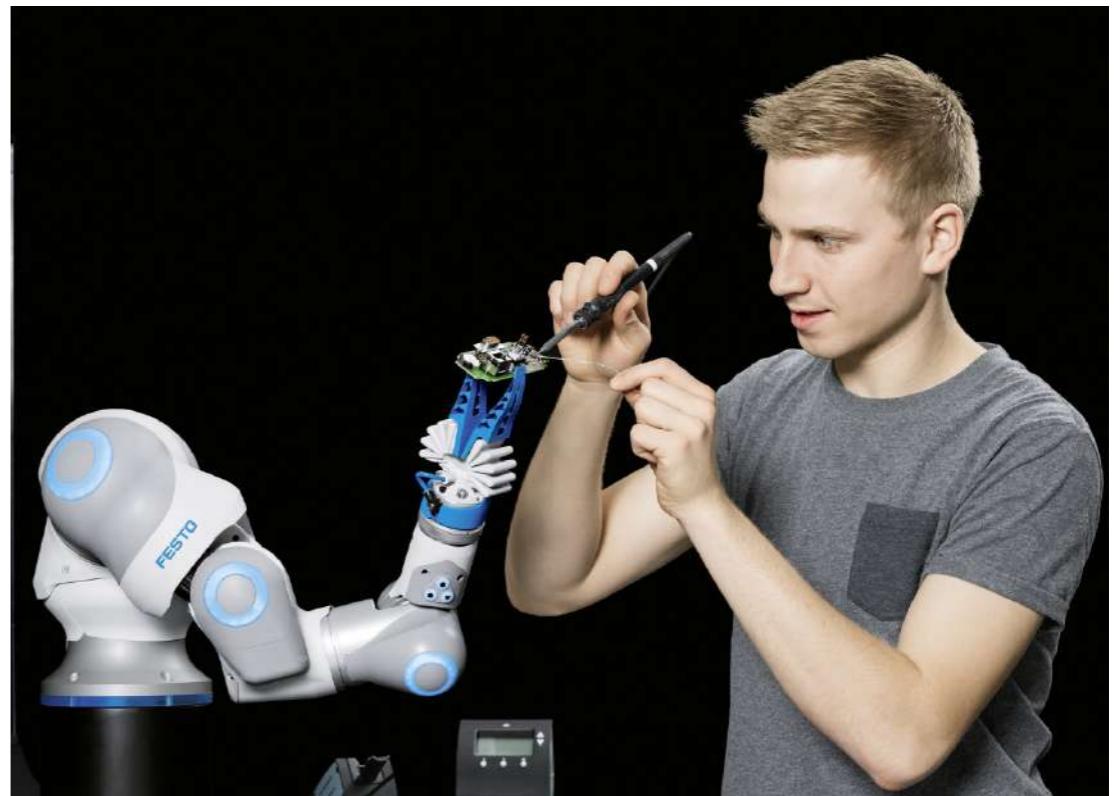
## Learning from Demonstration



with expert's actions

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

## Learning from Observation

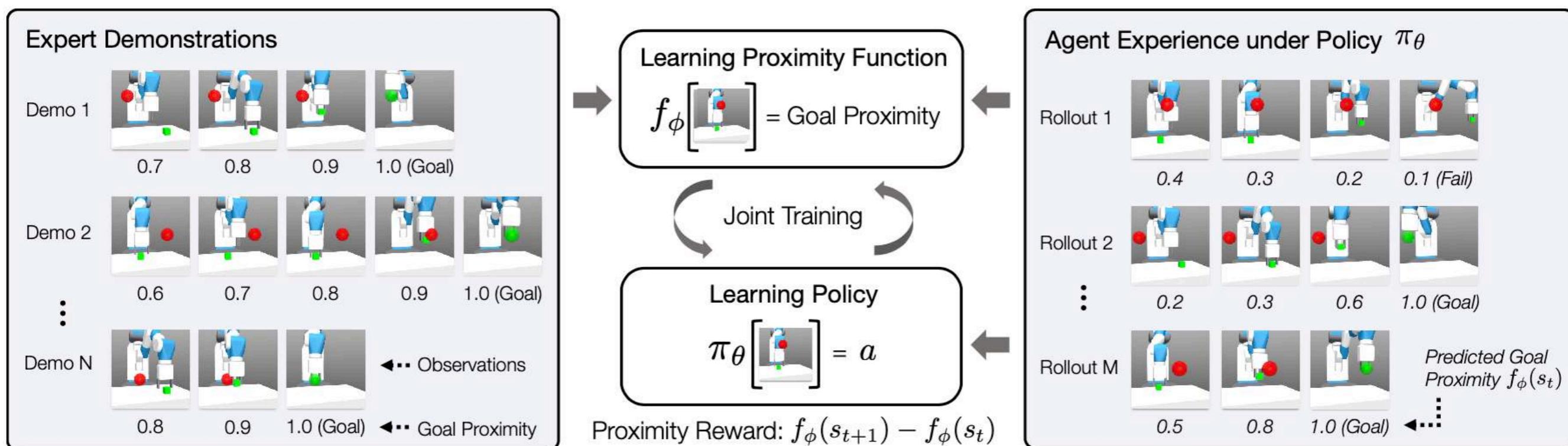


vs.

