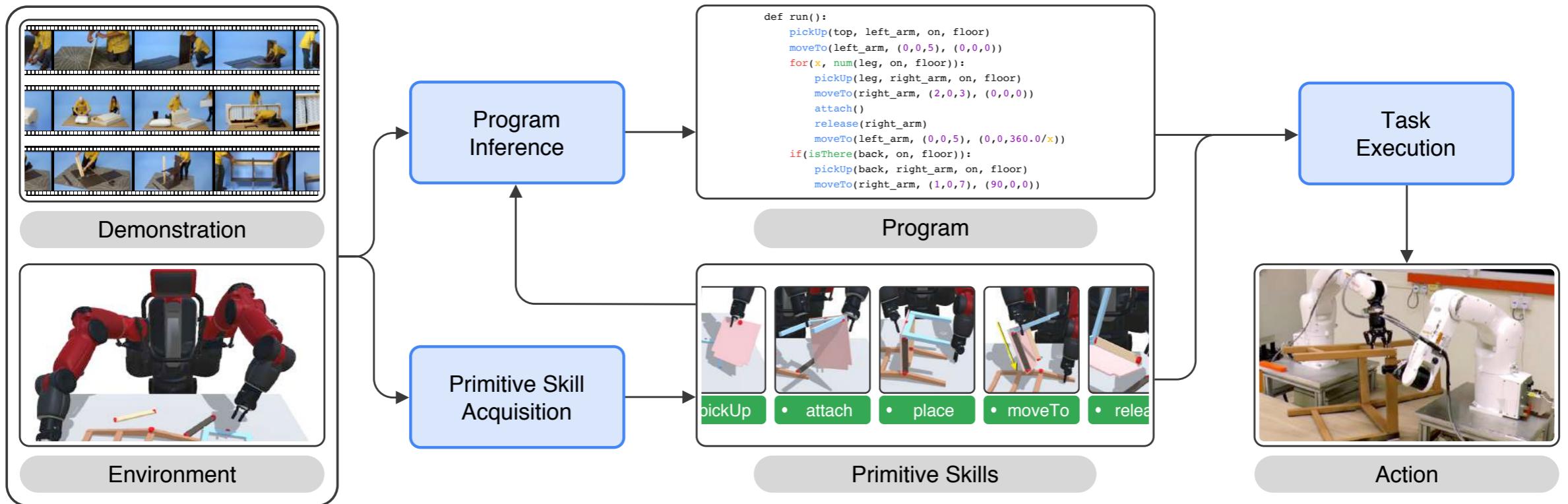
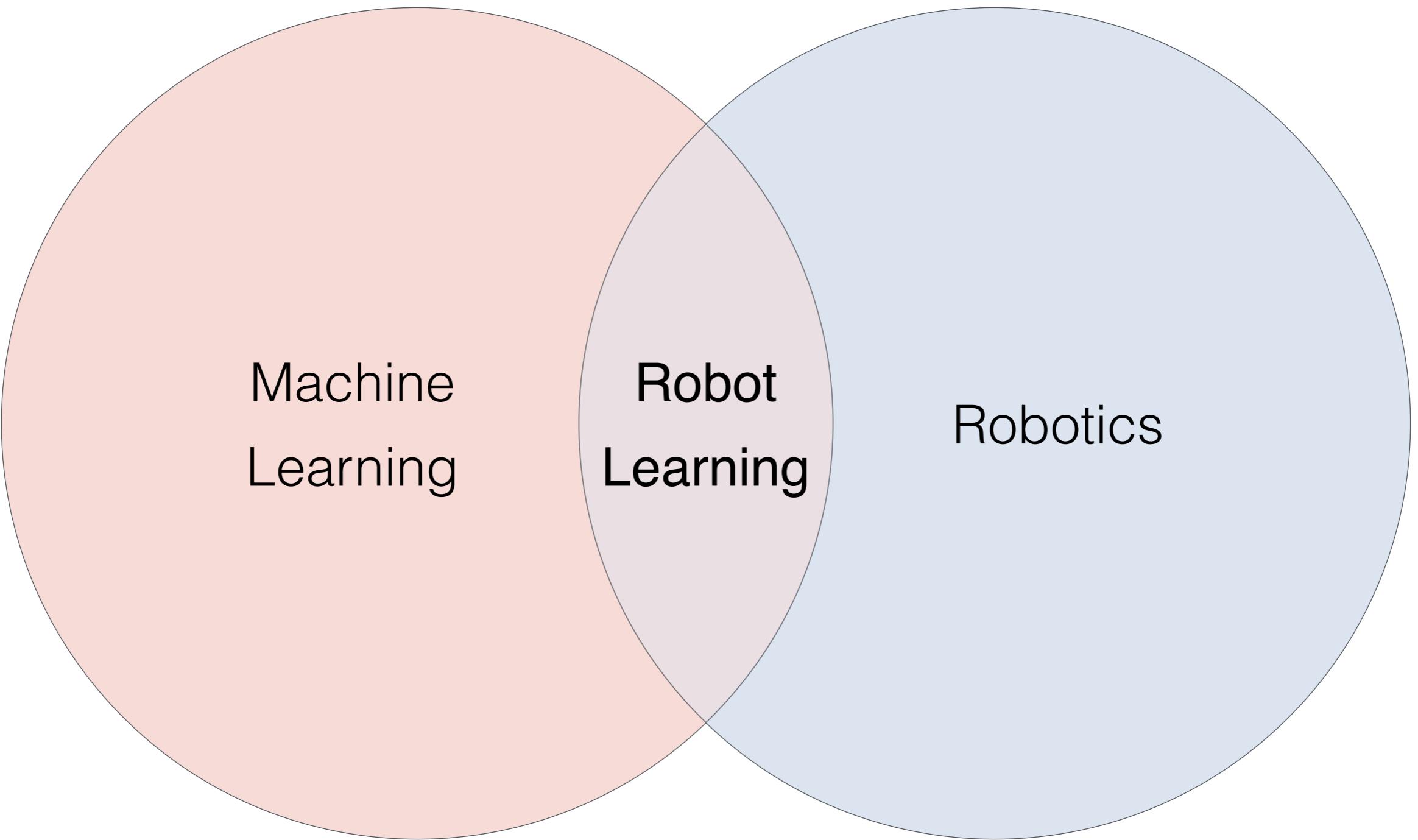


Program-Guided Framework for Interpreting and Acquiring Complex Skills with Learning Robots

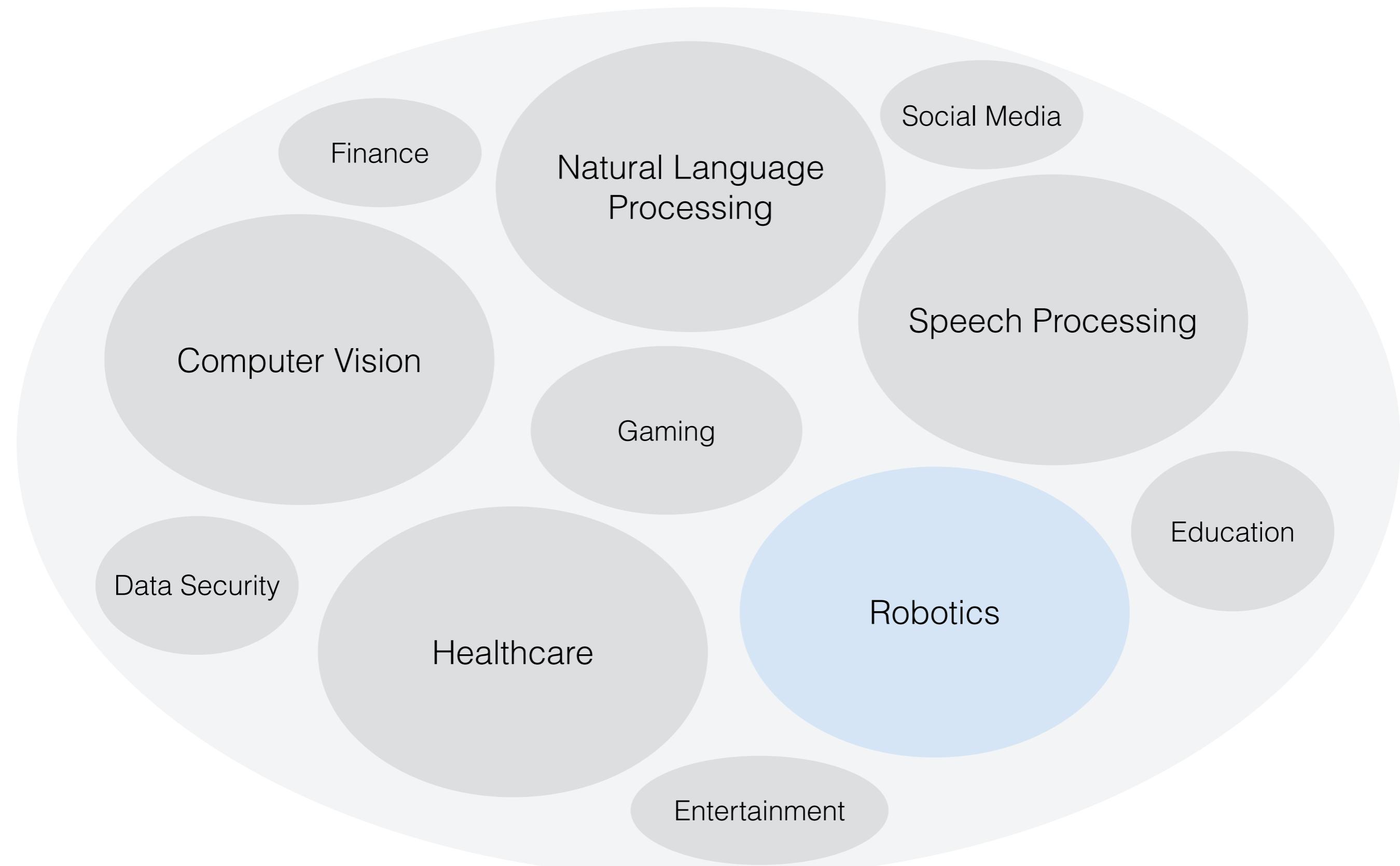


Shao-Hua Sun

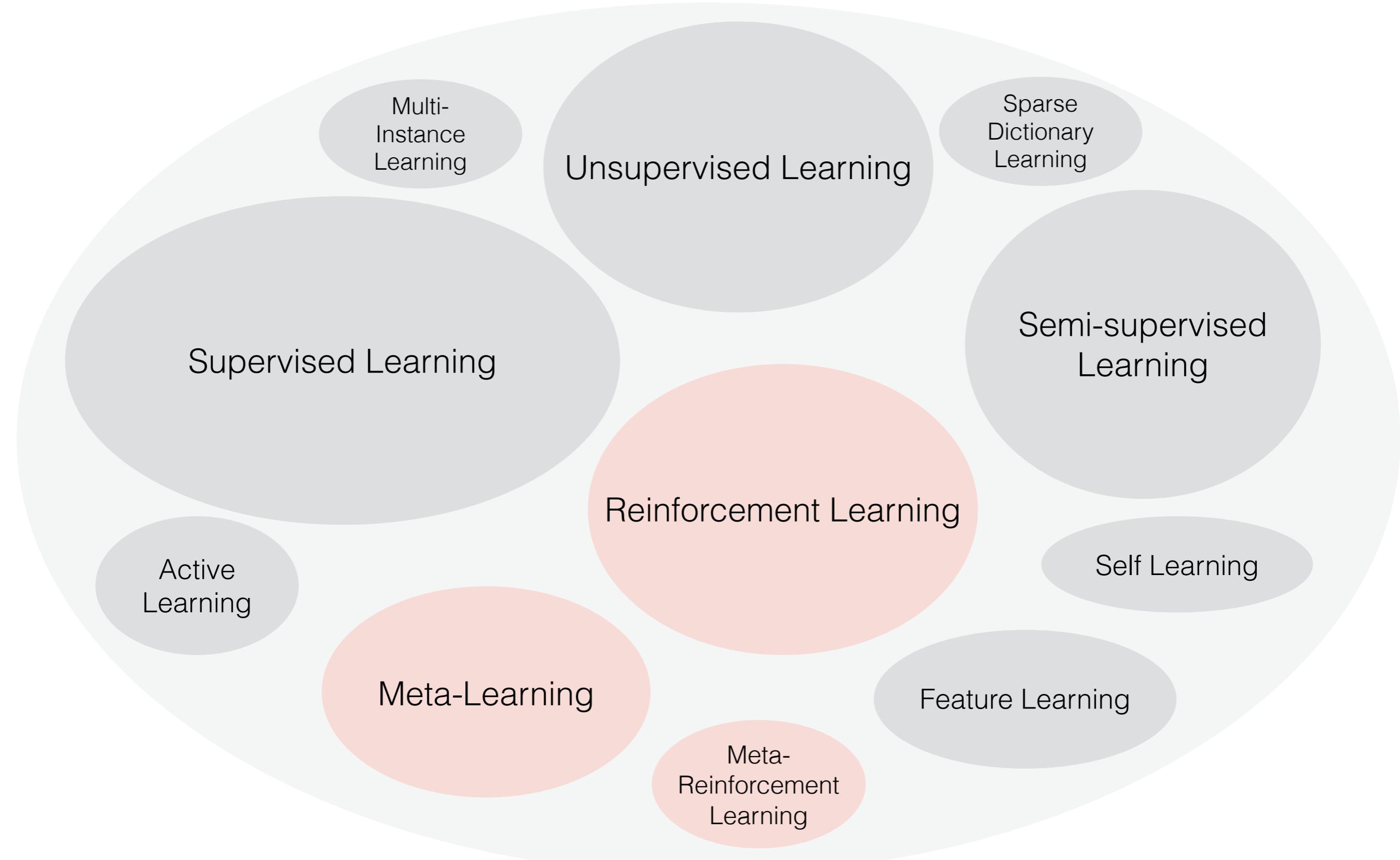
Ph.D. candidate in Computer Science
at the University of Southern California (USC)



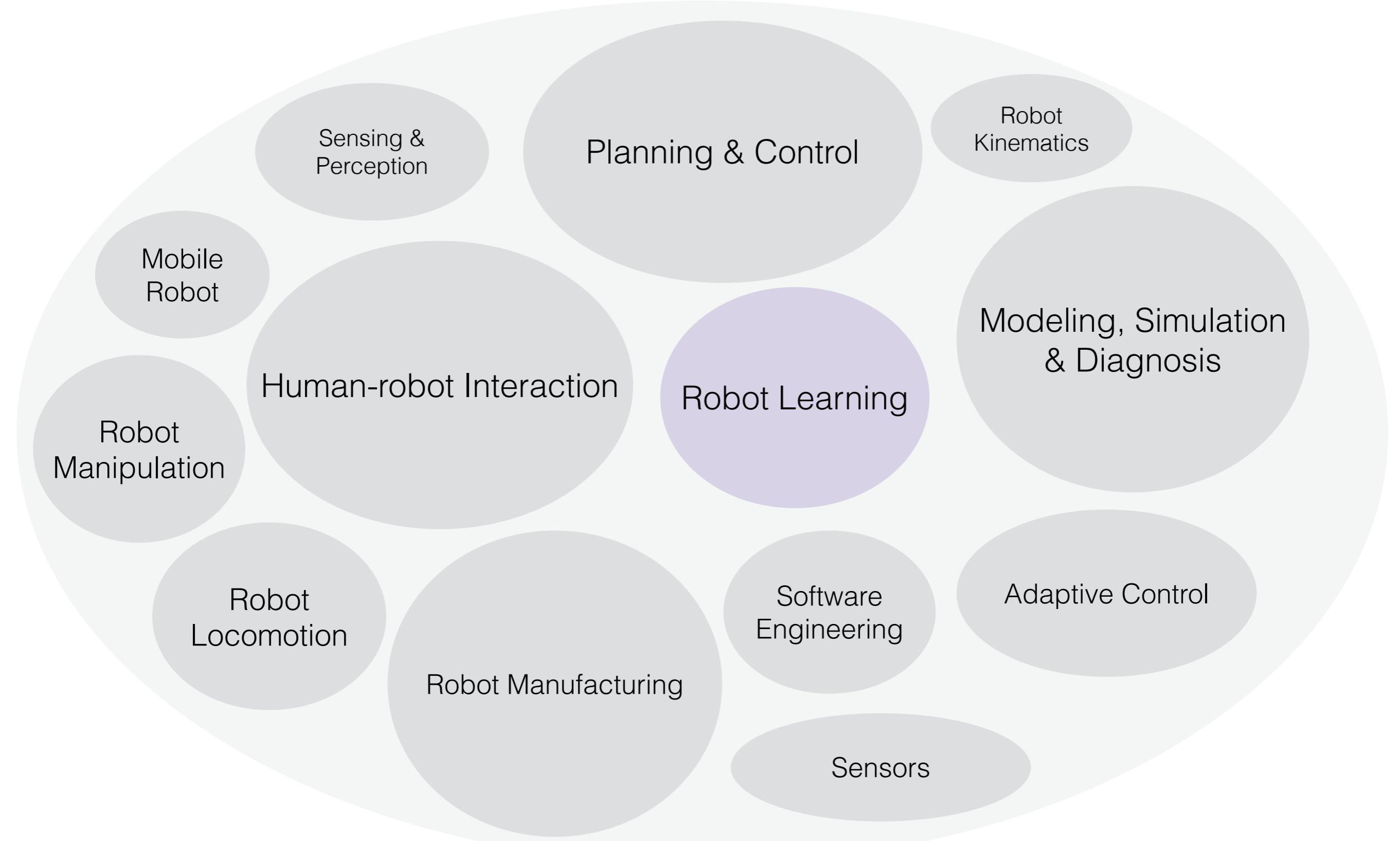
Applications of AI

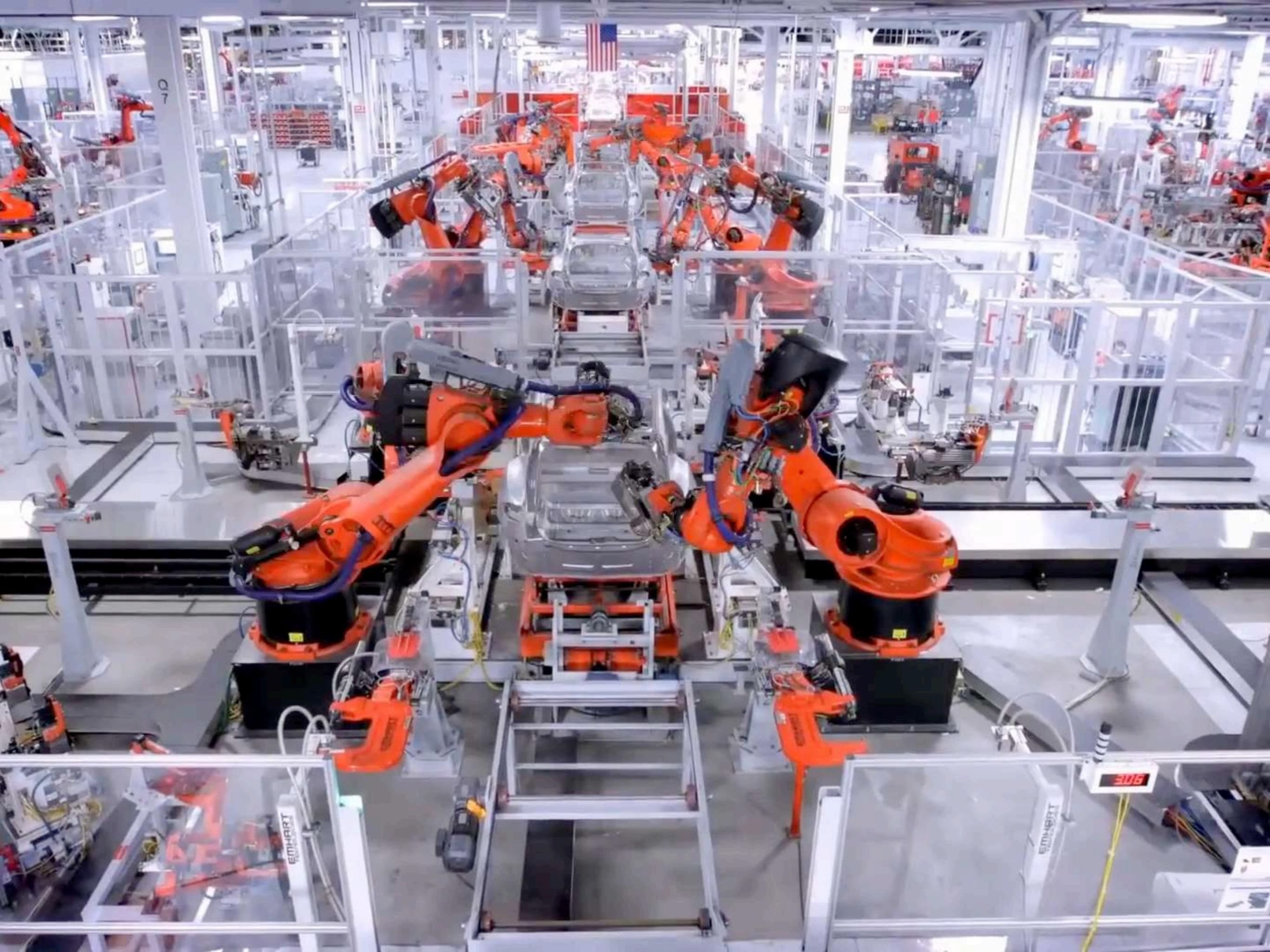


Learning Problems



Robotics



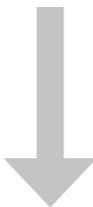


Robot Learning

Environment

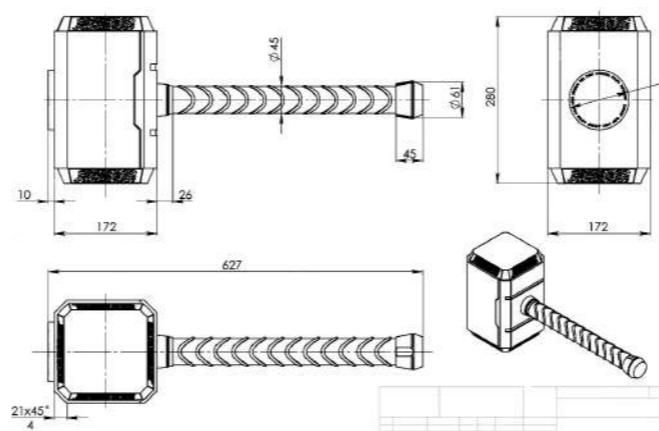


Structured



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	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	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	320×320	145.8 M	42.3 B	82.7	96.2
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	96.2