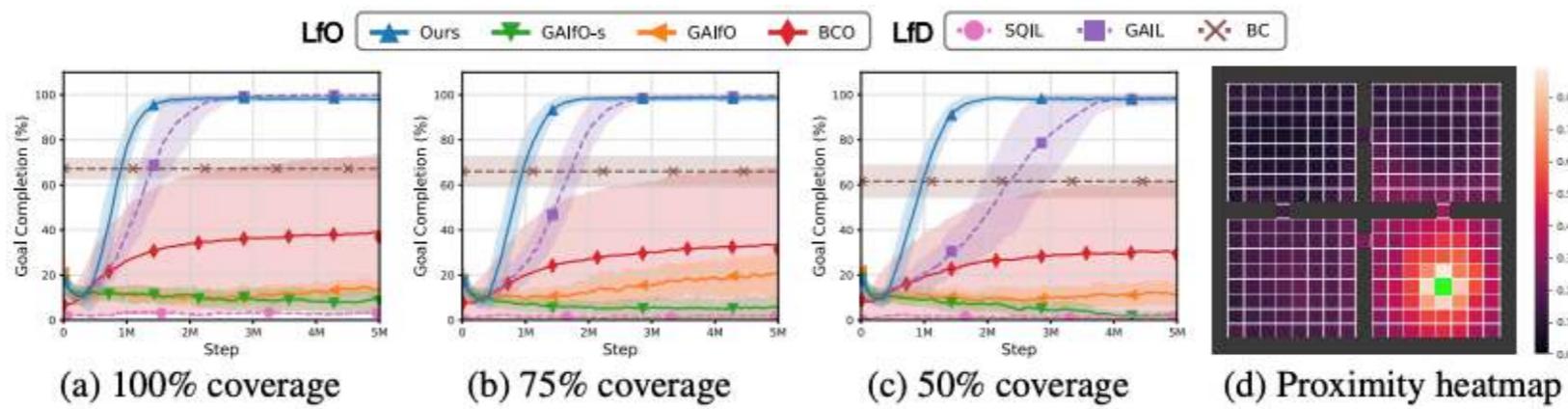
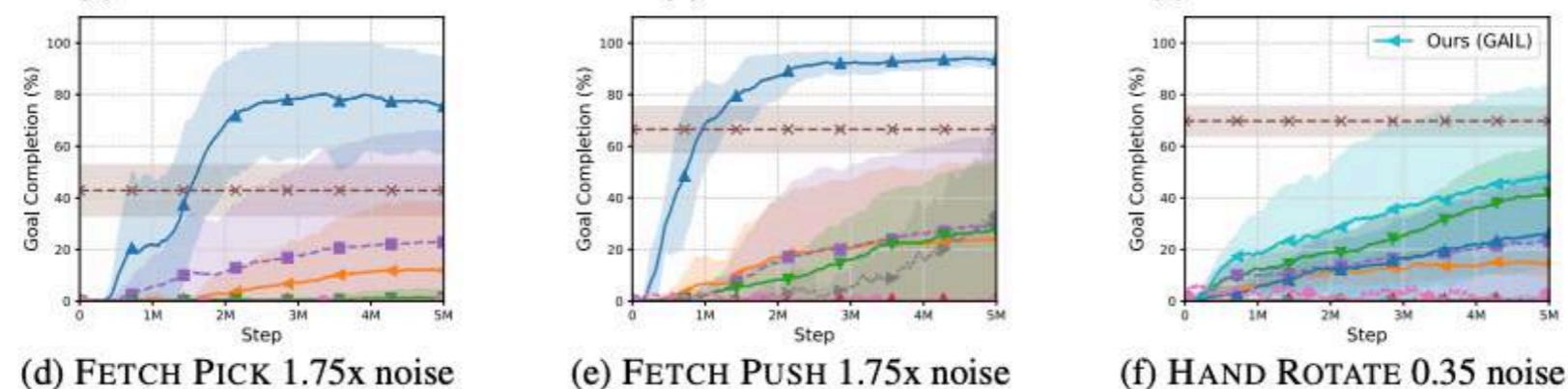
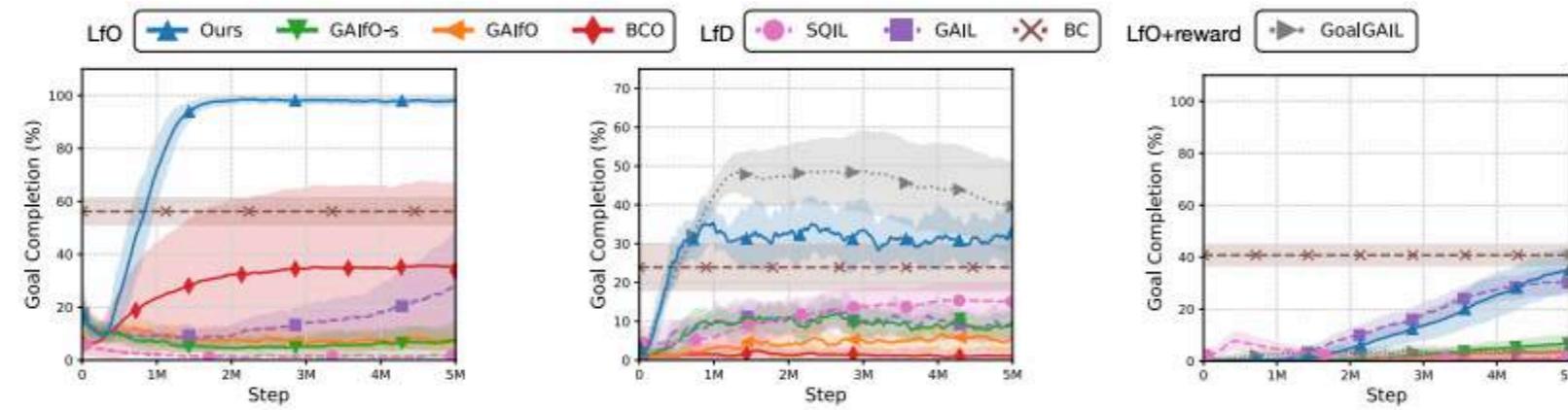
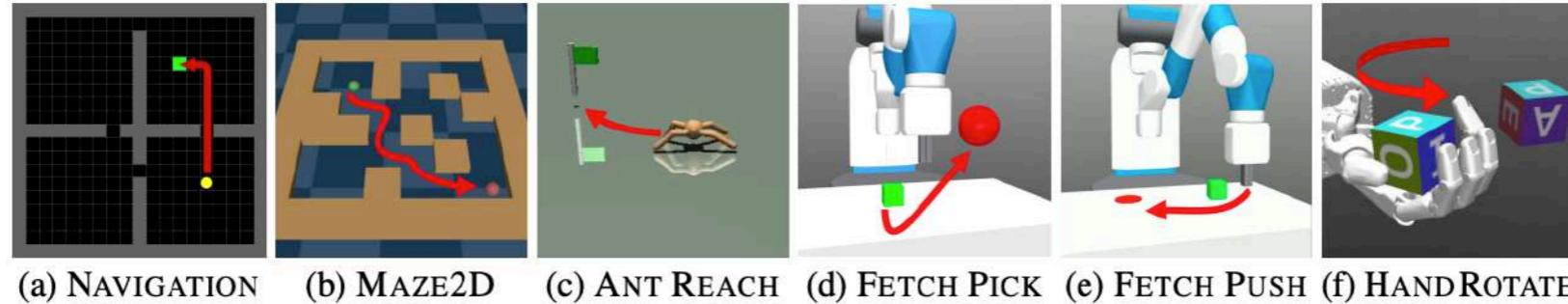
without expert's actions

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

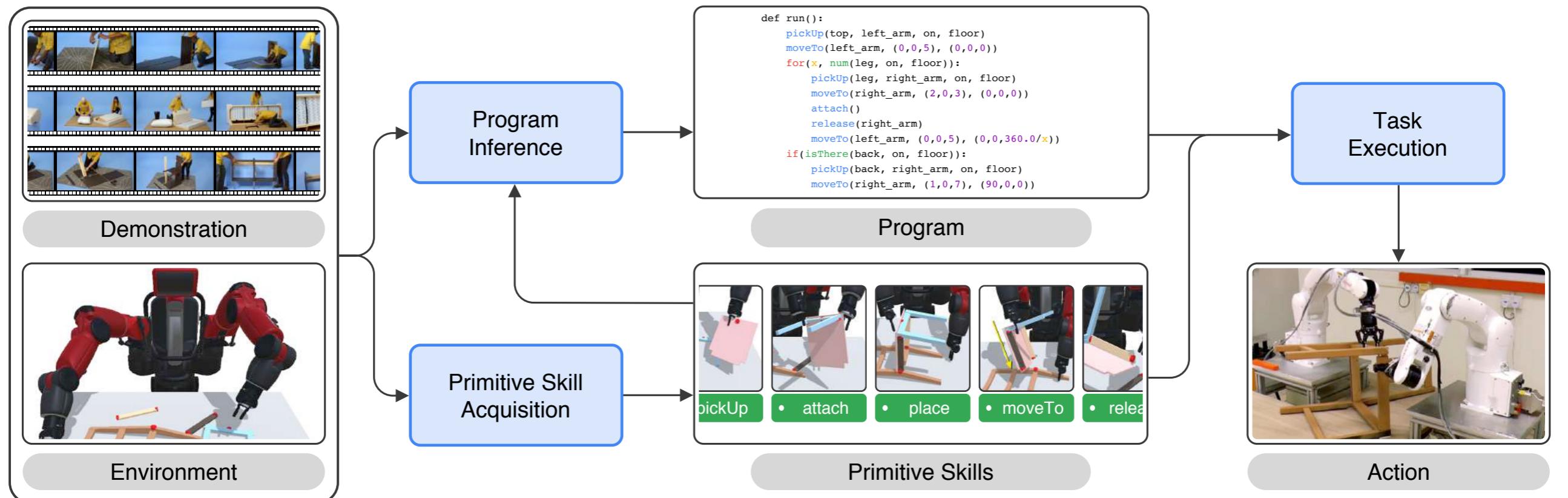
# Learning from Observation via Inferring Goal Proximity