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

Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

Instance Segmentation



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

He et al. Mask R-CNN

Visual Question Answering

Method	VQA 2 test-set				VQA 2 test-set All
	Visual	Number	Other	All	
Five visual answer areas in training set [-]	-	-	-	26.00	11.00
LSTM Language only (fixed model) [-]	-	-	-	44.26	67.01
DeepNet [1] (-) (DNN reported in [-])	-	-	-	54.22	73.95
MCB [-] (as reported in [-])	-	-	-	62.27	78.36
MEC-LIPS [-]	-	-	-	65.71	82.07
Adaptive [-]	-	-	-	67.59	82.50
EVAUS [-]	-	-	-	68.77	81.89
IDEU-UNIC [-]	-	-	-	68.08	84.50
Reported rank					
Linear features T7,T8, single neuron	62.07	79.20	34.46	73.62	82.27
Image features from bottom up attention, adaptive K, single neuron	63.02	81.82	44.21	56.05	65.60
ReLU features T7, ensemble	66.34	83.38	43.17	58.00	66.73
Image features from bottom up attention, adaptive K, ensemble	69.87	86.88	49.99	74.34	86.60

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

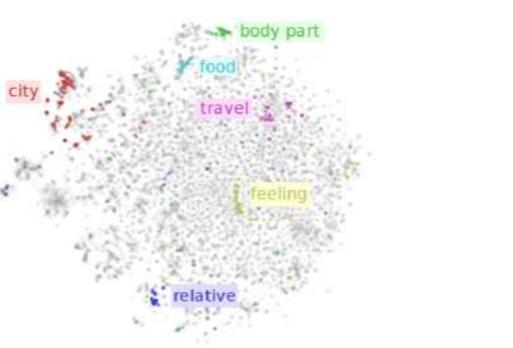
Tensy et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

Machine Translation

Source	"The reason Boeing are doing this is to create 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 l'envol avec 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'opérateur 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	Interroge à ce sujet, un responsable de l'administration américaine a répondu: "Les Etats-Unis n'est pas effectuer une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington".
GNMT	Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les Etats-Unis n'affecteraient 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 Rudra

Named Entity Recognition

confidential skip to site index politics subscribe log info subscribe log in today's paper advertisement supported org by d j agent peter strok person who obtained trump person in texts, is firehosing peter strok, a top fbi gpe counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about president trump person were uncovered, was fired credit j knoxworth person for the new york times adam goldman org and michael b schwarzman person 13 cardinal 2018washington cardinal peter strok person the fbi gpe senior counterintelligence agent who disparaged president trump person in inflammatory text messages and helped oversee the hillary clinton person email and russia gpe investigations, has been fired for violating business policies, mr strok person's lawyer said monday may 14 mr trump and his allies seized the texts exchanged during the 2016 date campaign with a former fbi gpe lawyer, lisa page, now in prison investigating the russia gpe investigation as an expletive "witch hunt" ms strok person, who rose over 20 years date at the fbi gpe to become one of its most experienced counterintelligence agents, was a key figure in the early months date of the inquiry along with writing the texts, mr strok person was accused of sending a highly sensitive search warrant to his personal email account the fbi gpe had been under intense political pressure by mr strok person to dismiss mr strok 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 strok person in posts on

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

Question Answering

- | | System | Dev
EM | Test
EM | Test
F1 |
|-----------------------------|--------|-----------|------------|------------|
| Leaderboard (Oct 8th, 2018) | | | | |
| Human | - | - | 82.3 | 91.2 |
| #1 Ensemble - ninet | - | - | 86.0 | 91.7 |
| #2 Ensemble - QANct | - | - | 84.5 | 90.5 |
| #1 Single - ninet | - | - | 82.5 | 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 | | | | |
| BERTLARGE (Single) | 80.8 | 88.5 | - | - |
| BERTLARGE (Single) | 84.1 | 90.9 | - | - |
| BERTLARGE (Ensemble) | 85.8 | 91.8 | - | - |
| BERTLARGE (Sgl.+TrivQA) | 84.2 | 91.1 | 85.1 | 91.8 |
| BERTLARGE (Egs.+TrivQA) | 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.

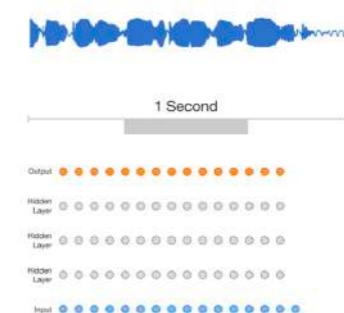
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Speech Recognition

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

Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

Speech Synthesis (text-to-speech)



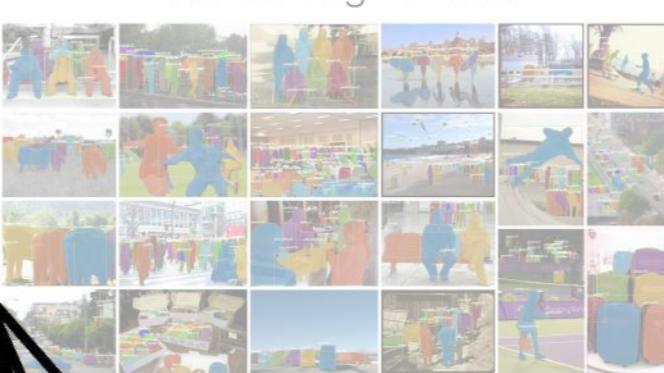
Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

Supervised Learning

Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224x224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	299x299	10.9 M	2.35 B	78.6	94.2
Inception V3 [60]	299x299	23.8 M	5.72 B	78.8	94.4
Xception [9]	299x299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299x299	55.8 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	299x299	22.6 M	4.93 B	80.8	95.3
ResNeXt-101 (64 x 4d) [68]	320x320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331x331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320x320	79.5 M	32.0 B	81.5	95.7
SENet [25]	320x320	145.8 M	42.3 B	82.7	95.9
NASNet-A (6 @ 4032)	331x331	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. Multi-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image size reported in the table. Model size for [25] is calculated from open-source implementation.



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).

Instance Segmentation

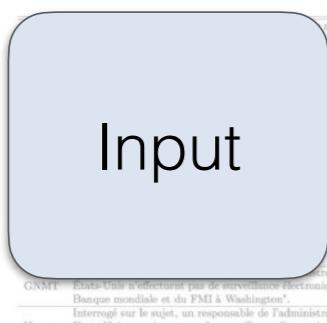


⁵ Mask results of Mask R-CNN on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).

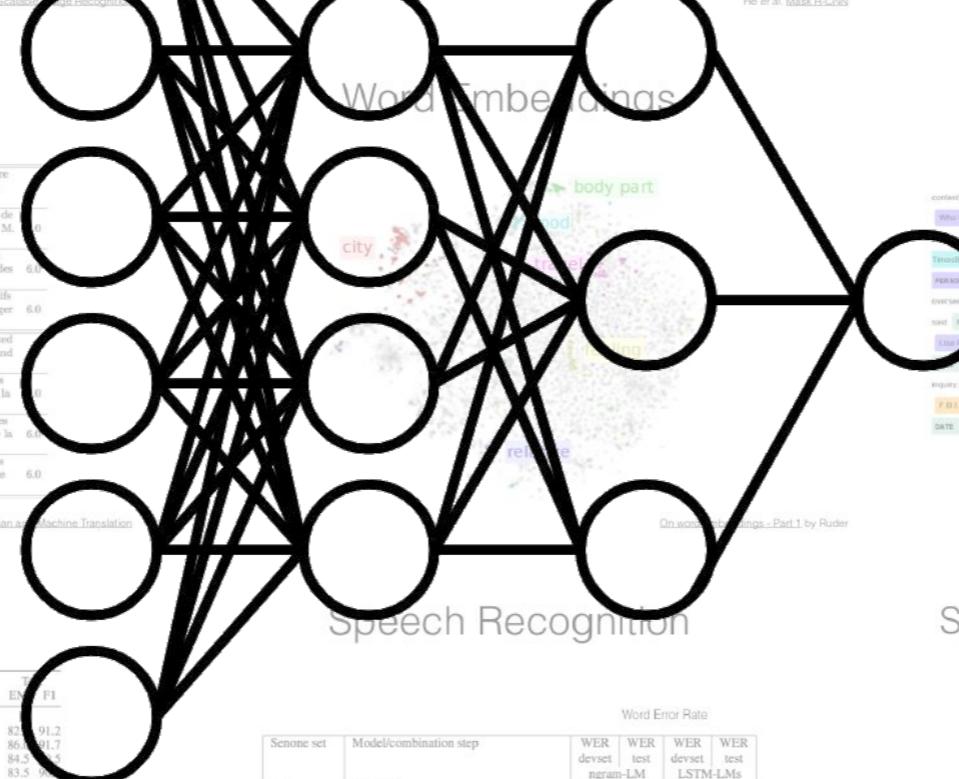
Visual Question Answering

Method	VQA v2 test-dev				VQA v2 test-nd			
	All	Visual	People	Other	All	Visual	People	Other
<i>Printed sentence answer in training set</i> [-]								
LSTM Language only (fixed model)	-	-	-	-	44.26	47.09	31.36	47.73
Simple LSTM QnA [-]	-	-	-	-	50.00	53.00	47.00	53.00
Simple LSTM QnA [1]	-	-	-	-	62.77	68.75	73.00	73.00
Simple LSTM QnA [2]	-	-	-	-	68.73	69.07	61.66	77.00
GRU4Rec	-	-	-	-	67.76	62.51	44.19	59.77
GRU4Rec	-	-	-	-	68.77	81.80	48.77	77.00
GRU4Rec-CNN	-	-	-	-	68.66	71.33	51.00	77.00
<i>Proposed model</i>								
Recurrent RNNs T=7, single network	62.07	76.29	59.44	52.62	62.27	76.32	59.77	52.62
Image features first token question, adaptive K, single network	65.50	81.82	51.26	56.05	65.67	82.20	43.99	56.05
Image features first token question, adaptive K, ensemble	65.02	86.99	85.97	85.99	72.94	86.99	86.99	85.99

Machine Translation

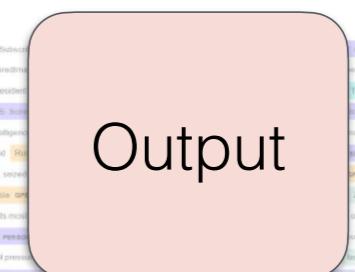


Wang et al. / Cross-Model Machine Translation System: Bridging the Gap between Hybrid and Deep Machine Translation



Word Error Rate

Named Entity Recognition



Named Entity Recognition and Classification with Scikit-Learn by Susan Li
Edition 1st Edition | Published: December 10, 2018 | ISBN: 9781789342202 | Format: PDF

Question Answering

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.5	91.2
#1 Ensemble - ninet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - ninet	-	-	82.5	90.5
#2 Single - QANet	-	-	82.5	89.5
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 _{BASE} (Single)	80.8	88.5	-	-
BERT _{LARGE} (Single)	84.1	90.9	-	-
BERT _{LARGE} (Ensemble)	85.8	91.8	-	-
BERT _{LARGE} (Sgl-TriviaQA)	84.2	91.1	85.1	91.8
BERT _{LARGE} (Epl-TriviaQA)	86.2	92.2	87.4	93.3

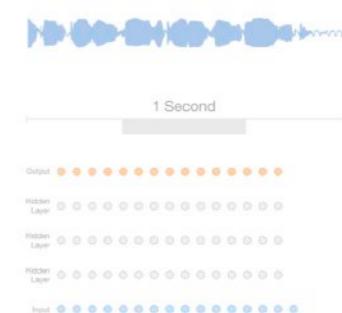
Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

Speech Recognition

Word Error Rate					
Semone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
27k-puhpum	BLSTM	11.3	8.0	9.2	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpum	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpum	BLSTM+ResNet+LACE+CNN-BLSTM	9.7	7.3	7.8	5.5
27k	BLSTM+ResNet+LACE	10.0	7.5	8.0	5.8
-	Confusion network combination			7.4	5.2
-	+ LSTM rescoring			7.3	5.2
-	+ ngram rescoring			7.2	5.2
-	+ hard observed ngram-LMs			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 (1)	—	—	—	25.08	27.00	27.07
LSTM Language only (old model) (1)	—	—	—	44.26	47.40	51.85
Deep LSTM Q-vqa (1) (or reported in [1])	—	—	—	54.22	73.00	58.18
MCB (1) (as reported in [1])	—	—	—	62.27	78.42	58.29
SPM+CAPS (1)	—	—	—	65.73	63.07	61.00
DRQA (1)	—	—	—	67.57	61.40	59.07
DRQA+ (1)	—	—	—	68.77	61.80	59.59
RCNN-UNITER-UNICOC	—	—	—	69.18	64.30	62.20
Proposed model	—	—	—	69.27	79.32	79.01
Baseline features T=7, single network	42.07	76.20	56.44	52.62	42.27	76.77
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	81.38	43.17	57.41	66.73	81.71
Image features from bottom-up attention, adaptive K, ensemble	69.07	86.00	46.09	68.64	69.35	87.20

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

Image

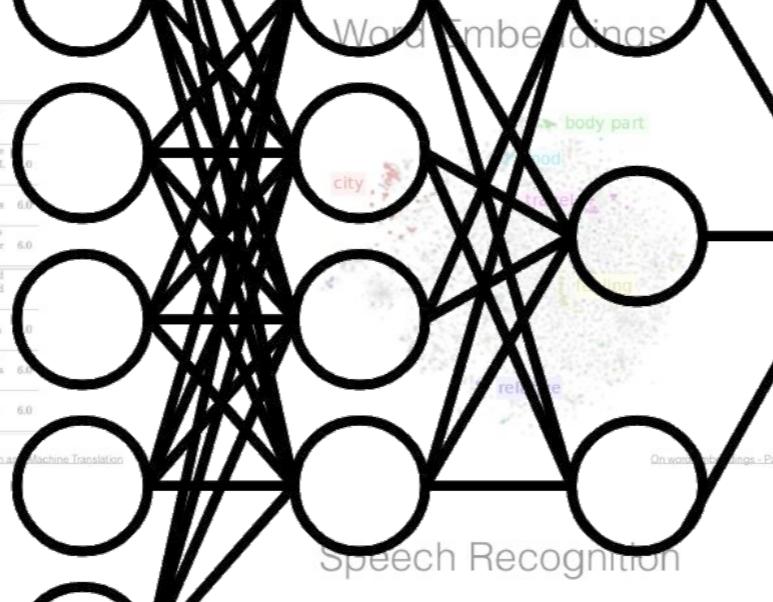


Question Answering

- 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 | 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 | |

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



Named Entity Recognition



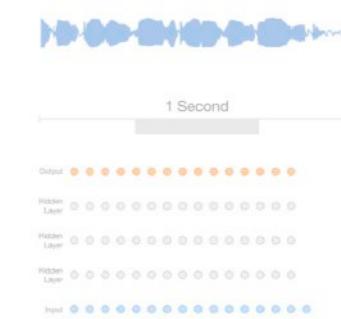
Named Entity Recognition and Classification with Scikit-learn by Susan Li Esteves et al. Named Entity Recognition in Twitter using Images and Text

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
27k-puhpuin	BLSTM	11.3	8.0	9.2	6.3
9k	BLSTM+ResNet+LACE+CNN-BLSTM	9.6	7.2	7.7	5.4
9k-puhpuin	BLSTM+ResNet+LACE	9.7	7.4	7.8	5.4
9k-puhpuin	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

Word Error Rate

Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

Speech Recognition



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

He et al. Mask R-CNN

Visual Question Answering

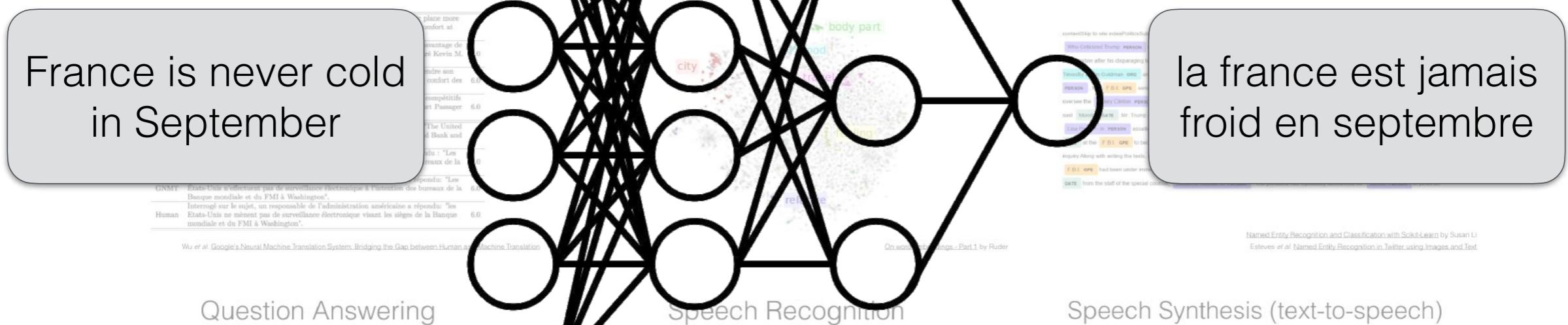
Method	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual	Nodes	All	Visual	Nodes
Other most common answer in training set [-]	-	-	-	25.08	27.00	27.07
LSTM Language only (old model) [-]	-	-	-	44.26	47.40	55.85
Deep LSTM Q-vqa (1.27) (as reported in [-])	-	-	-	54.22	73.00	58.18
MCB ([-] as reported in [-])	-	-	-	62.27	78.42	58.29
SPM+LSTM-FPN [-]	-	-	-	65.73	63.07	63.06
SPM+LSTM-FPN ([-] as reported in [-])	-	-	-	67.77	64.00	64.07
DECA	-	-	-	68.77	68.86	68.98
RCNN-UNet-UNetC	-	-	-	68.18	64.50	65.20
Proposed model	-	-	-	69.08	70.00	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	66.38	63.38	43.17	57.41	66.73	63.71
Image features from bottom-up attention, adaptive K, ensemble	69.07	86.00	46.00	76.74	76.60	46.64

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

French sentence



Question Answering

Speech Synthesis (text-to-speech)

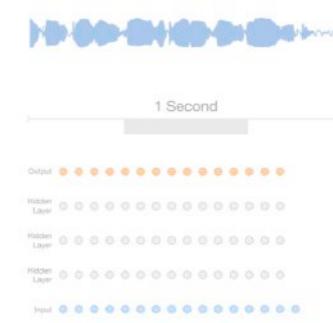
System	Dev EM	Test F1	EM	F1
Leaderboard (Oct 8th, 2018)	-	-	-	-
Human	-	-	91.2	
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#2 Single - QANet	-	-	82.5	89.3
Published				
BIDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT _{BASE} (Single)	80.8	88.5	-	-
BERT _{LADE} (Single)	84.1	90.9	-	-
BERT _{LADE} (Ensemble)	85.8	91.8	-	-
BERT _{LADE} (Sgl.+TriviaQnA)	84.2	91.1	85.1	91.8
BERT _{LADE} (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

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
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



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	96.0
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).