# Experiments



# Program-Guided Framework for Interpreting and Acquiring Complex Skills



Interpretable

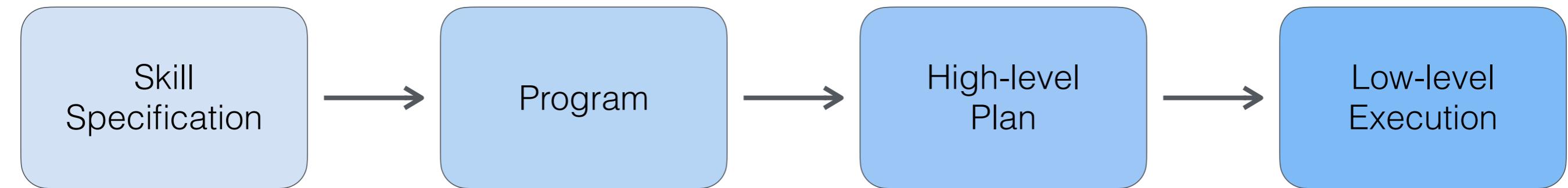
Programmatic / Generalizable

Hierarchical

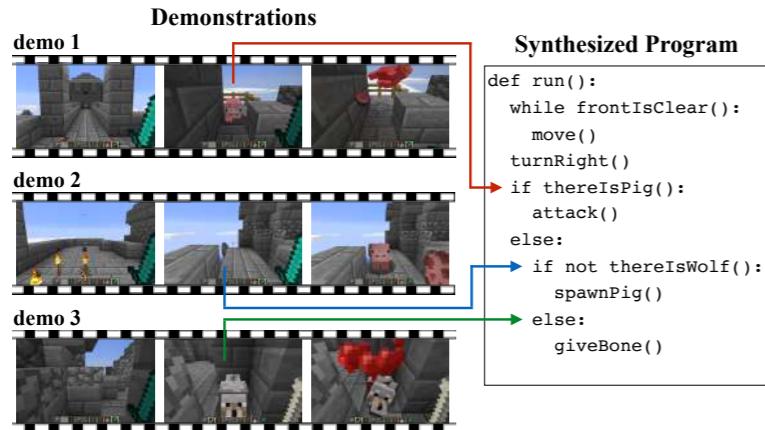
Modular

# Program-Guided Framework for Interpreting and Acquiring Complex Skills

---



# Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

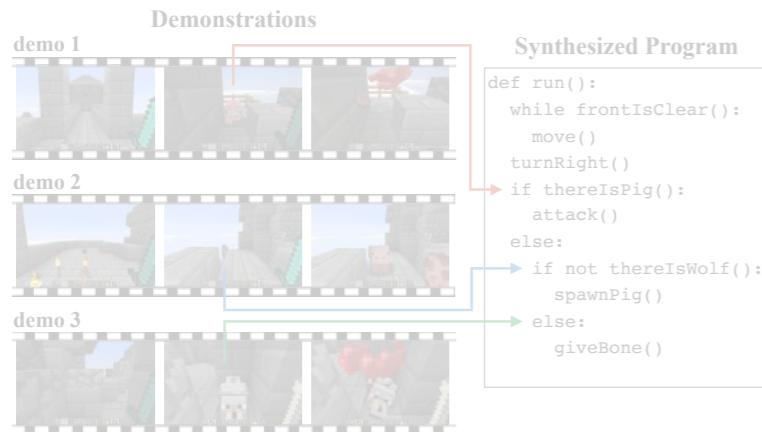
Skill  
Specification

Program

High-level  
Plan

Low-level  
Execution

## Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

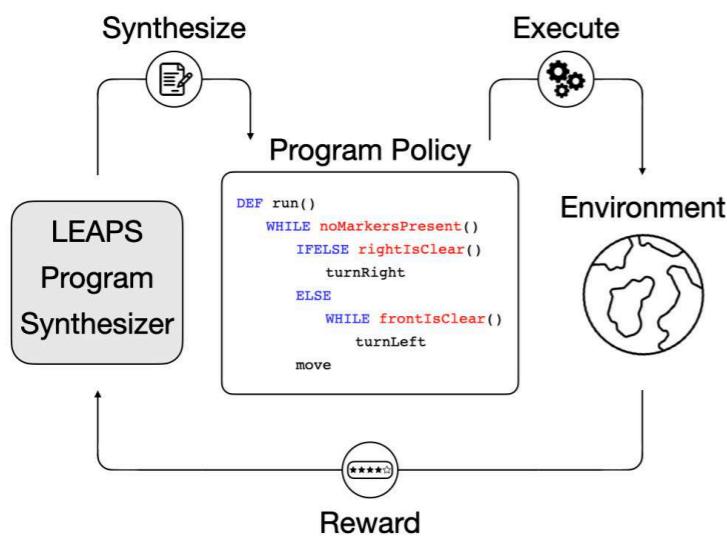
Skill Specification

Program

High-level Plan

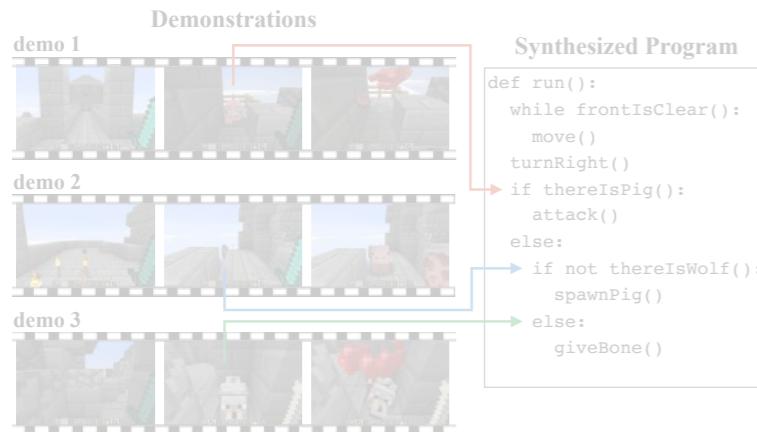
Low-level Execution

## 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



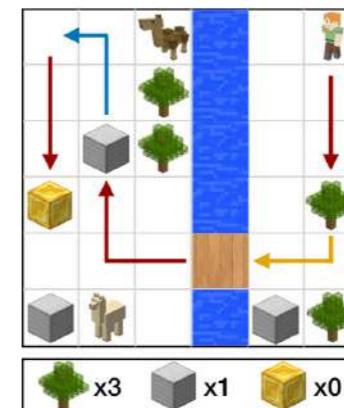
NeurIPS 2021

Program Guided Agent

**Program**

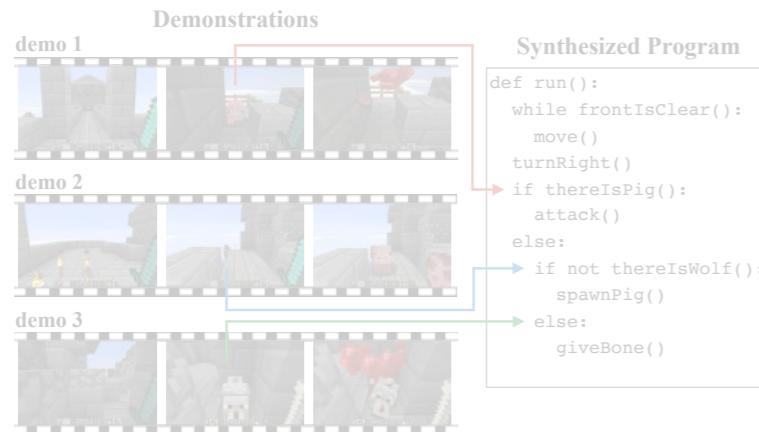
```

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)
    
```



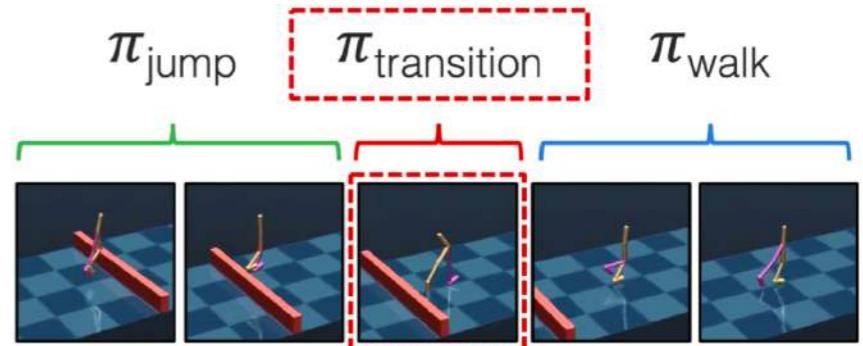
ICLR 2020 (Spotlight)

## 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

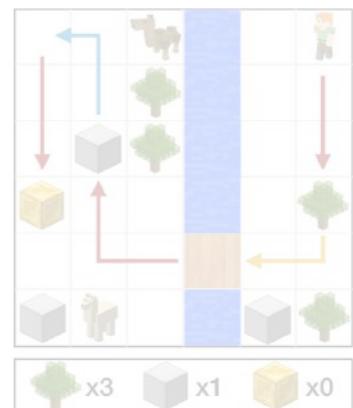


NeurIPS 2021

## Program Guided Agent

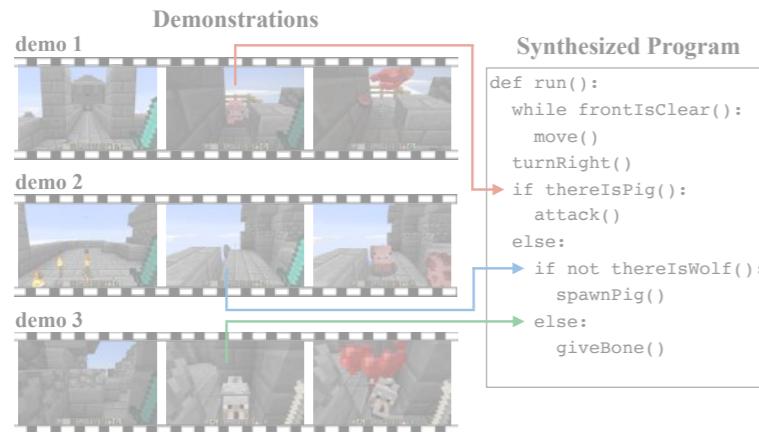
Program

```
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)
```