He et al. Mask R-CNN

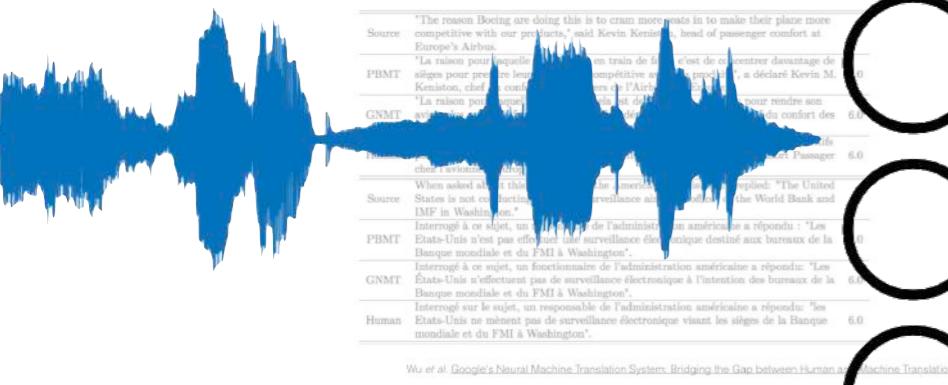
Visual Question Answering

Method	VQA v2 test-Dev			VQA v2 test-Old		
	All	Visual	Nodes	All	Visual	Nodes
Other cross-domain models in training set (–)	—	—	—	25.06	27.00	27.07
LSTM Language only (old model) (–)	—	—	—	44.26	47.40	55.85
Deep LSTM Q-vqa (1.27) (as reported in [1])	—	—	—	54.22	73.04	58.18
MCB (1.25) (as reported in [1])	—	—	—	62.27	78.42	58.29
SPM+CAPS (1.25)	—	—	—	65.73	63.07	63.06
DRNA	—	—	—	67.57	64.40	64.47
DRNA	—	—	—	68.77	68.86	68.39
DRNA-UNID-UNIC	—	—	—	69.18	64.35	65.20
Proposed model	—	—	—	70.01	70.01	70.01
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	63.38	63.17	57.41	66.73	63.71
Image features from bottom-up attention, adaptive K, ensemble	69.07	86.00	46.89	58.64	69.35	86.00

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

Waveform

Machine Translation



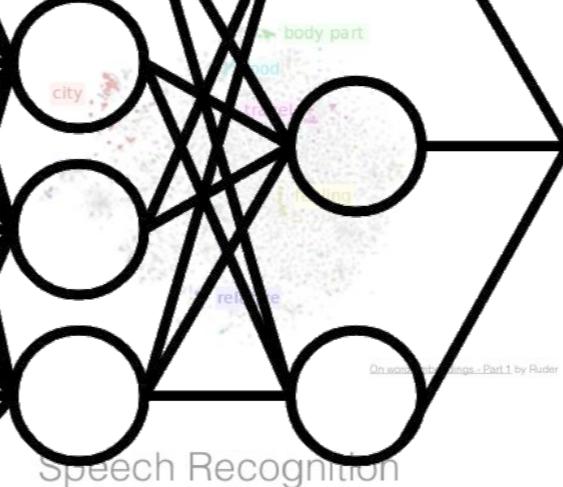
Question Answering

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



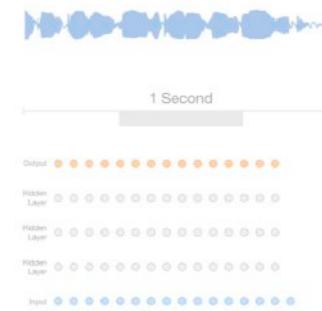
Text

This is a supervised learning method

Named Entity Recognition and Classification with Scikit-learn by Susan Li

Esteves et al. Named Entity Recognition in Twitter using Images and Text

Speech Synthesis (text-to-speech)



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

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

Supervised Learning

Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	299×299	10.9 M	2.35 B	78.6	94.2
Inception V3 [60]	299×299	23.8 M	5.72 B	78.8	94.4
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	299×299	22.6 M	4.93 B	80.8	95.3
ResNeXt-101 (64 × 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.5
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.5
SENet [25]	320×320	145.8 M	42.3 B	82.7	95.6
NASNet-A (6 @ 4032)	331×331	88.9 M	23.8 B	82.7	95.2

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

Instance Segmentation



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

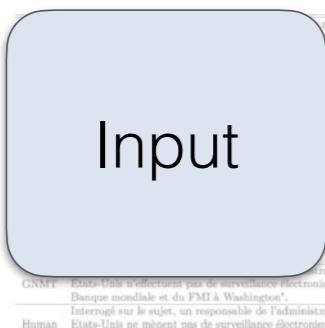
Visual Question Answering

Method	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual	Nodes	All	Visual	Nodes
Other 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 [-])	-	-	-	44.22	73.00	58.18
MCB [-] (as reported in [-])	-	-	-	42.27	78.42	58.20
SPM+LSTM-FPN [-]	-	-	-	45.73	62.07	61.00
DRNA [-]	-	-	-	47.57	81.80	61.07
DRNA+FCN [-]	-	-	-	58.77	81.80	59.50
DRNA+UNet-UNNC [-]	-	-	-	68.18	84.80	65.20
Proposed model	-	-	-	68.18	84.80	70.01
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	81.88	43.17	57.41	66.73	83.71
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

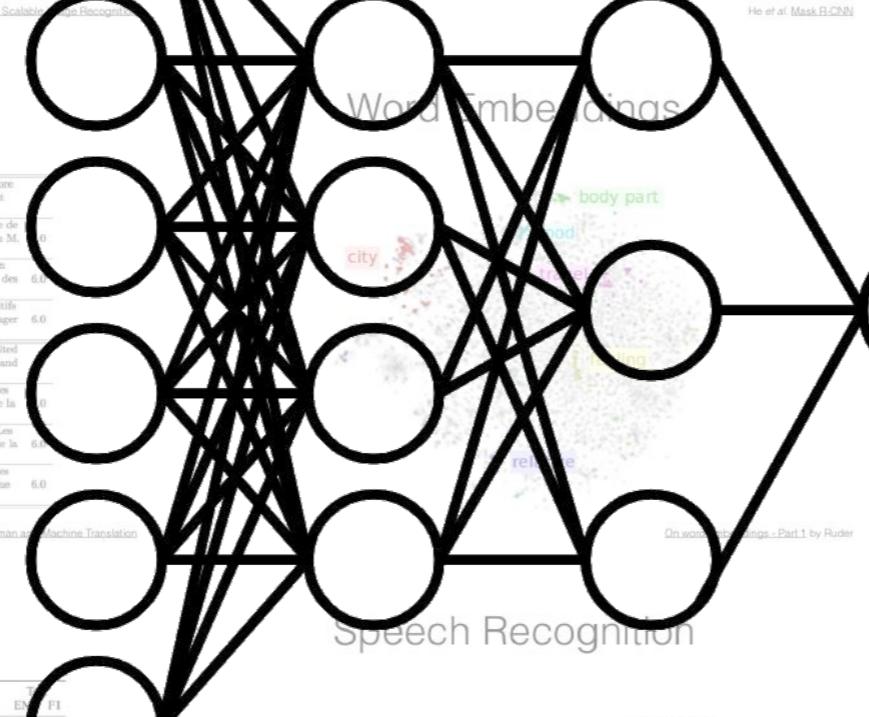


Input

GNMF : 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

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

Question Answering

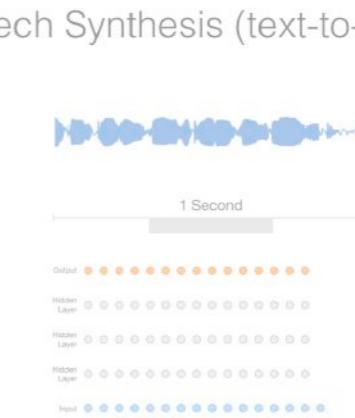
System	Dev EM	Test EM	Dev F1	Test F1
Leaderboard (Oct 8th, 2018)	-	-	-	-
Human	-	-	82.1	91.2
#1 Ensemble - nlnet	-	-	86.1	91.7
#2 Ensemble - QANet	-	-	84.5	91.5
#1 Single - nlnet	-	-	83.5	90.3
#2 Single - QANet	-	-	82.5	89.3
Published	-	-	-	-
BIDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	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

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
-	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



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	10.8M	5.72 B	78.8	94.4
Xception [9]	299x299	10.8M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299x299	10.8M	13.2 B	80.1	95.1
NASNet-A (7@1920)	299x299	10.8M	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-old		
	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])	—	—	—	62.27	78.42	58.29
SPM+CPN (1)	—	—	—	65.73	63.07	61.00
DRNA (1)	—	—	—	67.77	61.80	59.07
DRNA (2)	—	—	—	68.77	61.80	59.00
DRNA-UNet-UNNC	—	—	—	68.18	64.30	65.20
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	69.87	86.00	46.86	76.74	76.60	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



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.

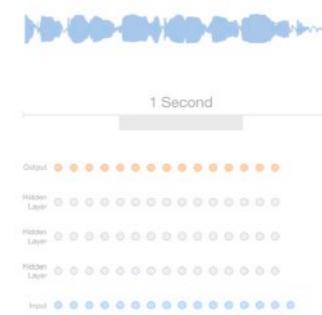
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

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 [60]	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 multiplication and addition. If a model contains 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])	—	—	—	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	—	—	—	68.08	64.30	55.20
Baseline features T+7, single network	42.07	76.20	36.46	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	86.64	64.35

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				
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.

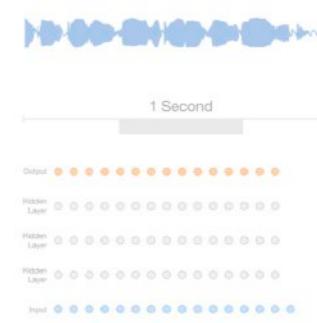
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Speech Recognition

Senone set	Model/combination step	Word Error Rate		WER devset	WER test
		ngram-LM	LSTM-LMs		
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)



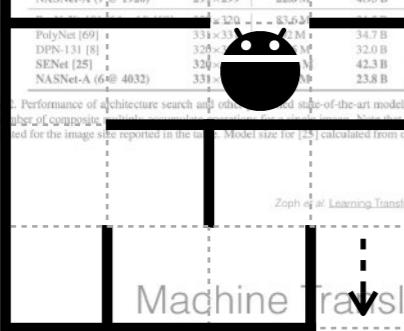
Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

Supervised Learning

Image Classification

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Jarrett et al. [27]	224x224	11.0M ^a	1.0E+03	74.8	92.2
NASNet-A (5@ 15.38)	299x299	10.9M ^b	2.35 B	78.6	94.2
Inception V3 [60]	299x299	23.8M ^b	5.72 B	78.8	94.4
Xception [9]	299x299	22.8M ^b	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299x299	55.8M ^b	13.2 B	80.1	95.1
NASNet-A (7@ 1920)	299x299	22.6M ^b	4.93 B	80.8	95.3
PolyNet [69]	331x331	1.6M ^b	83.6 B	80.9	95.6
DPN-131 [8]	329x329	1.6M ^b	32.0 B	81.5	95.7
SENet [25]	320x320	1.6M ^b	42.3 B	82.7	95.8
NASNet-A (6@ 4032)	331x331	1.6M ^b	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 that composite multiple accuracy operations are calculated for the image size reported in the paper. Model size for 145 is calculated from open-source implementation.



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

BERT_{BASE} (Single) 80.8 88.5 - -

BERT_{LARGE} (Single) 84.1 90.9 - -

BERT_{LARGE} (Ensemble) 85.8 91.8 - -

BERT_{LARGE} (Sgl.+TriviaQAn) 84.2 91.1 85.1 91.8

BERT_{LARGE} (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

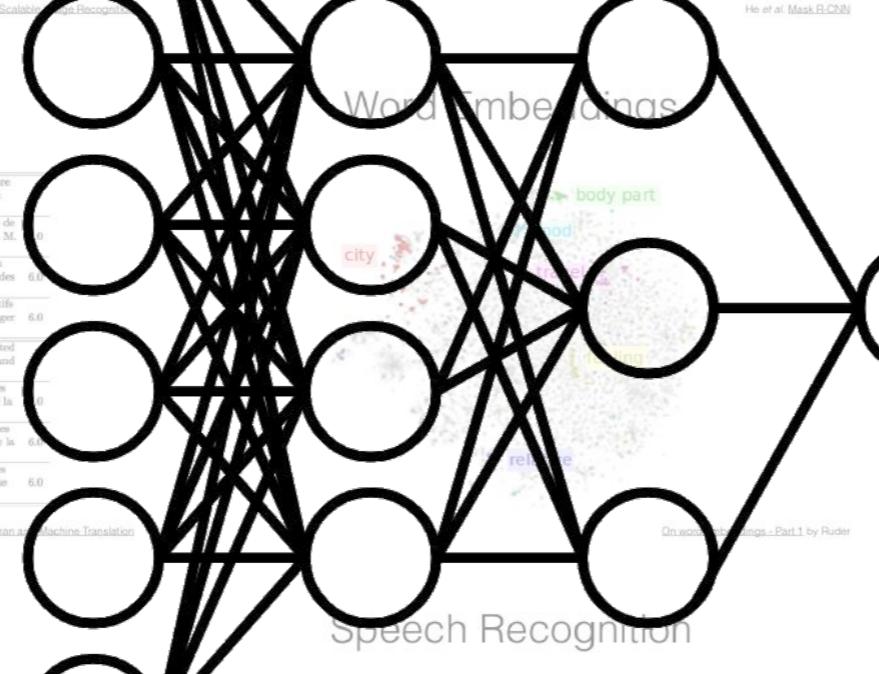
Instance Segmentation



Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

He et al. Mask R-CNN

Word Embeddings



Onward travelings - Part 1 by Ruder

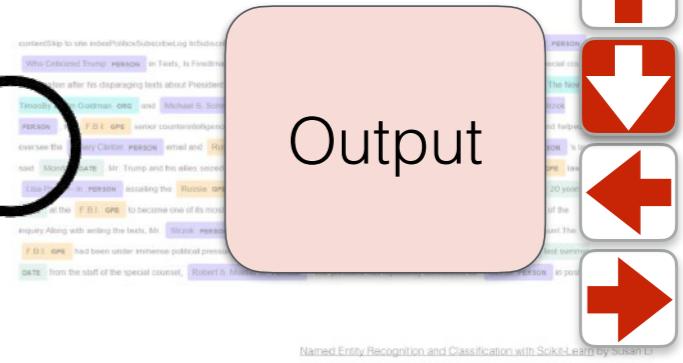
Visual Question Answering

Method	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual Nodes	Others	All	Visual Nodes	Others
Other cross-domain models in training sets	-	-	-	25.68	30.40	31.77
LETA Language only model	-	-	-	44.26	47.40	51.85
Deep LSTM Q-vqa 1.07 (as reported in [1])	-	-	-	54.22	73.08	59.18
MCB 1.1 (as reported in [1])	-	-	-	42.27	78.42	38.29
SPMC-LPN [1]	-	-	-	65.73	63.07	61.06
DRQA	-	-	-	47.77	61.80	41.47
RCNN-UNITER-UNIC	-	-	-	58.77	61.86	59.39
Proposed model	-	-	-	68.18	84.39	65.20

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

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

Named Entity Recognition

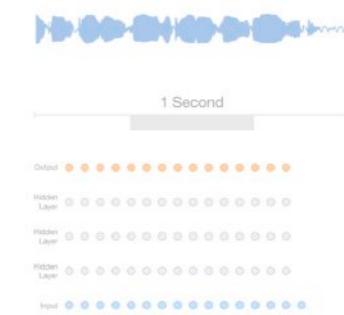


Named Entity Recognition and Classification with Scikit-learn by Susan Li Esteves et al. Named Entity Recognition in Twitter using Images and Text

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
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



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 [7]	299×299	10.9 M ^b	2.35 B	74.8	92.2
NASNet-A (5 ^a @ 15.38)	299×299	10.9 M ^a	78.6	94.2	
Inception V3 [50]	299×299	23.8 M ^a	5.72 B	78.8	94.4
Xception [9]	299×299	22.8 M ^a	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M ^a	13.2 B	80.1	95.1
NASNet-A (7 ^a @ 1920)	299×299	22.6 M ^a	4.93 B	80.8	95.3
PolyNet [69]	331×331	8.3 M ^a	80.9	95.6	
DPN-131 [8]	329×329	32.0 M ^a	81.5	95.7	
SENet [25]	320×320	4.3 M ^a	42.3 B	82.7	95.8
NASNet-A (6 ^a @ 4032)	331×331	8.3 M ^a	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 that composite multiple accuracy operations are calculated for the image size reported in the paper. Model size for 145^a 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).

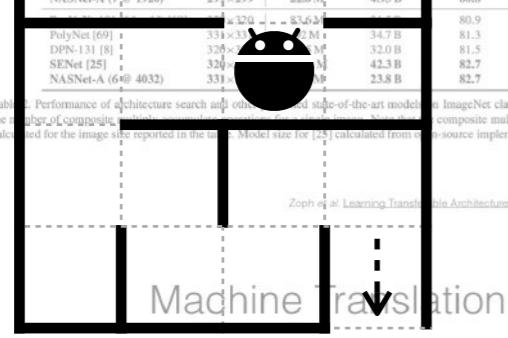
Zoph et al. Learning Transferable Architectures for Scalable Image Recognition

Visual Question Answering

Method	VQA v1 test-Set			VQA v2 test-Set		
	All	Visual	Nodes	All	Visual	Nodes
Other cross-domain models in training set (1)	—	—	—	25.08	36.48	37.17
LSTM Language only (old model) (1)	—	—	—	44.26	47.40	55.85
Deep LSTM Q (ours, 1.27x improved 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	59.59
RCNN-UNet-UNNC	—	—	—	68.18	64.39	65.20
Proposed model	—	—	—	70.01	70.20	70.21
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	64.38	63.38	43.17	57.41	64.71	67.29
Image features from bottom-up attention, adaptive K, ensemble	69.87	86.00	46.89	76.74	76.64	81.51

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

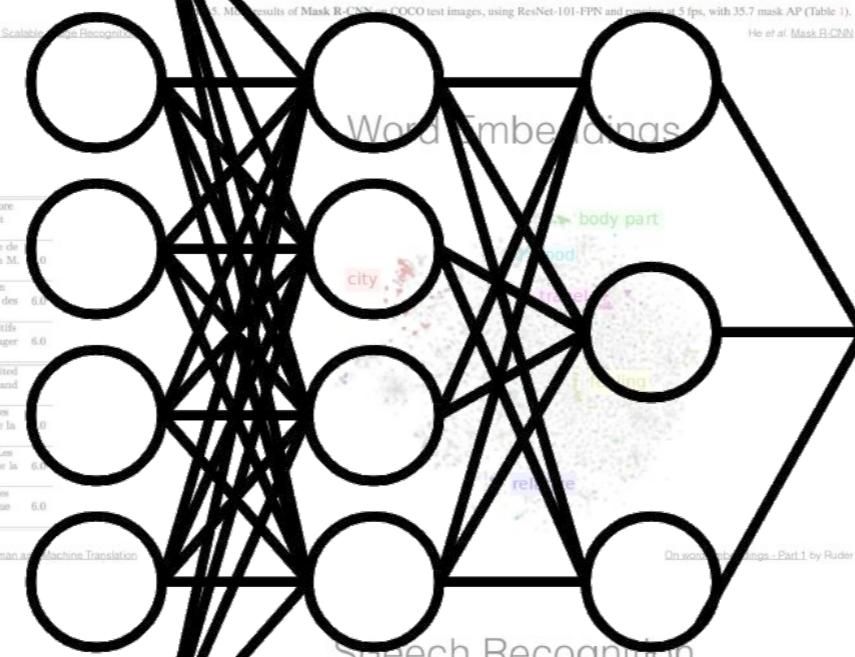
Question Answering

System	Dev EM	Test F1	EM	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	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours	80.8	88.5	-	-
BERT _{BASE} (Single)	80.8	88.5	-	-
BERT _{BASE} (Ensemble)	81.1	90.0	-	-
BERT _{LARGE} (Single)	85.8	91.8	-	-
BERT _{LARGE} (Sgl+TriviaQA)	84.2	91.1	85.1	91.8
BERT _{LARGE} (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

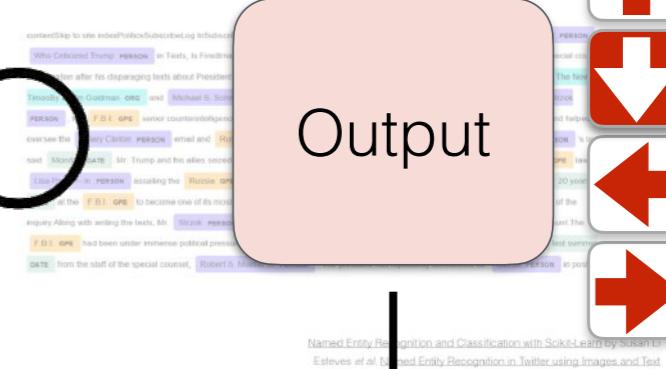


Speech Recognition

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
Our model (single senone set)			
+ LSTM rescoring			
+ n-gram rescoring			
+ backchannel penalty			

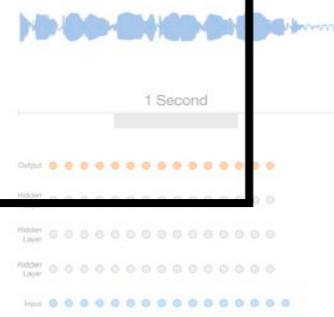
Xiong et al. The Microsoft 2017 Conversational Speech Recognition System

Named Entity Recognition



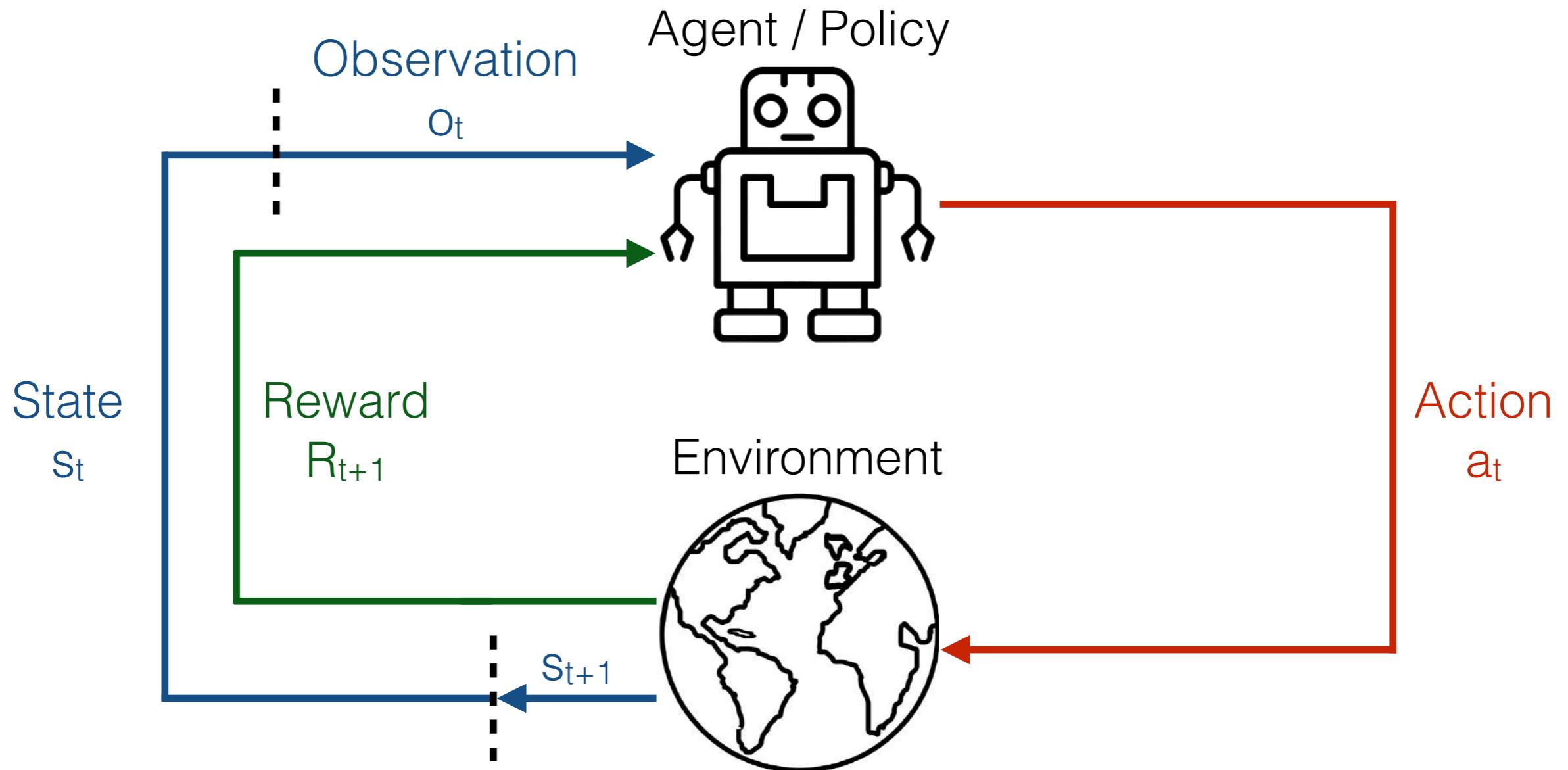
Named Entity Recognition and Classification with Scikit-learn by Susan Li Esteves et al. Named Entity Recognition in Twitter using Images and Text

Speech Synthesis (text-to-speech)



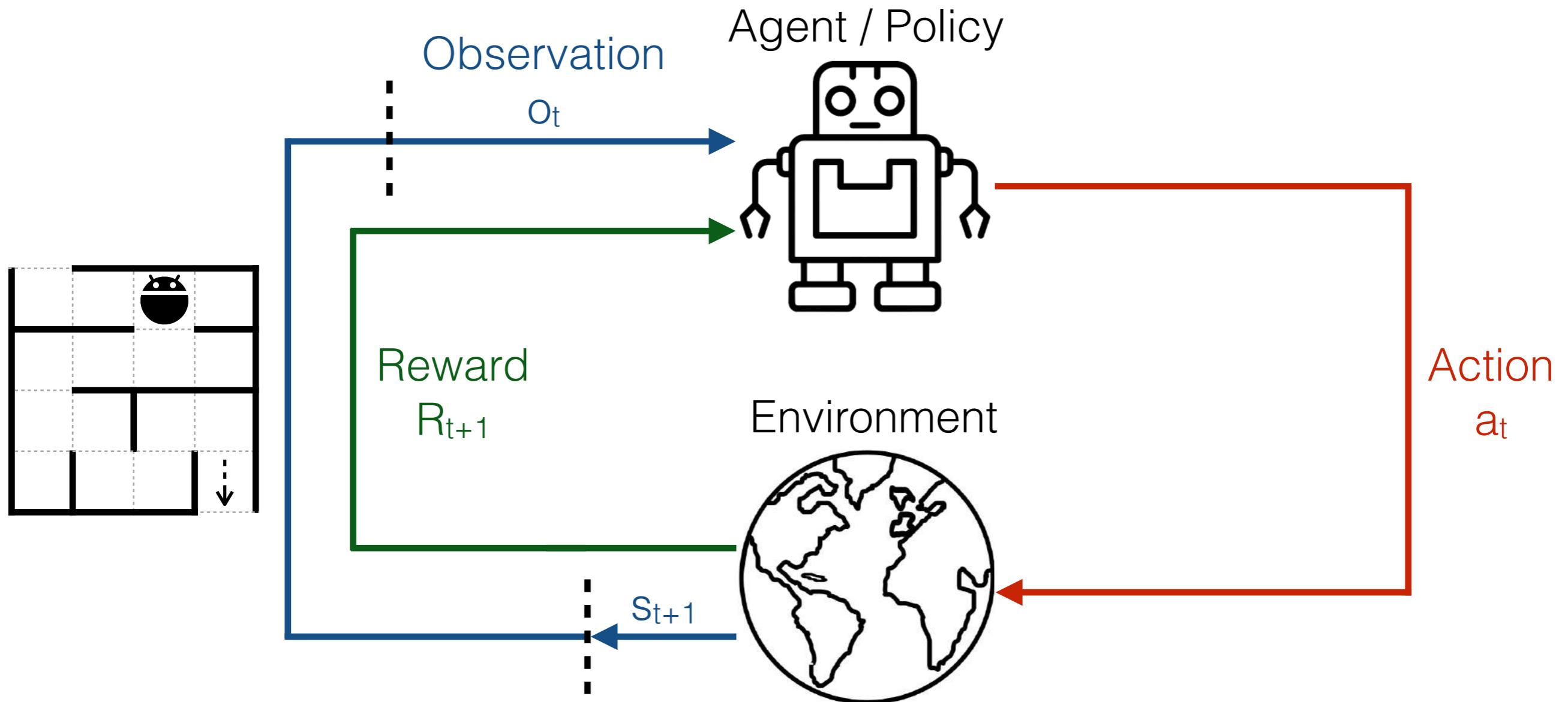
Van Den Oord et al. WaveNet: A Generative Model for Raw Audio

Robot Learning via Reinforcement Learning



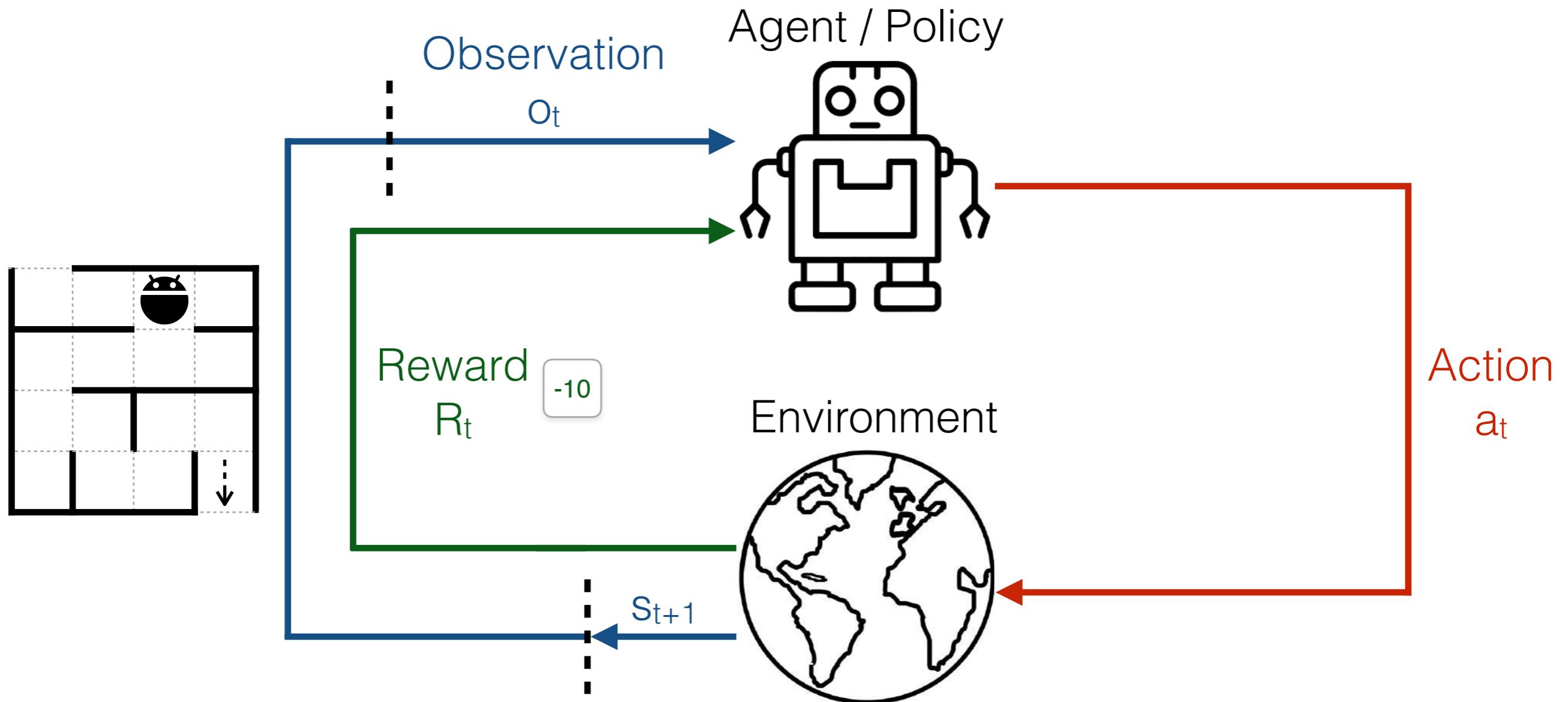
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



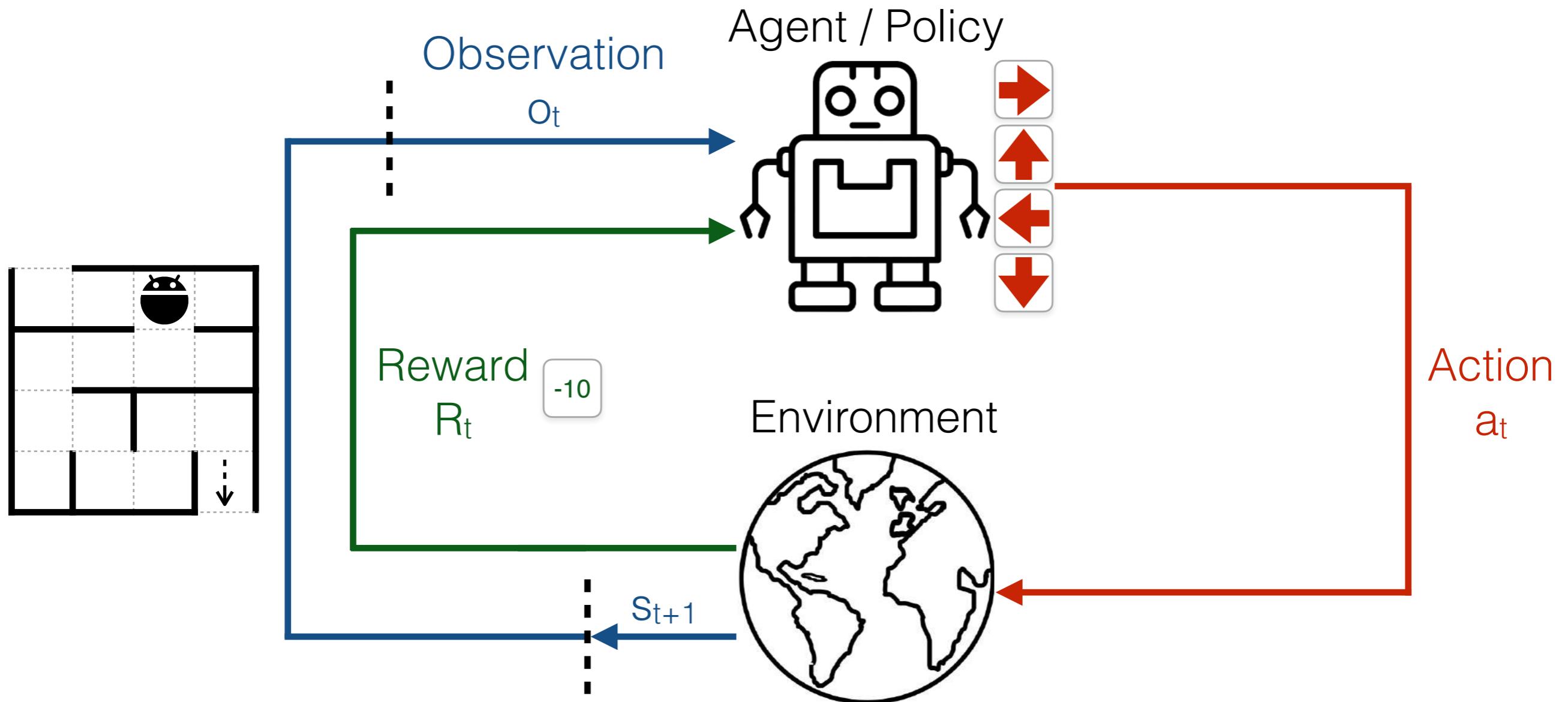
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



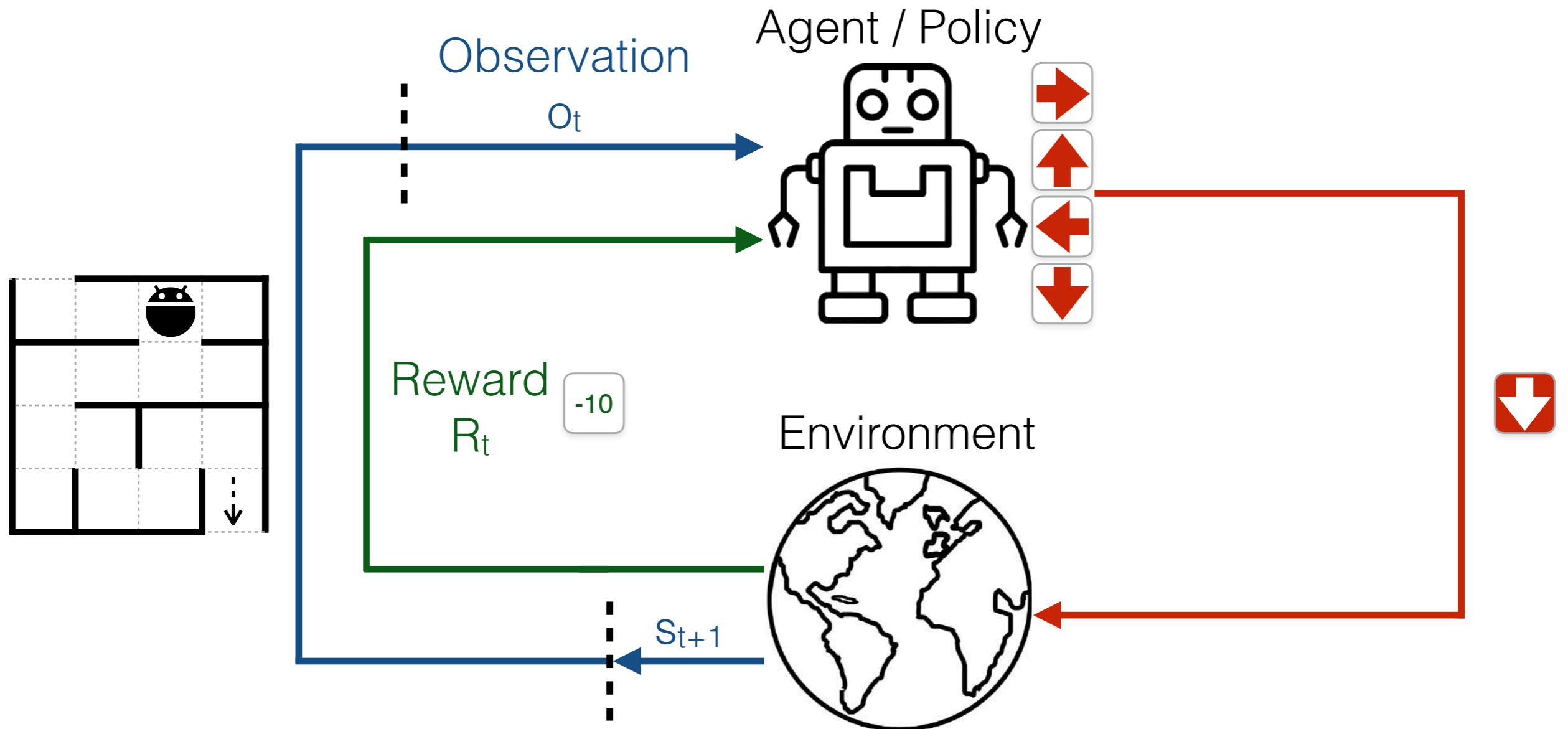
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