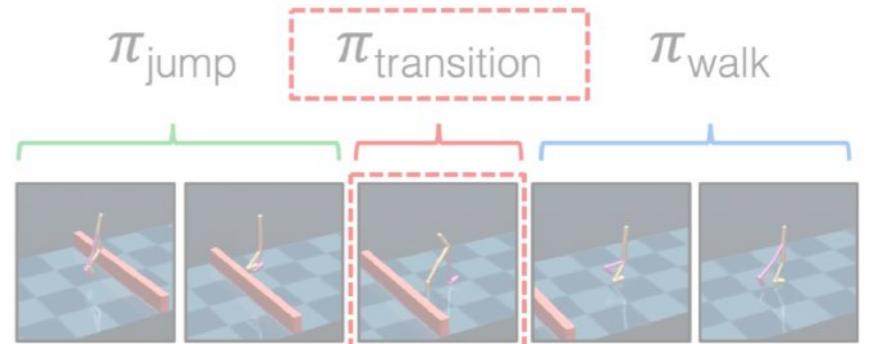


ICLR 2020 (Spotlight)

## Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018



ICLR 2019

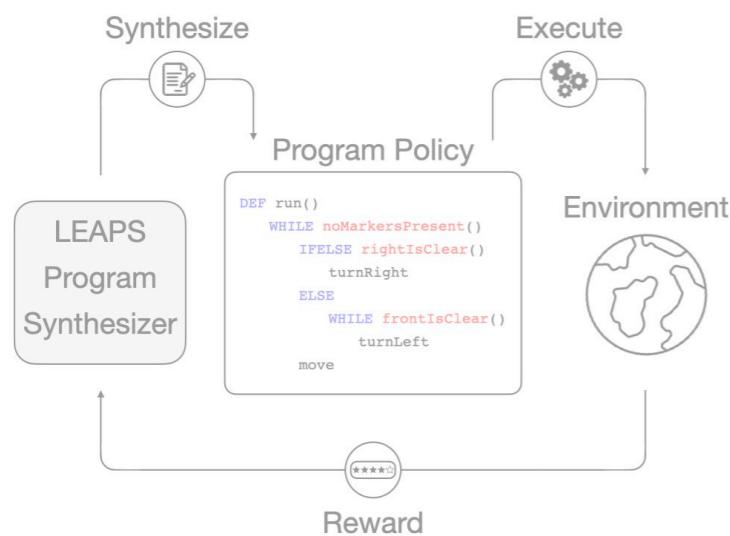
Skill  
Specification

Program

High-level  
Plan

Low-level  
Execution

## Learning to Synthesize Programs as Interpretable and Generalizable Policies

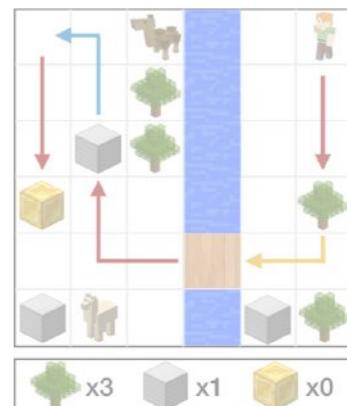


NeurIPS 2021

## Program Guided Agent

**Program**

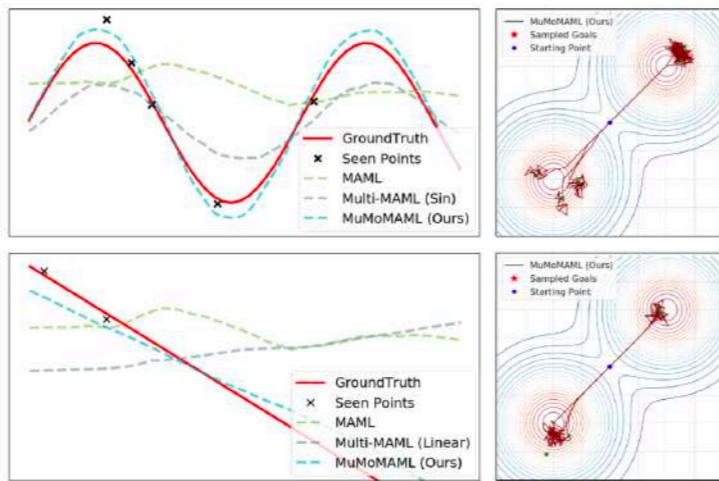
```
def Task():
    if is_there[River]:
        mine(Wood)
        build_bridge()
    if agent[Iron] < 3:
        mine(Iron)
        place(Iron, 2, 3)
    else:
        goto(4, 2)
    while env[Gold] > 0:
        mine(Gold)
```



ICLR 2020 (Spotlight)

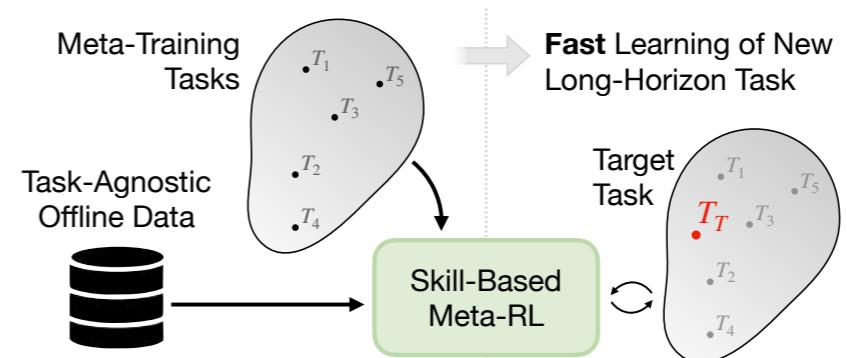
## Low-level Execution

### Toward Multimodal Model-Agnostic Meta-Learning



Meta-learning workshop @ NeurIPS 2018

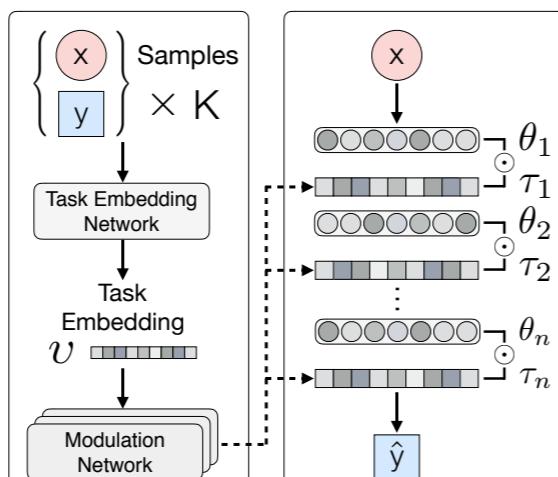
### Skill-based Meta-Reinforcement Learning