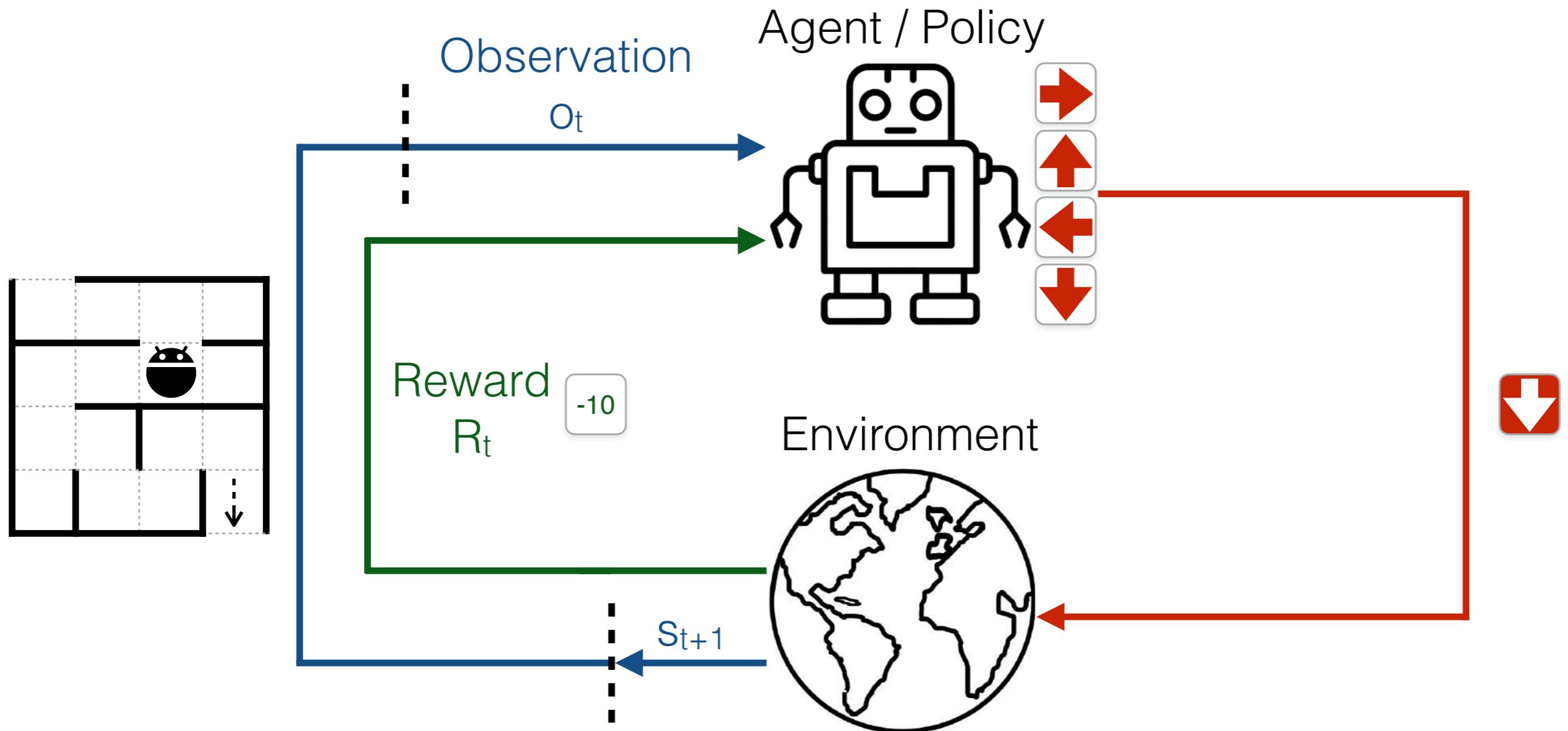
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



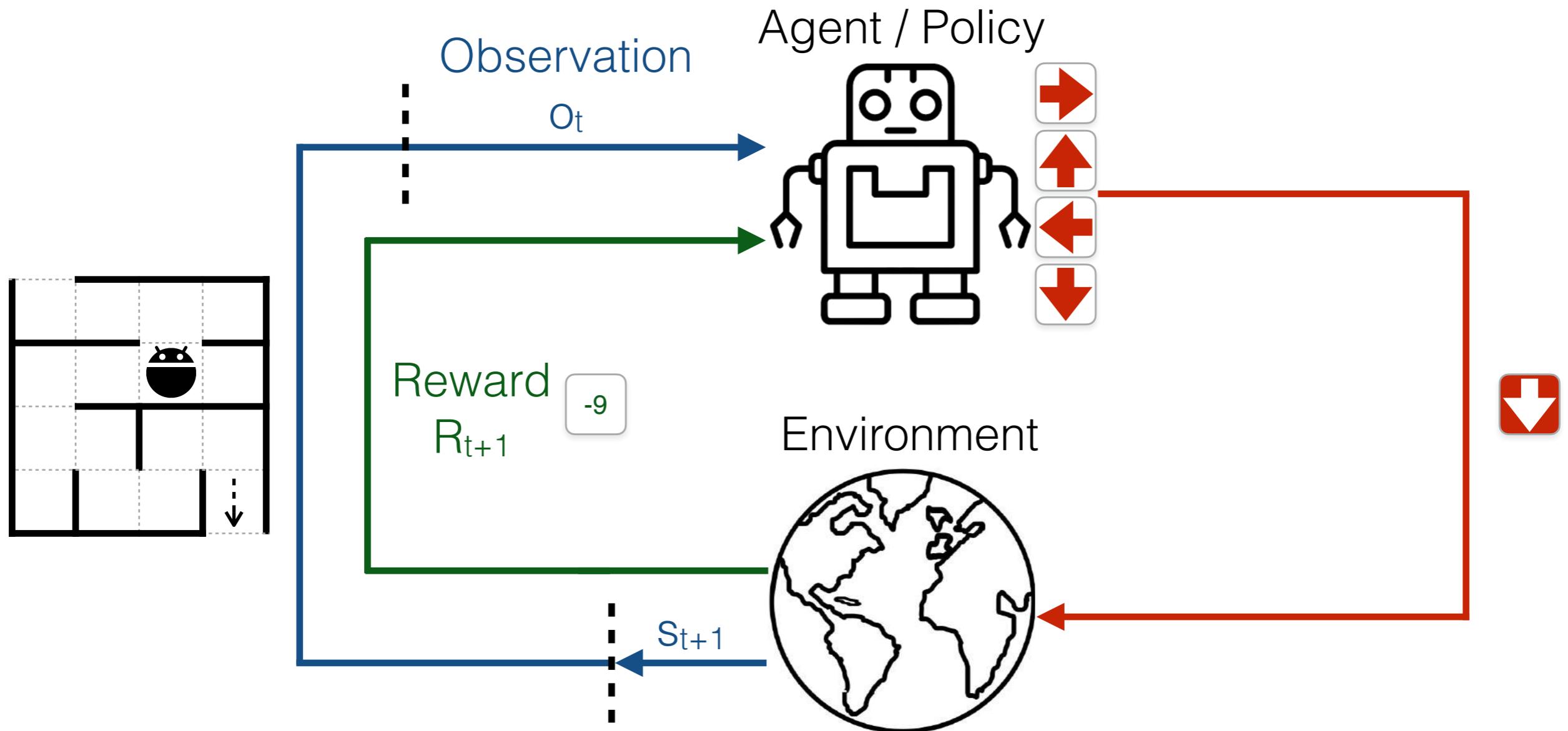
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



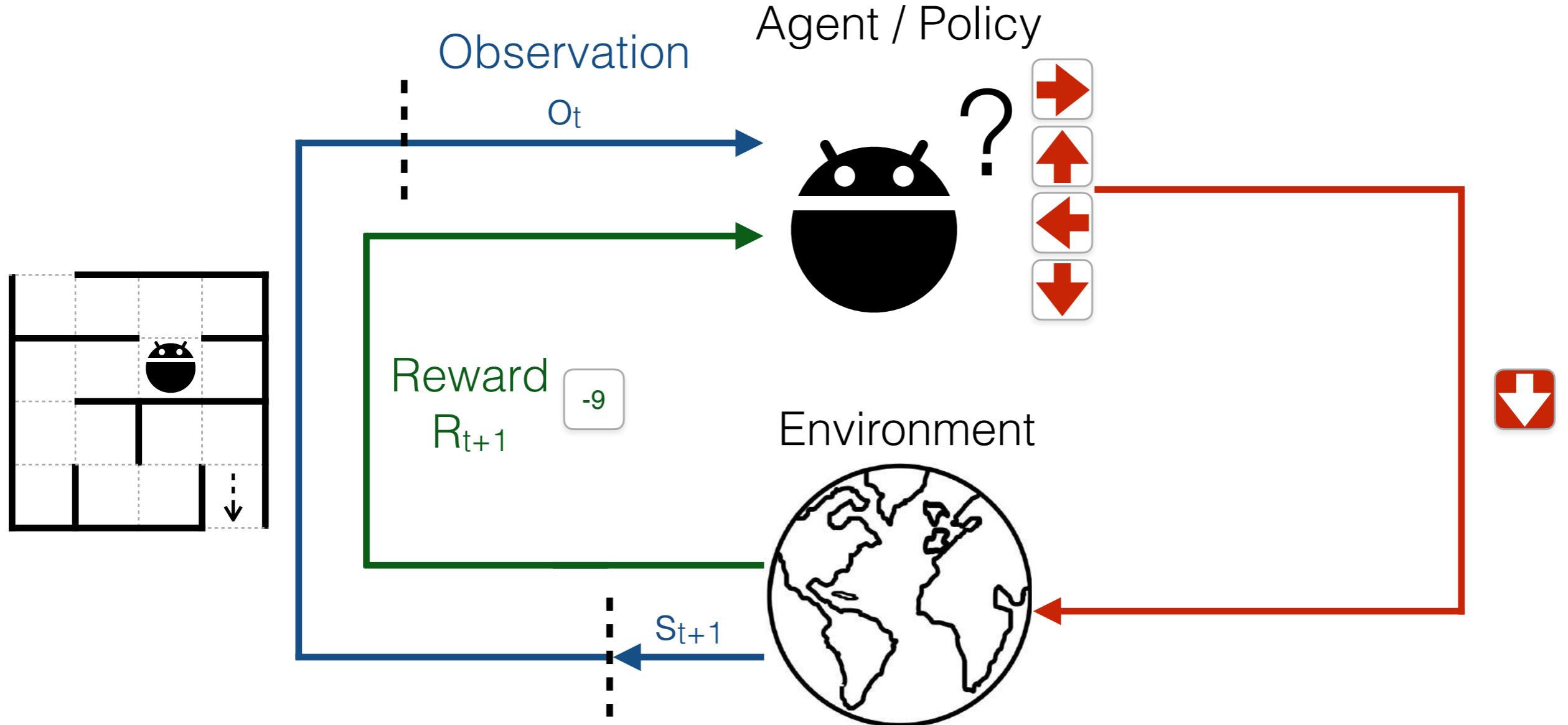
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



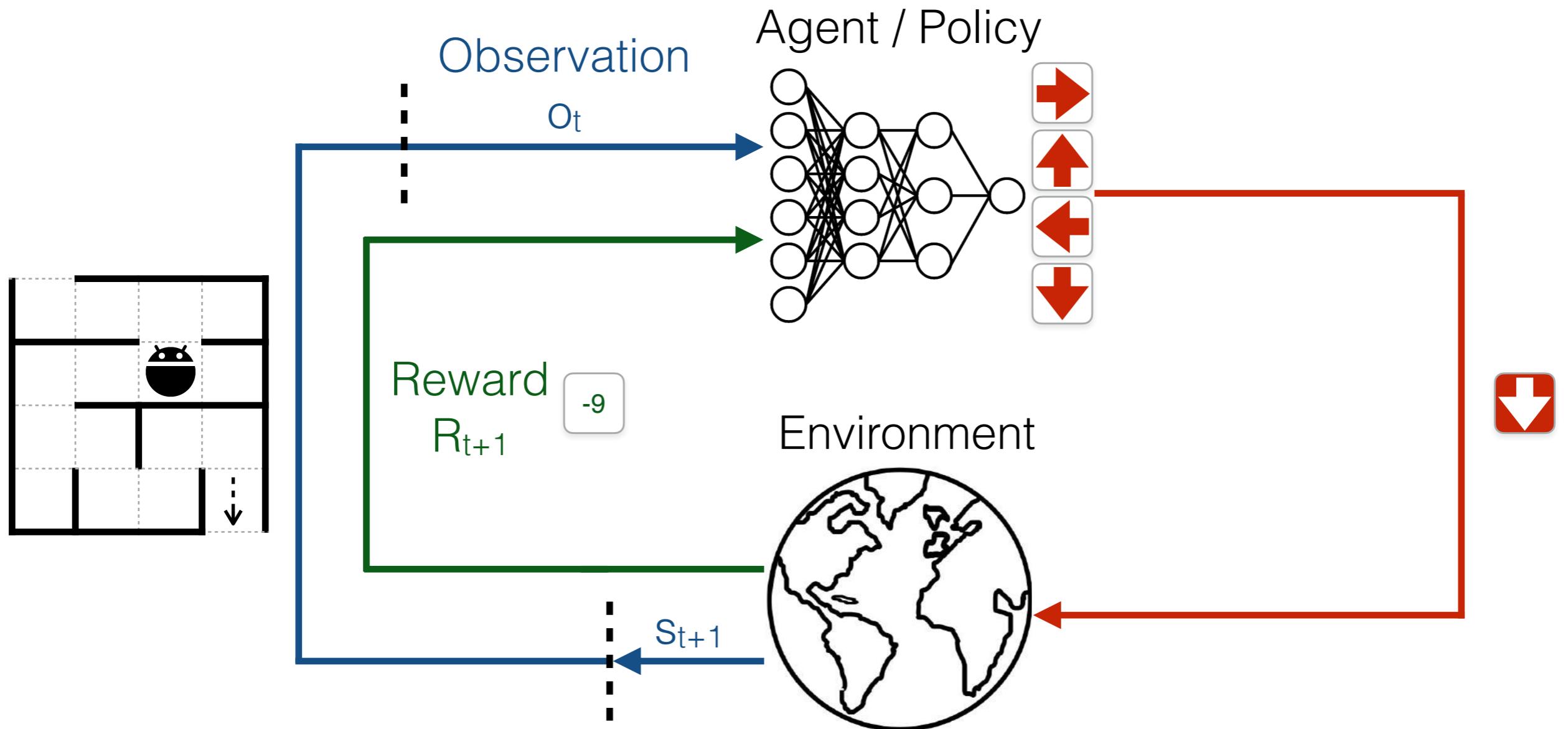
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



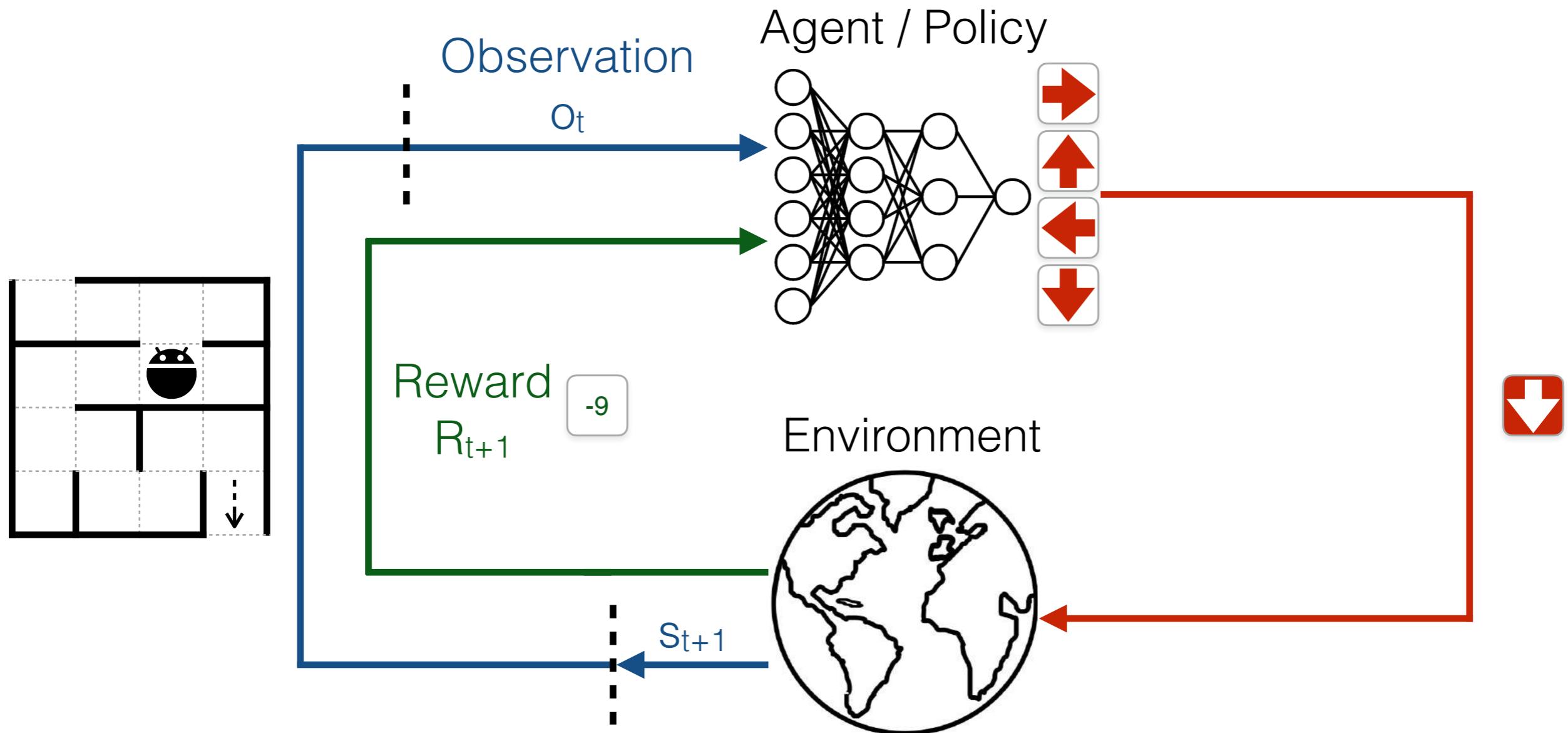
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

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



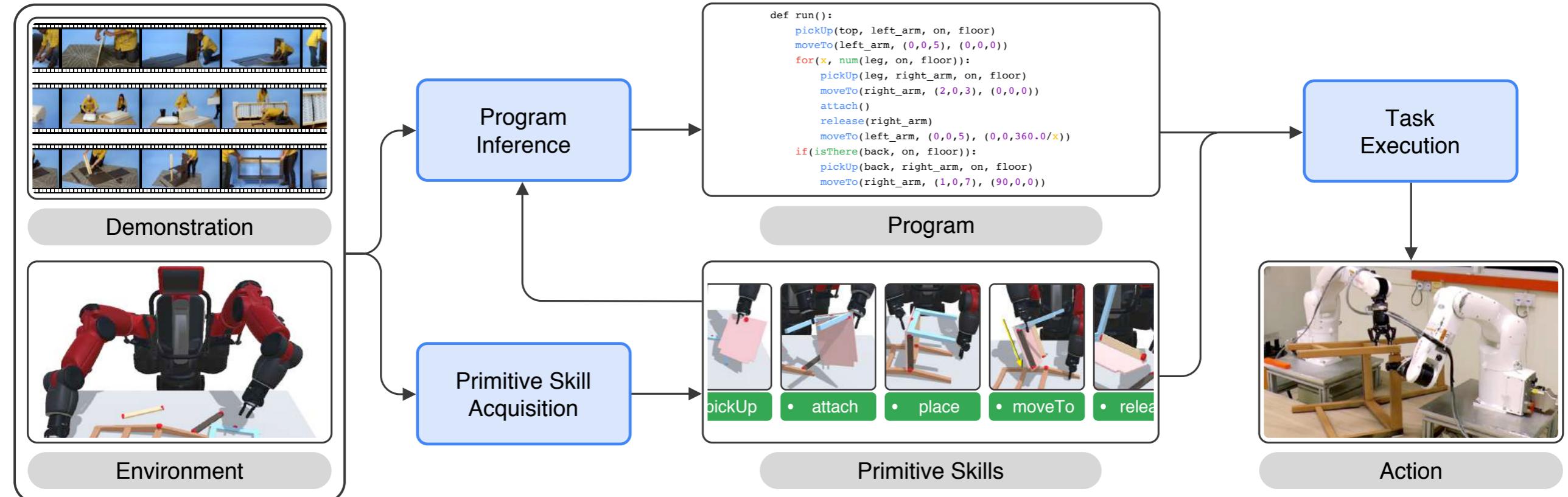
Uninterpretable

Not generalizable

Limited to short-horizon tasks

No skill-reuse

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



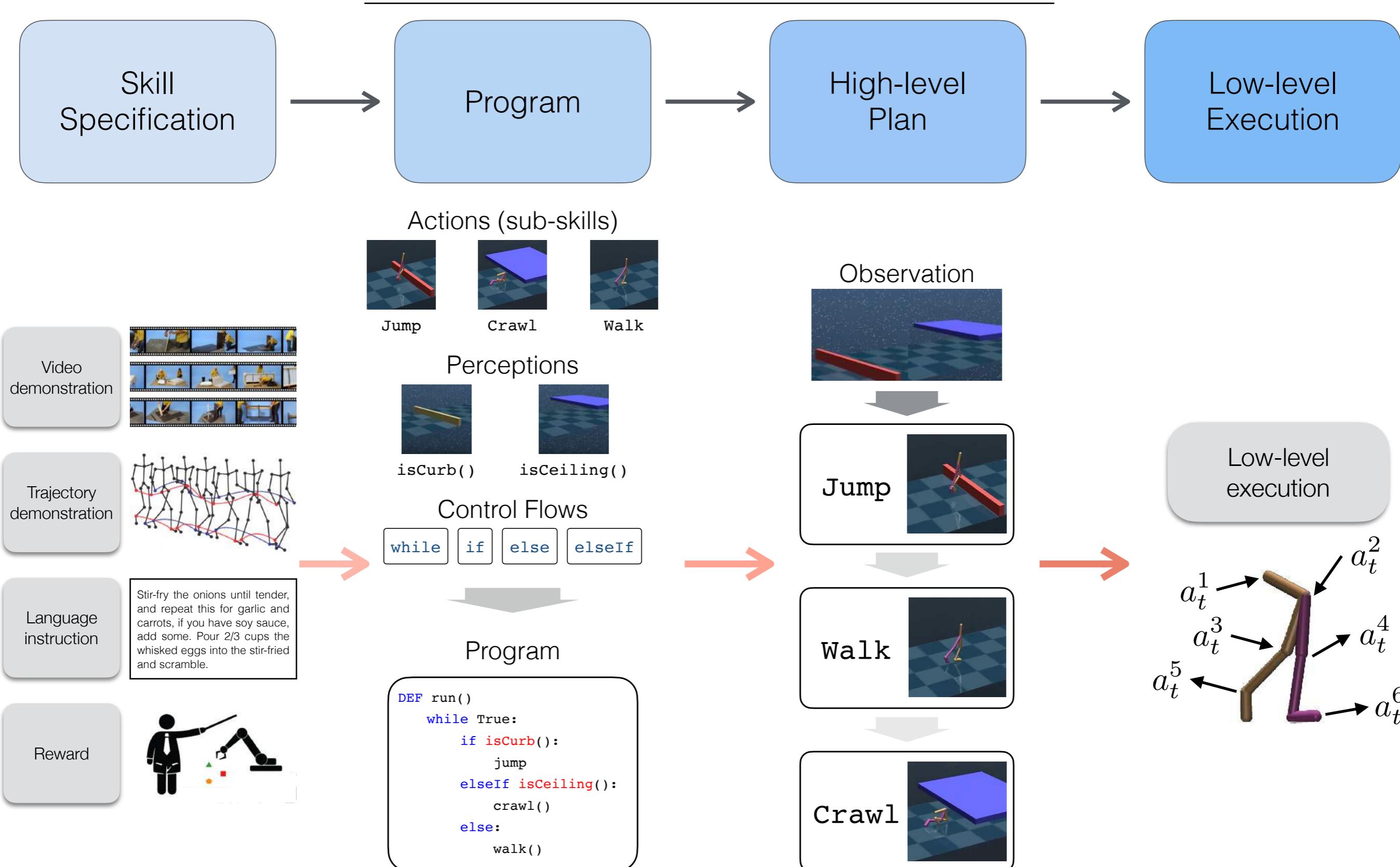
Interpretable

Programmatic / Generalizable

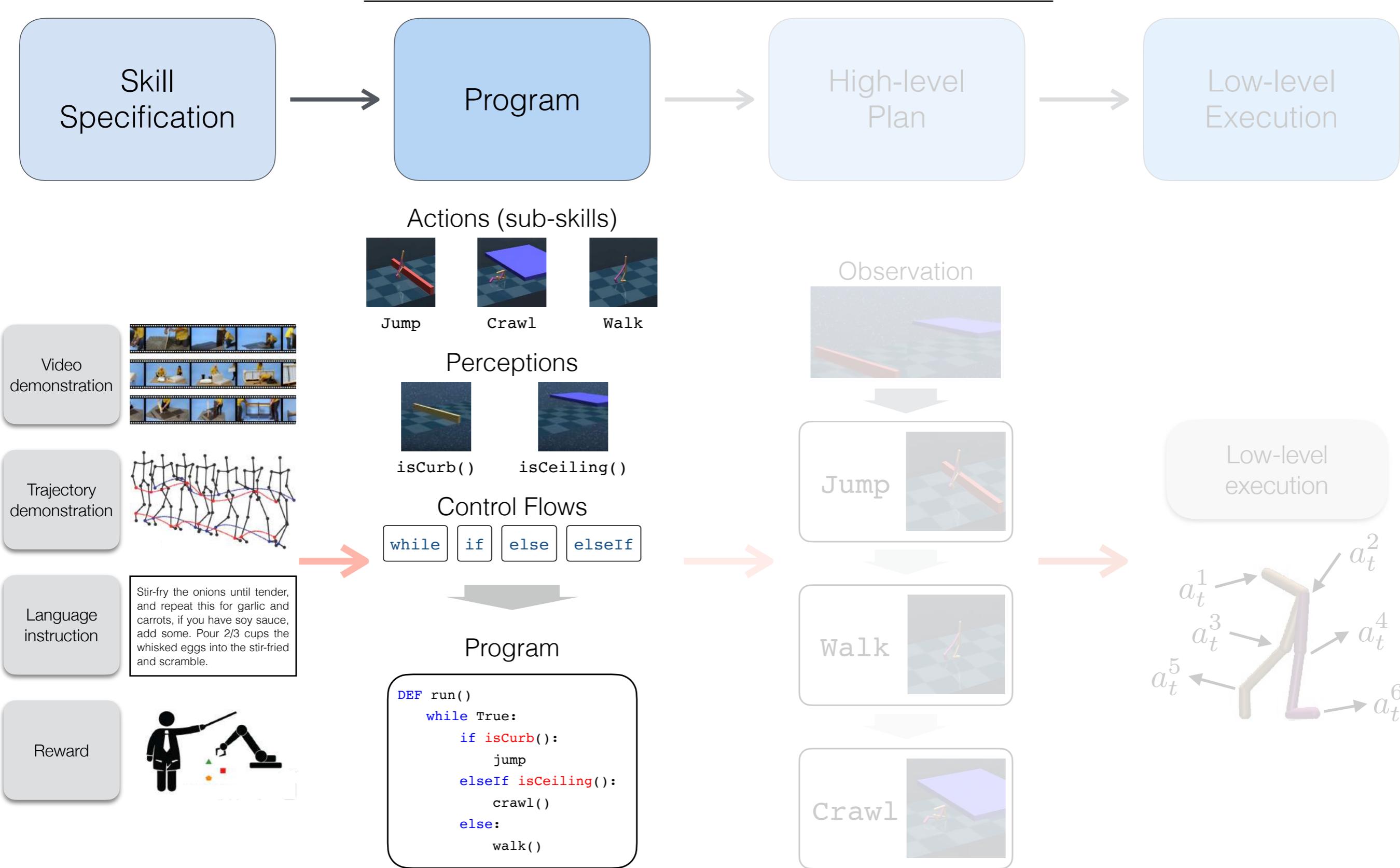
Hierarchical

Modular

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

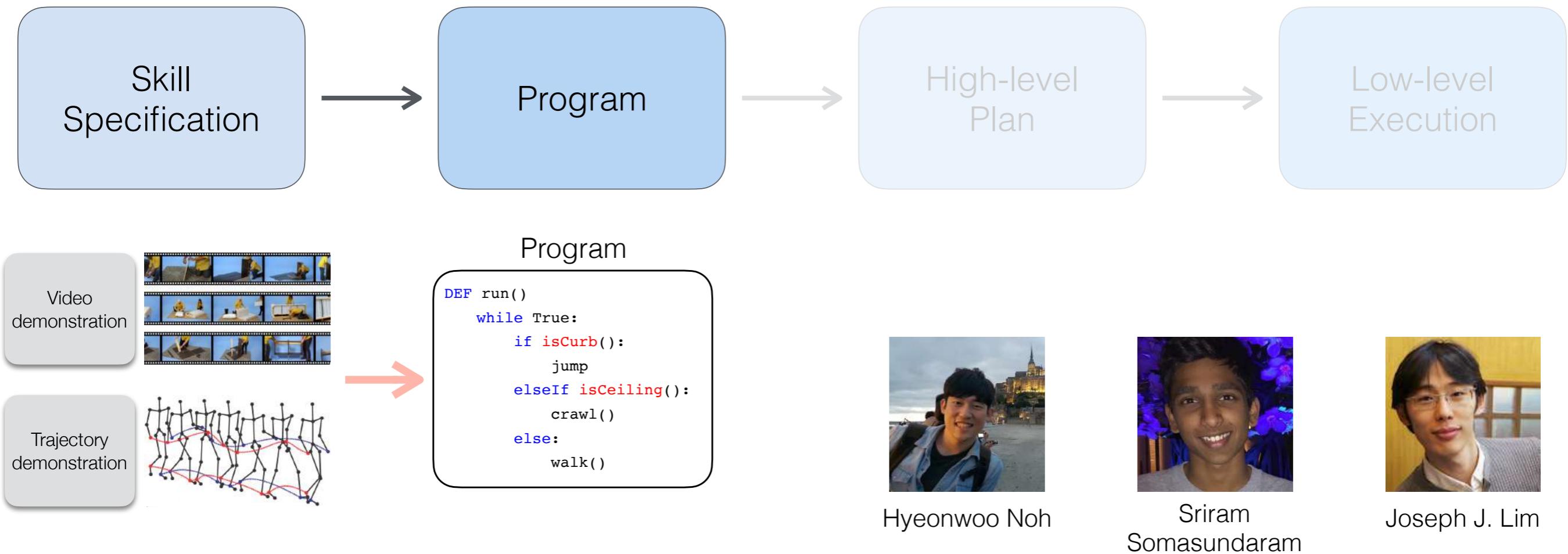


Program Inference



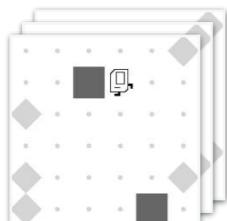
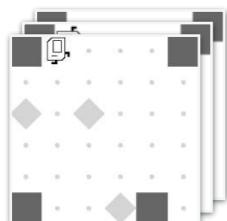
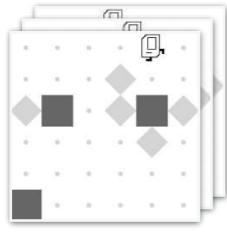
Neural Program Synthesis from Diverse Demonstration Videos

ICML 2018

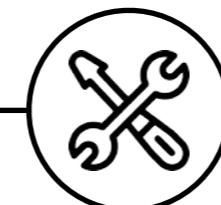


Imitation Learning

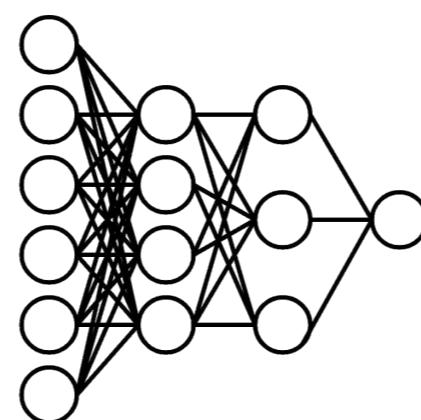
Demonstrations



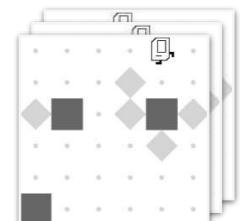
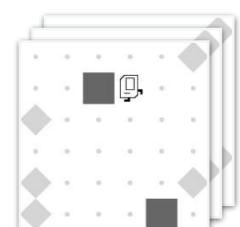
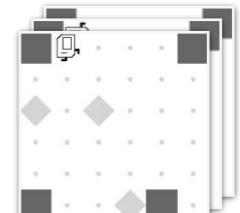
Imitate



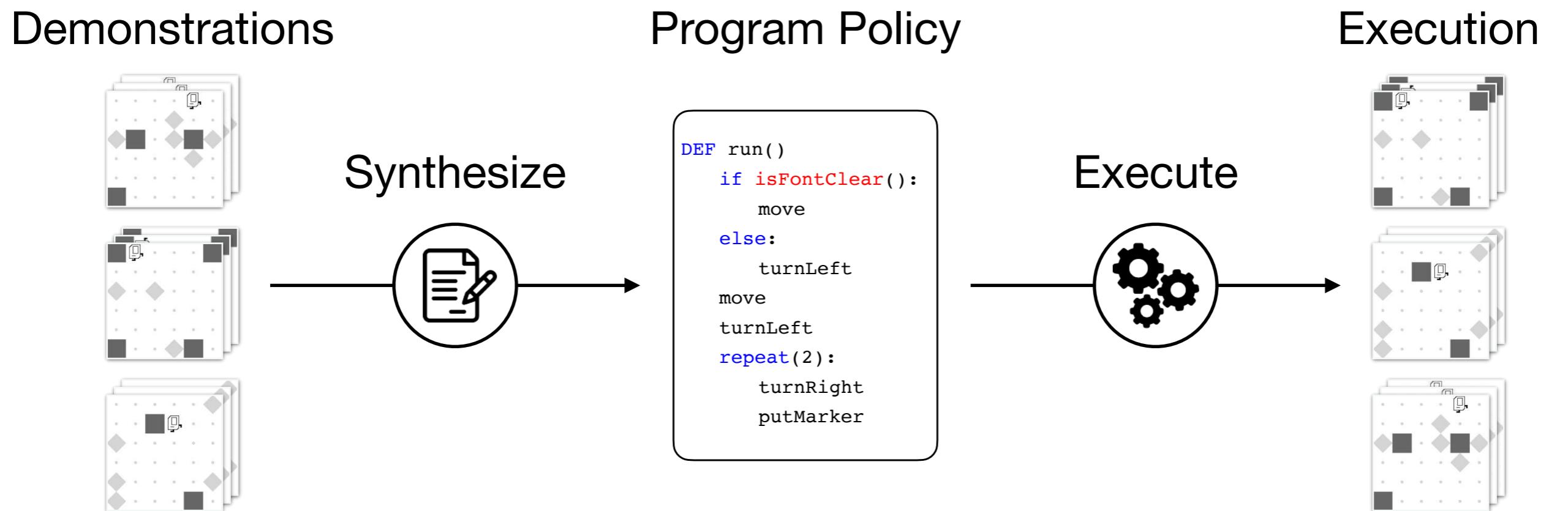
Neural Network
Policy



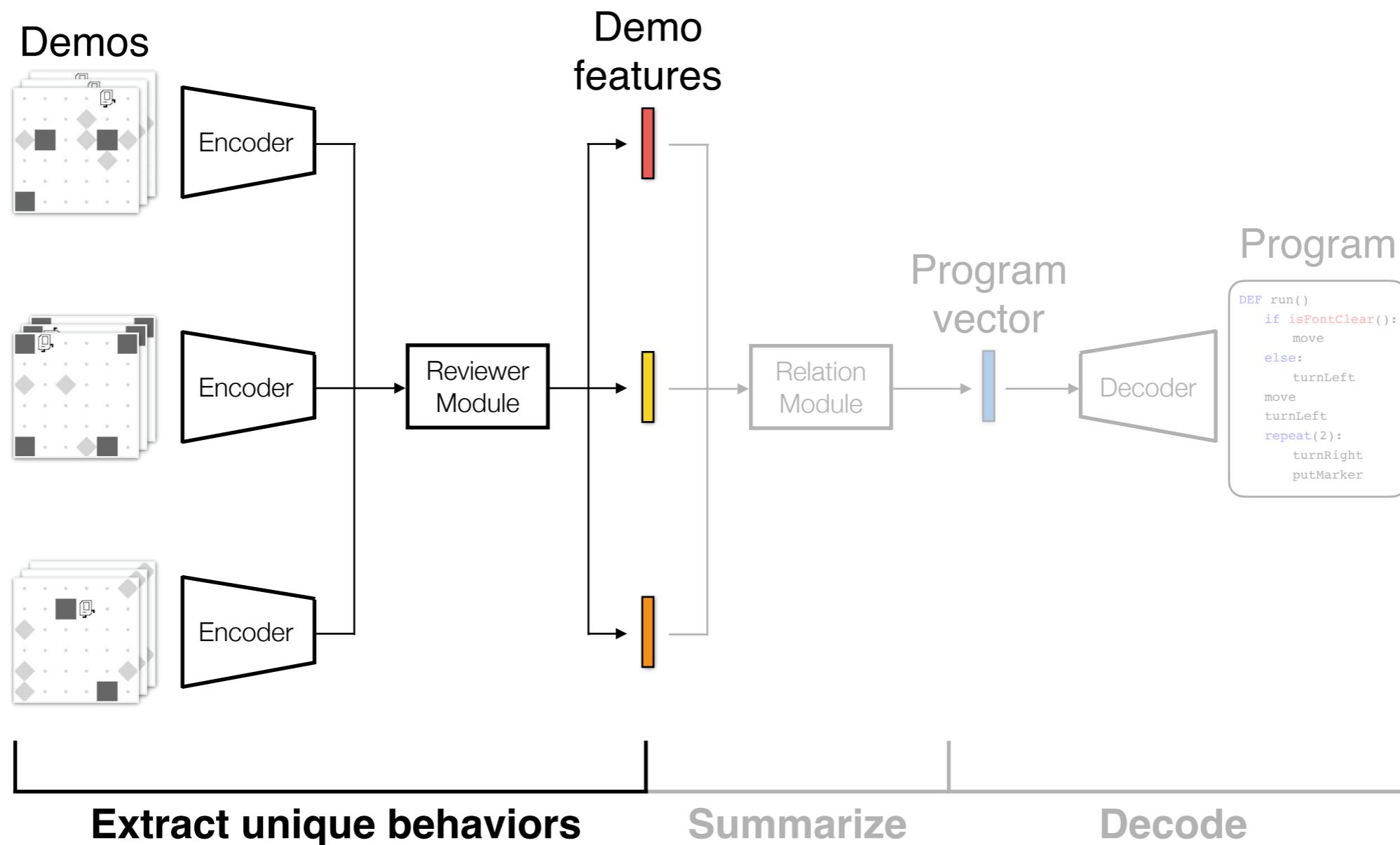
Execution



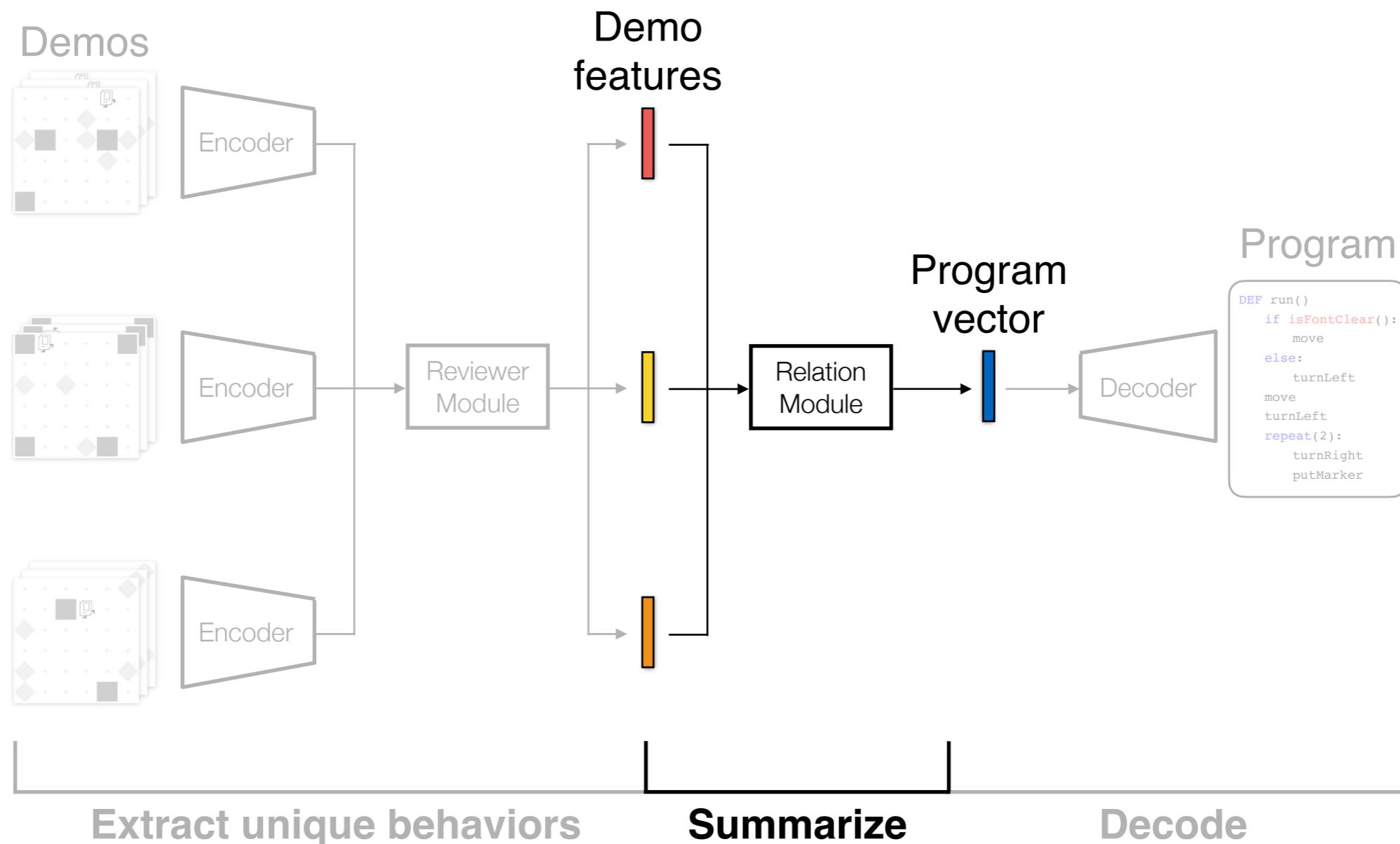
Imitation Learning by Synthesizing Programs