Deep RL workshop @ NeurIPS 2021

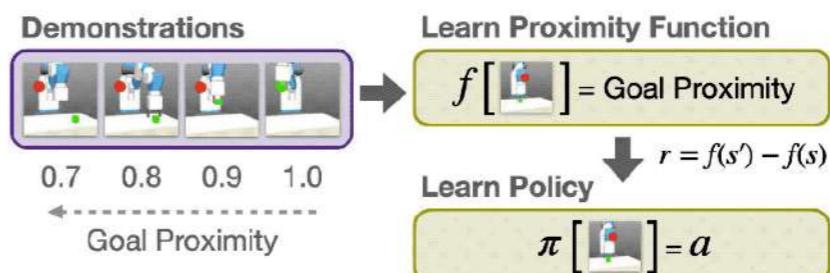
Meta-learning workshop @ NeurIPS 2021  
submitted to ICLR 2022

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



NeurIPS 2019 (Spotlight)

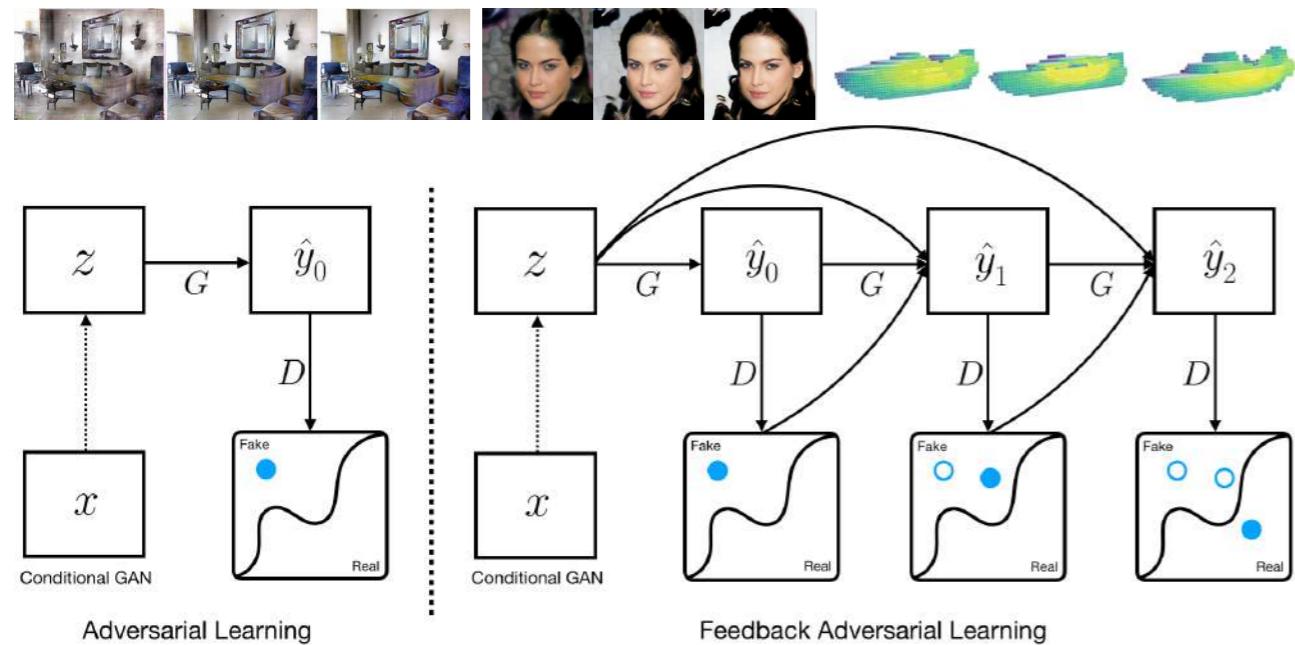
### Generalizable Imitation Learning from Observation via Inferring Goal Proximity



NeurIPS 2021

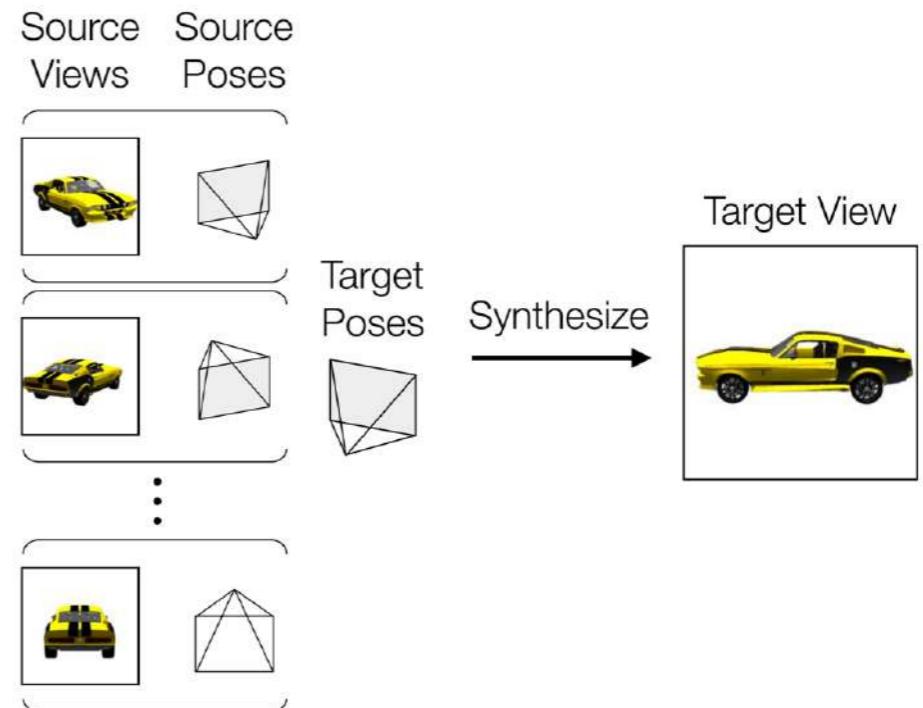
# Computer Vision / 3D Vision / Image Synthesis / GAN / Medical Imaging