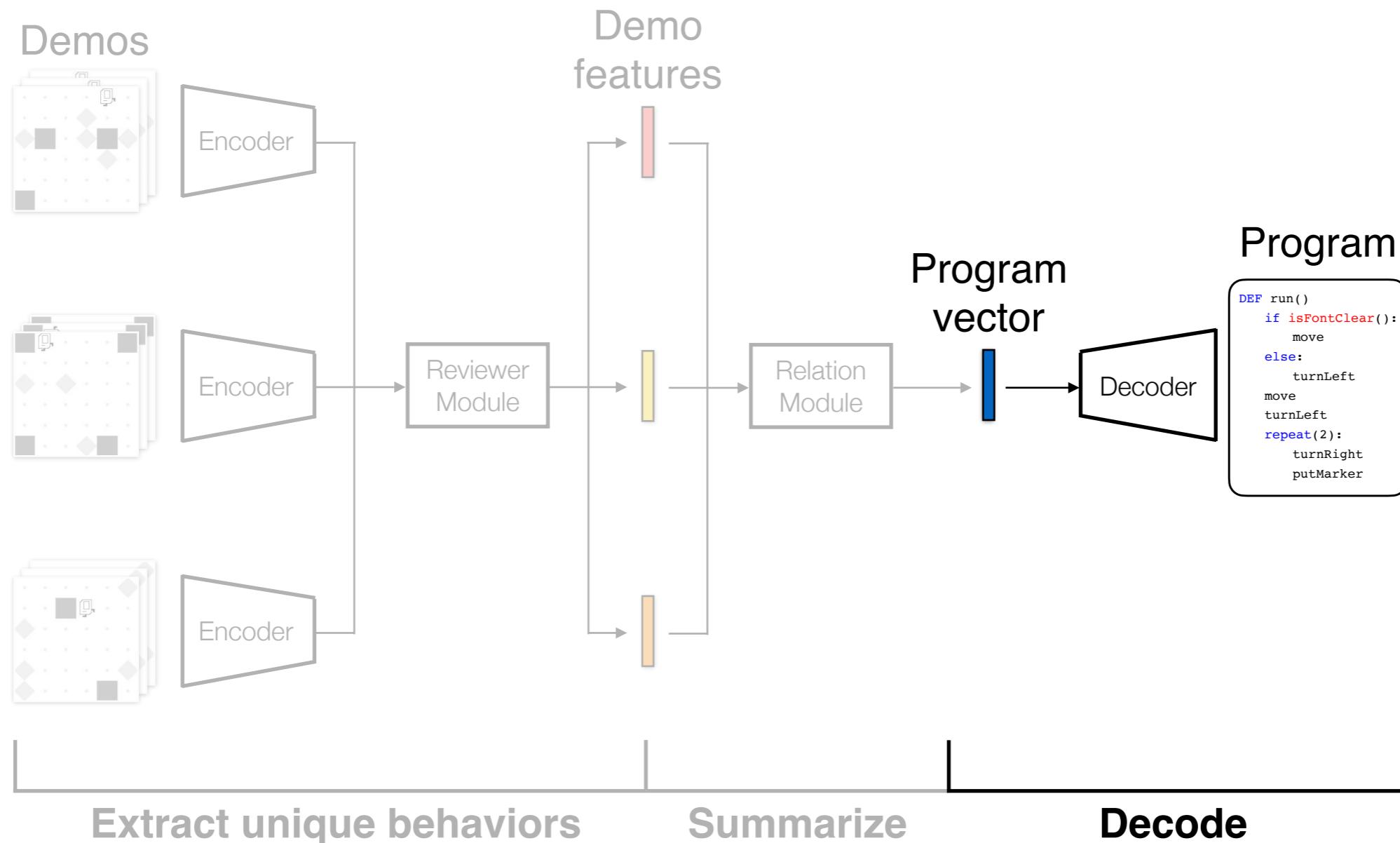
Model Overview



Model Overview



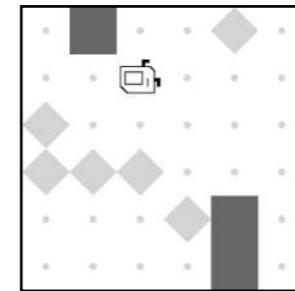
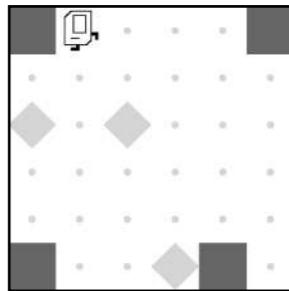
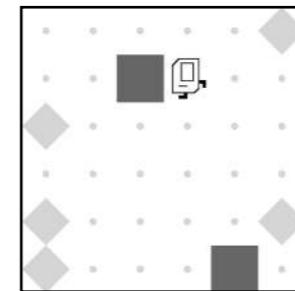
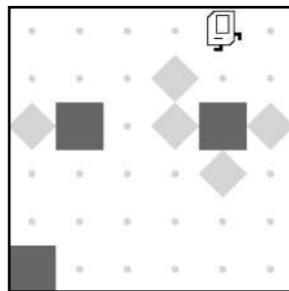
Model Overview



Environments

Karel

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



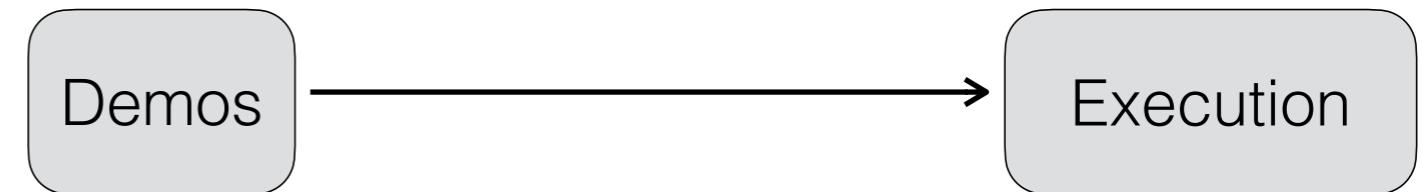
ViZDoom

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

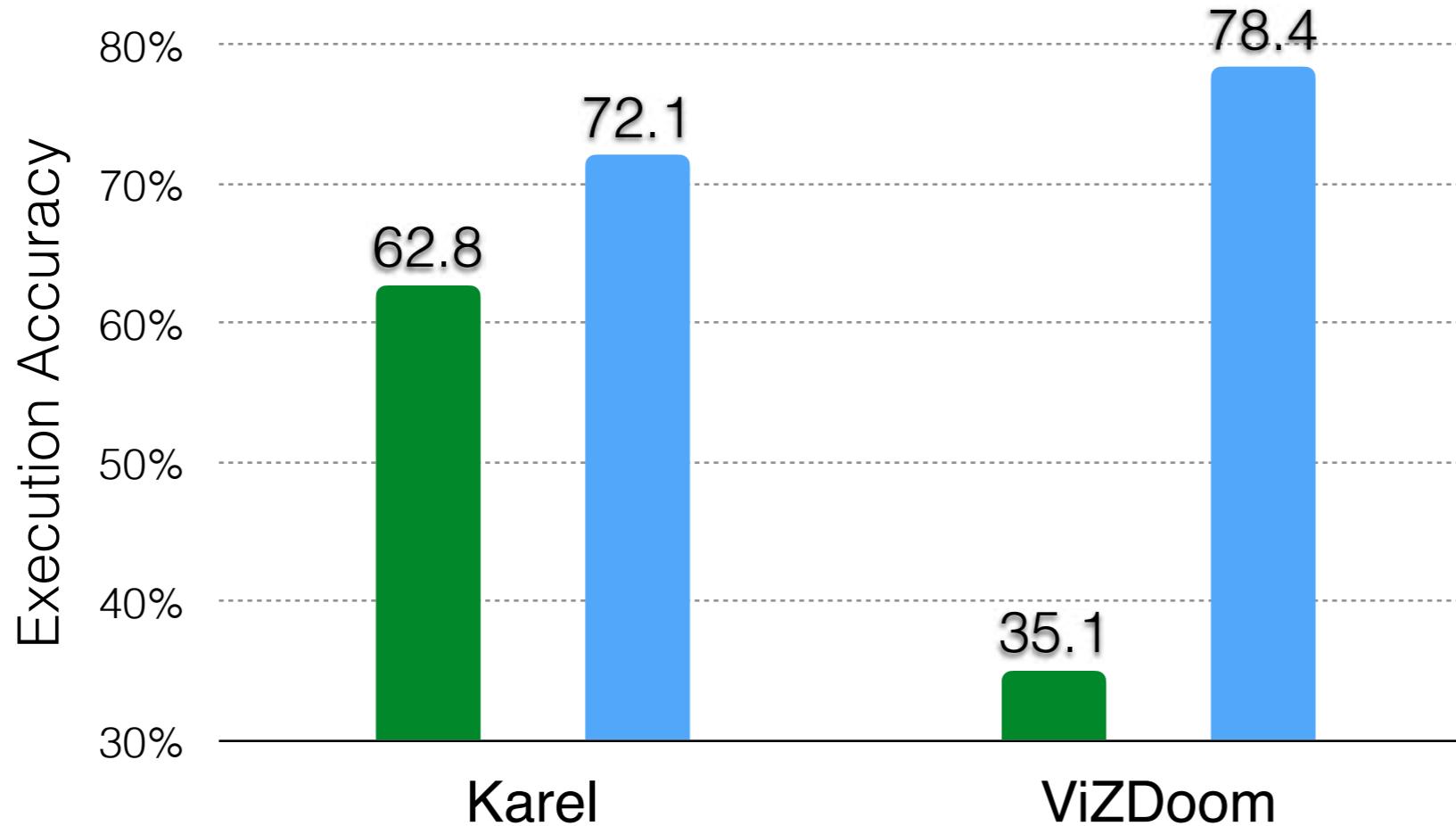
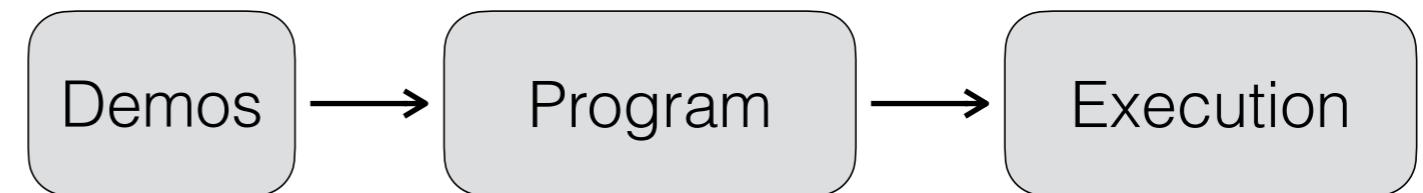


Quantitative Results

Neural Network Policy



Program Policy

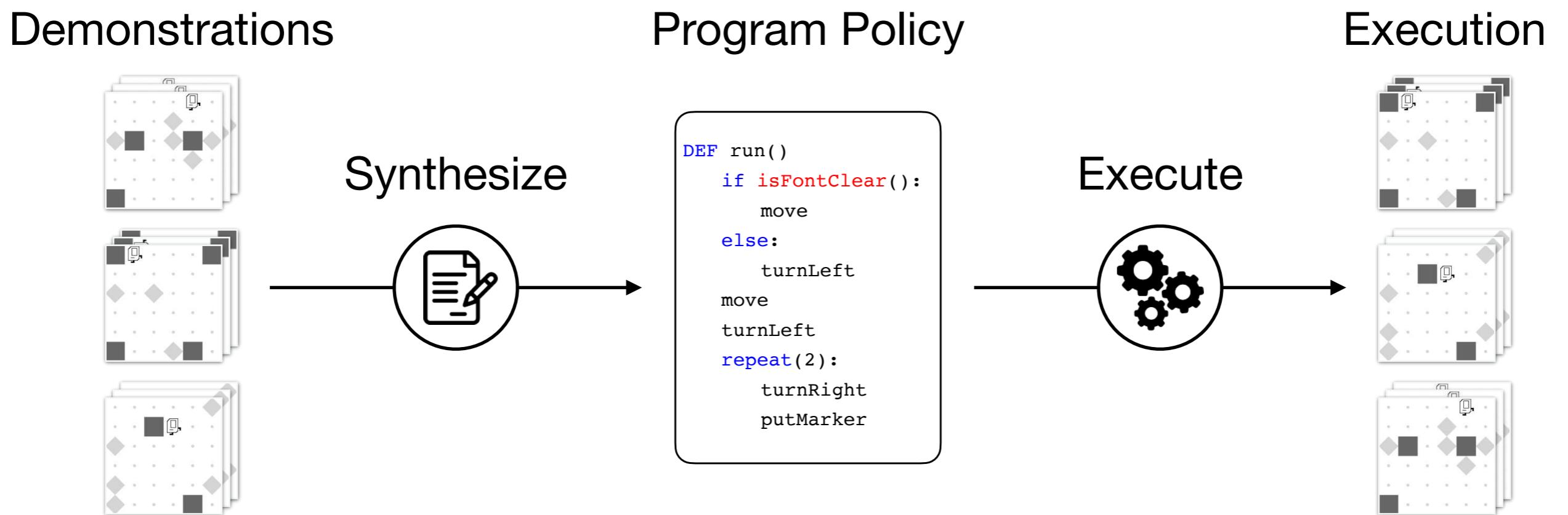


```
DEF run()
    if isFontClear():
        move
    else:
        turnLeft
    move
    turnLeft
    repeat(2):
        turnRight
        putMarker
```