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



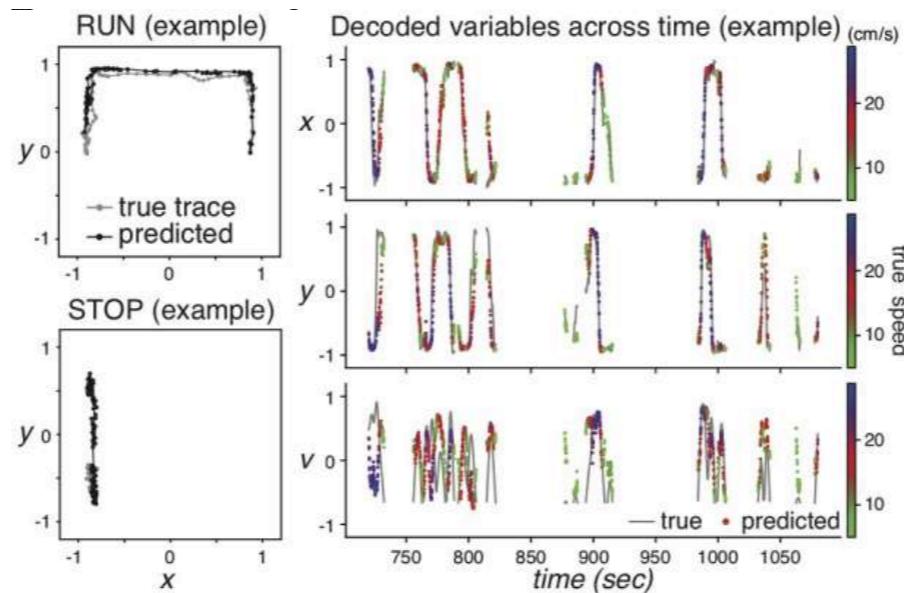
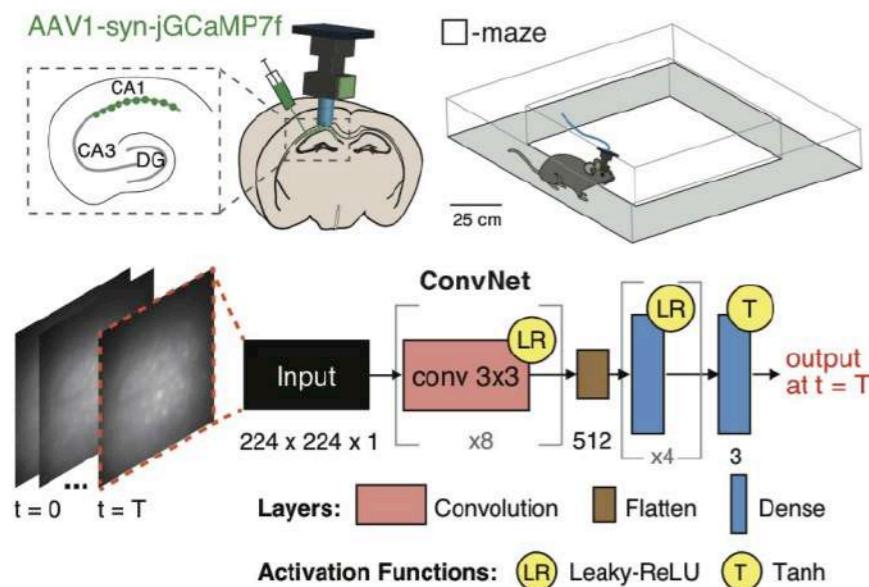
CVPR 2019

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

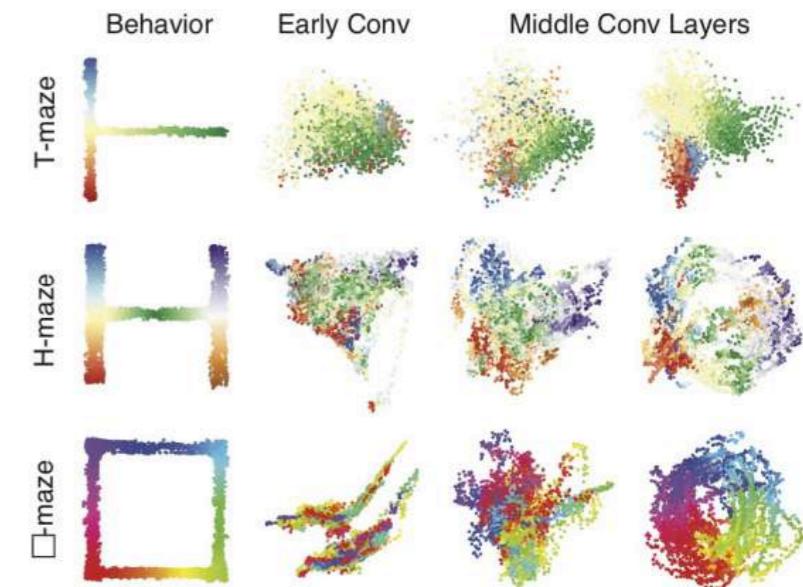


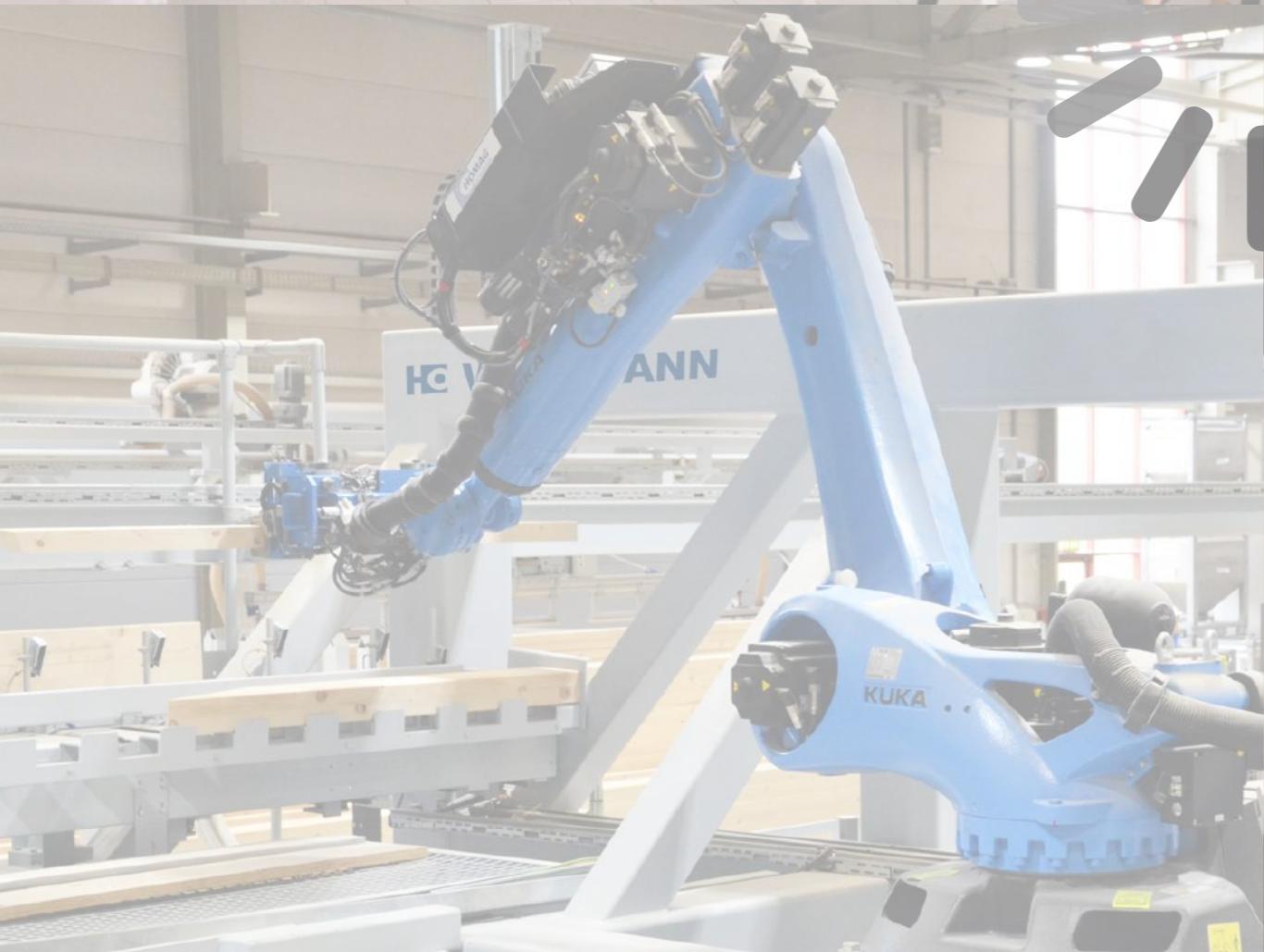
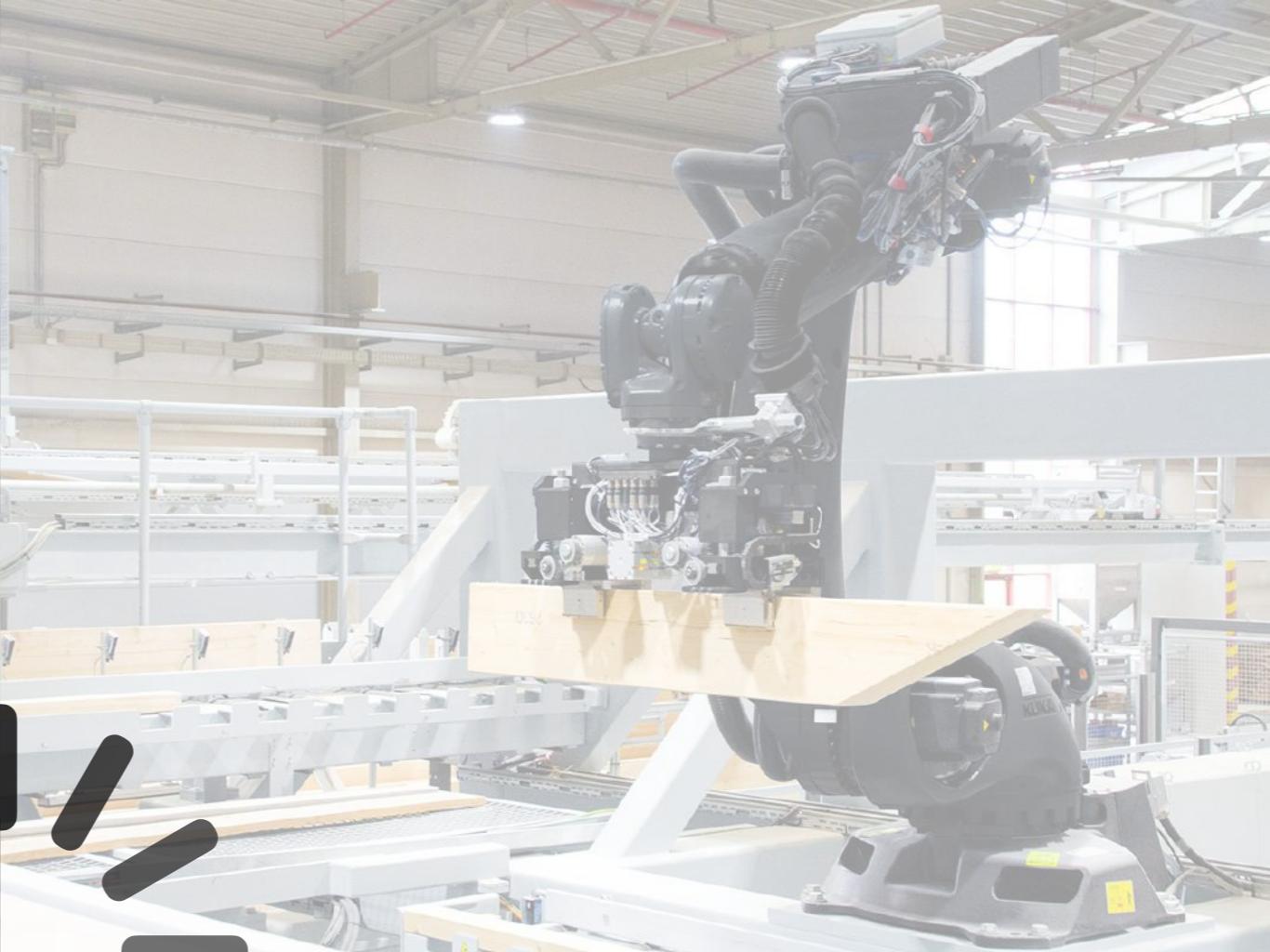
ECCV 2018

## Behavioral Clusters Revealed by End-to-end Decoding from Microendoscopic Imaging



bioRxiv 2021





Thank You

---

Questions?