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

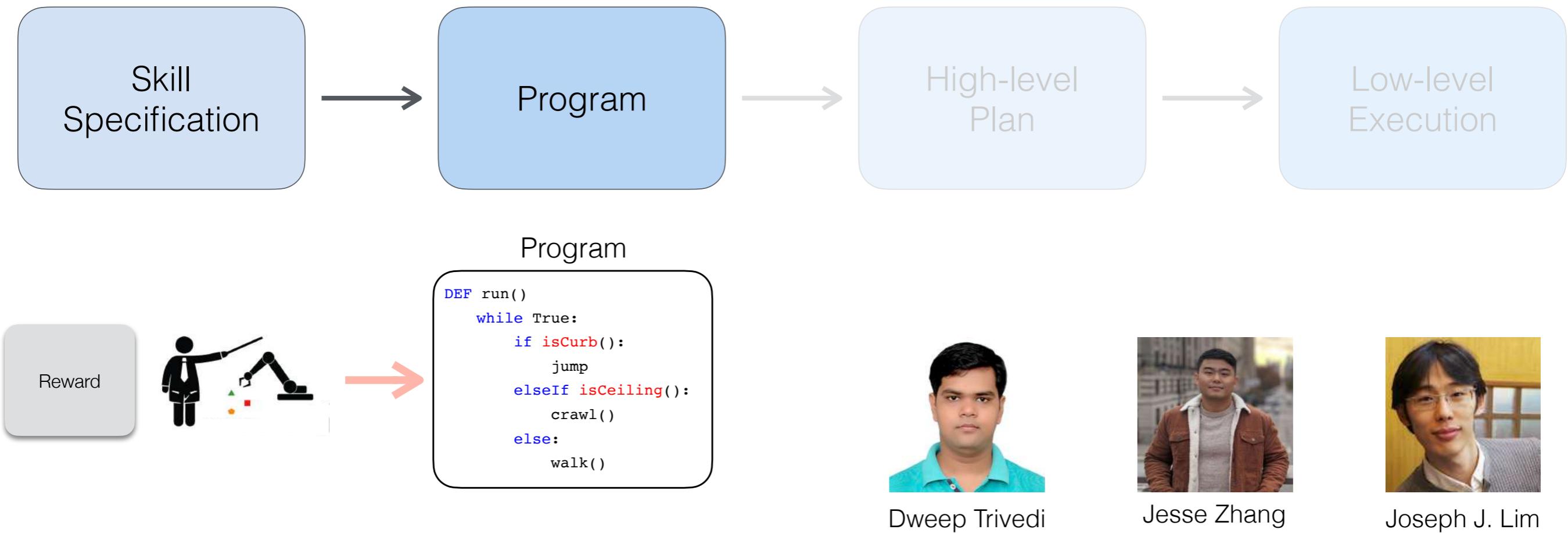
Takeaway

- Synthesize programs to imitate demonstrations

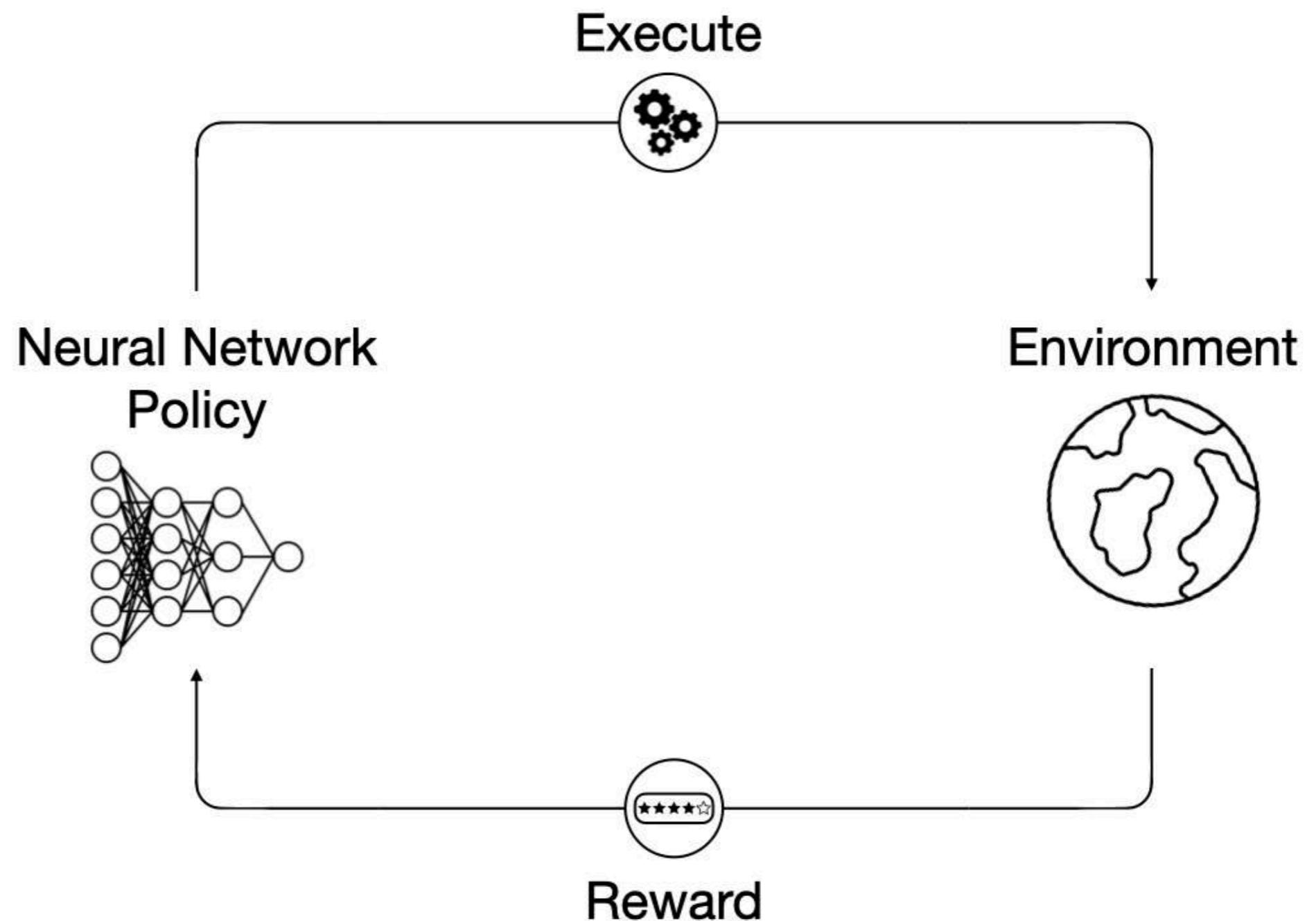


Learning to Synthesize Programs as Interpretable and Generalizable Policies

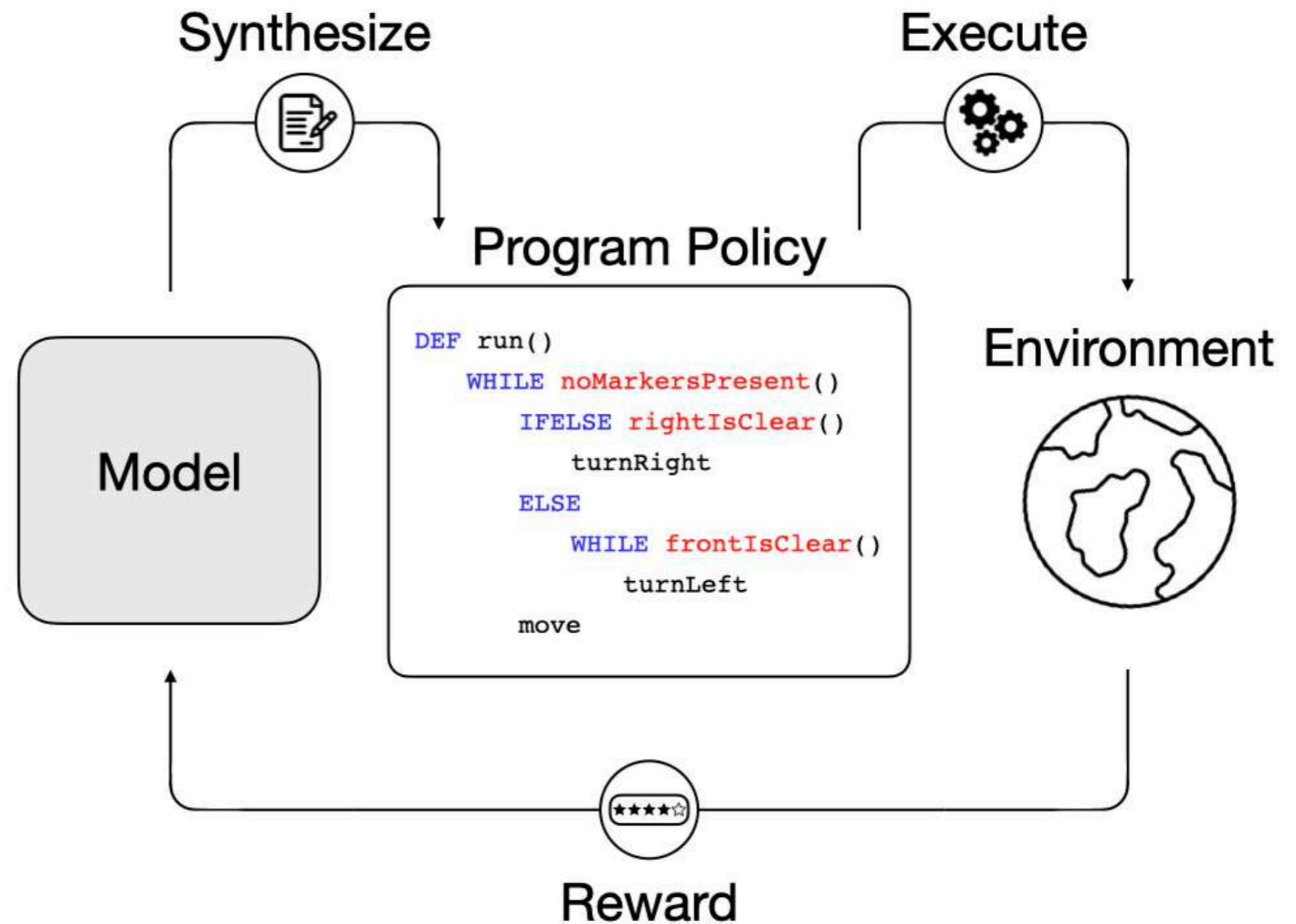
NeurIPS 2021



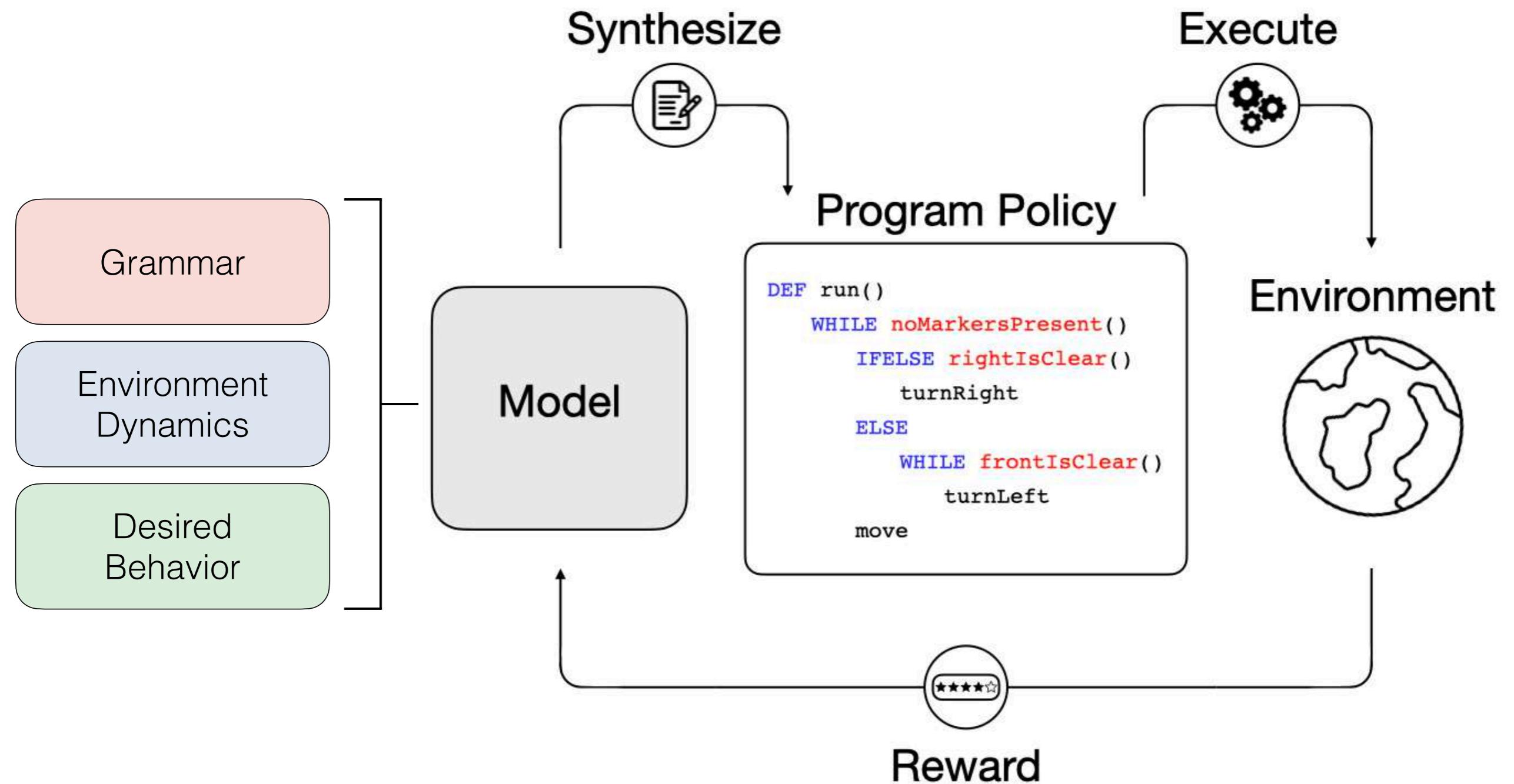
Reinforcement Learning



Reinforcement Learning by Synthesizing Programs



Reinforcement Learning by Synthesizing Programs



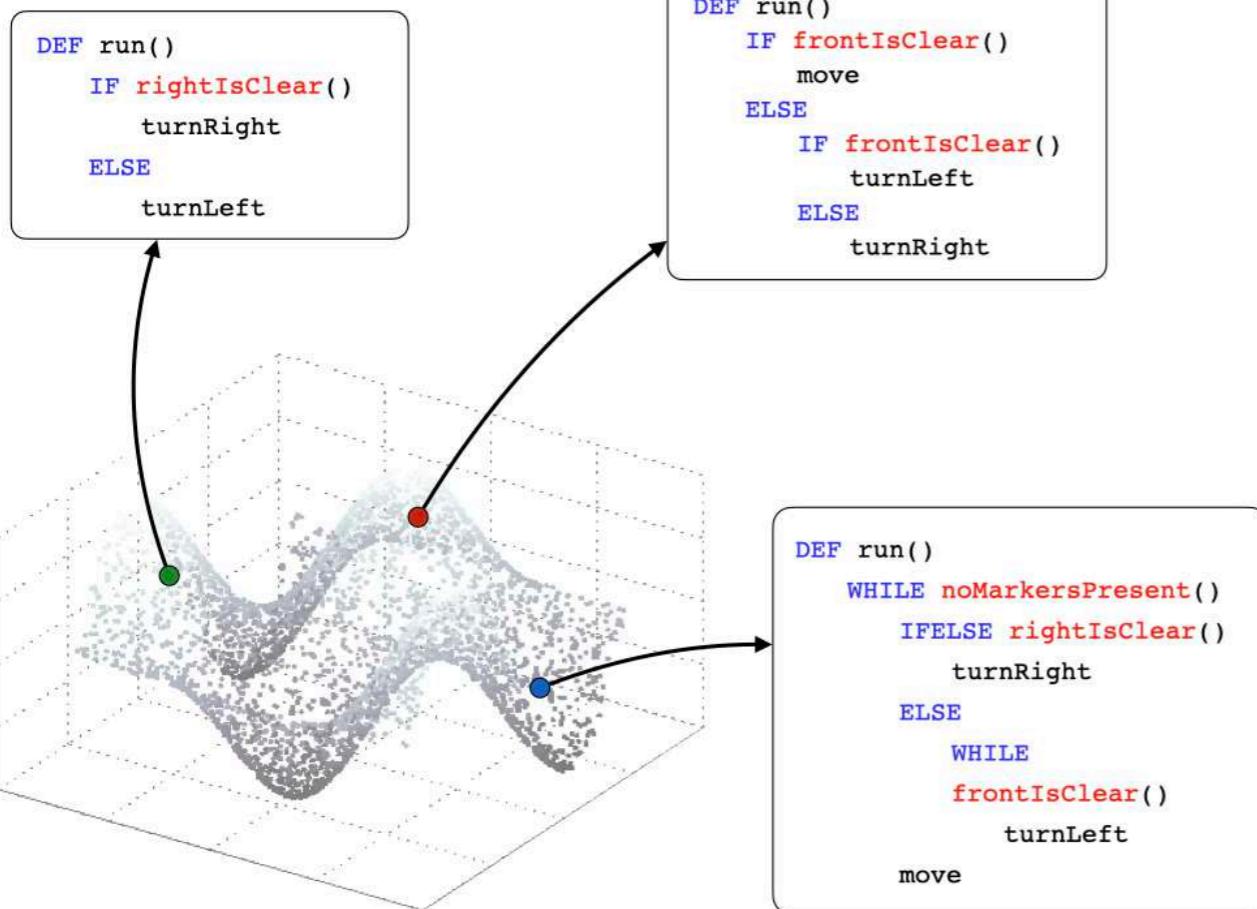
Method Overview

Stage 1

Learn a program embedding space from randomly generated programs

Grammar

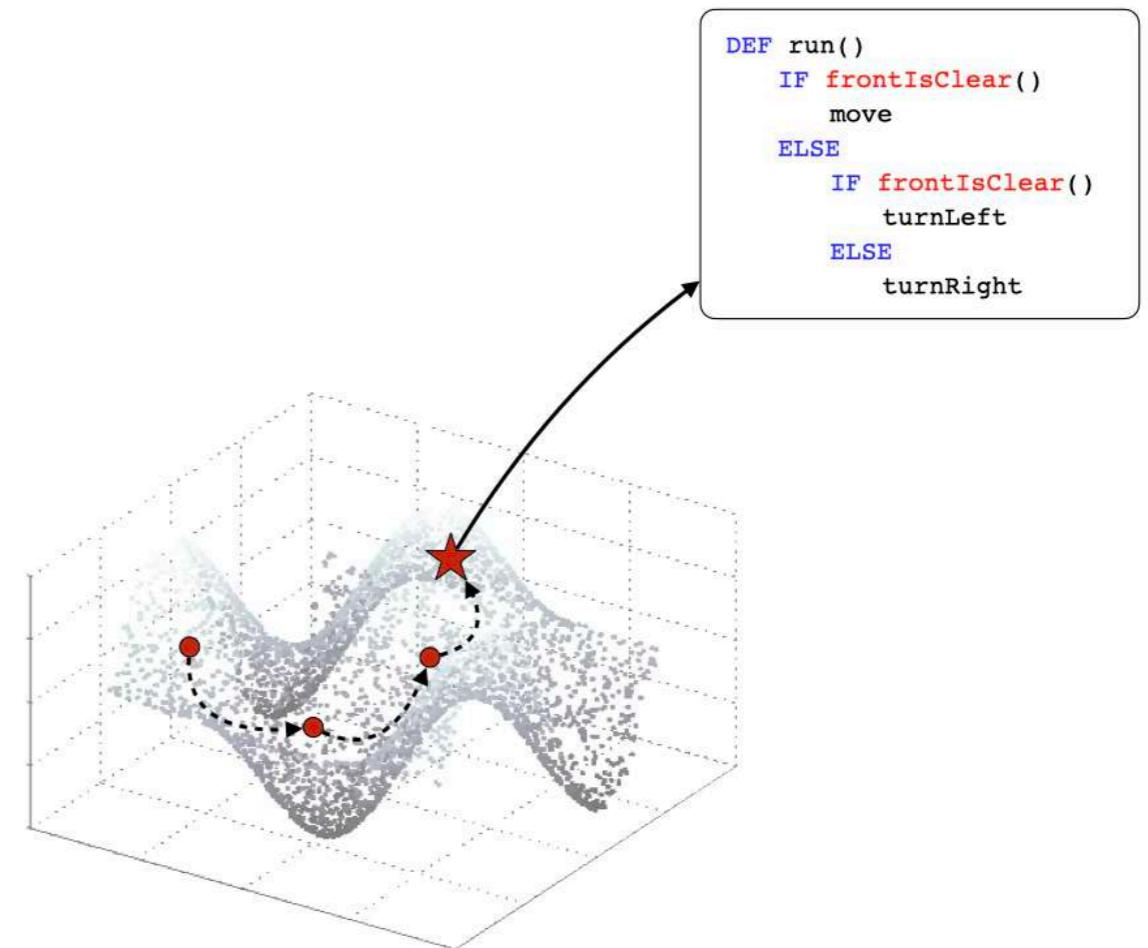
Environment
Dynamics



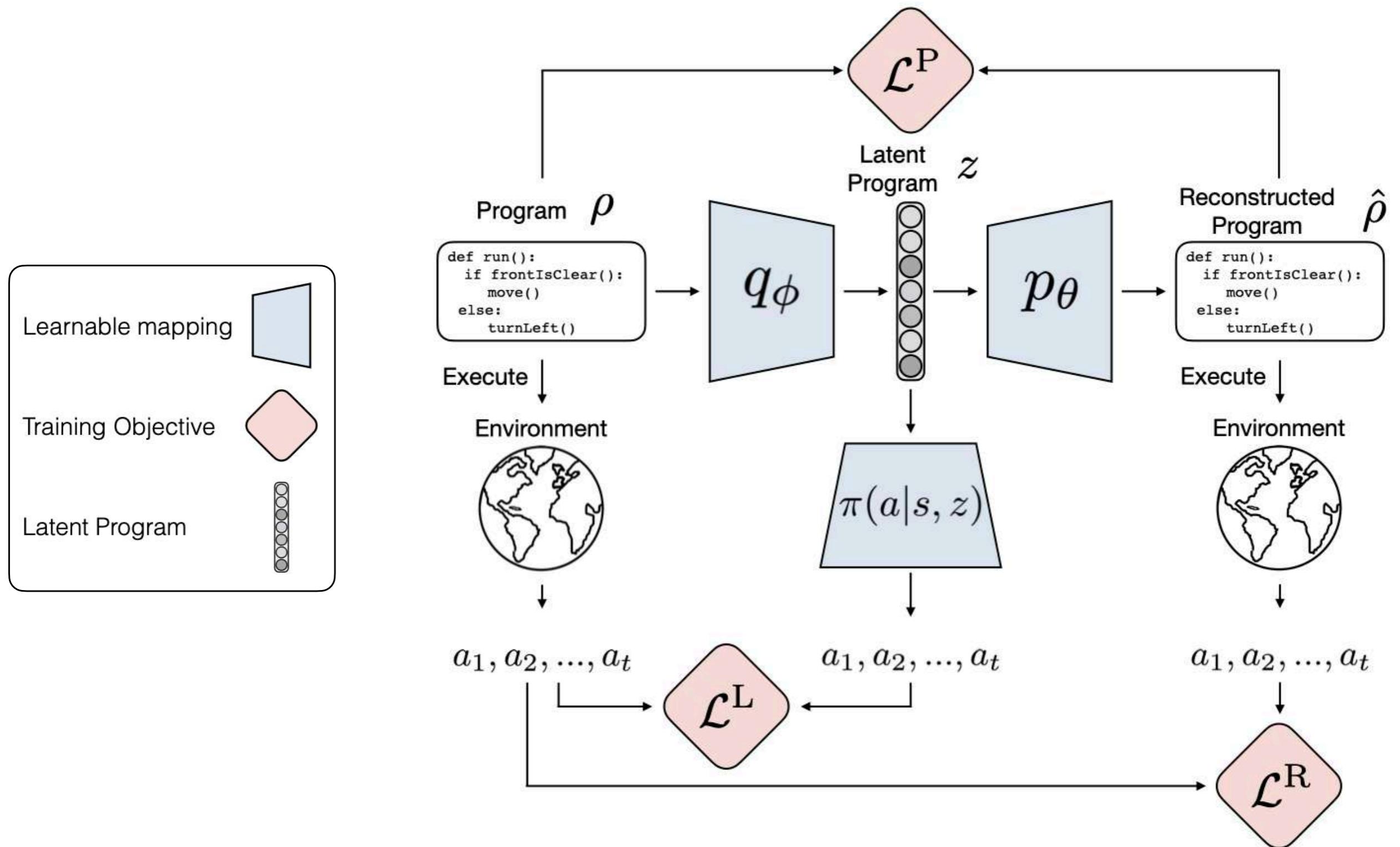
Stage 2

Search for a task-solving program

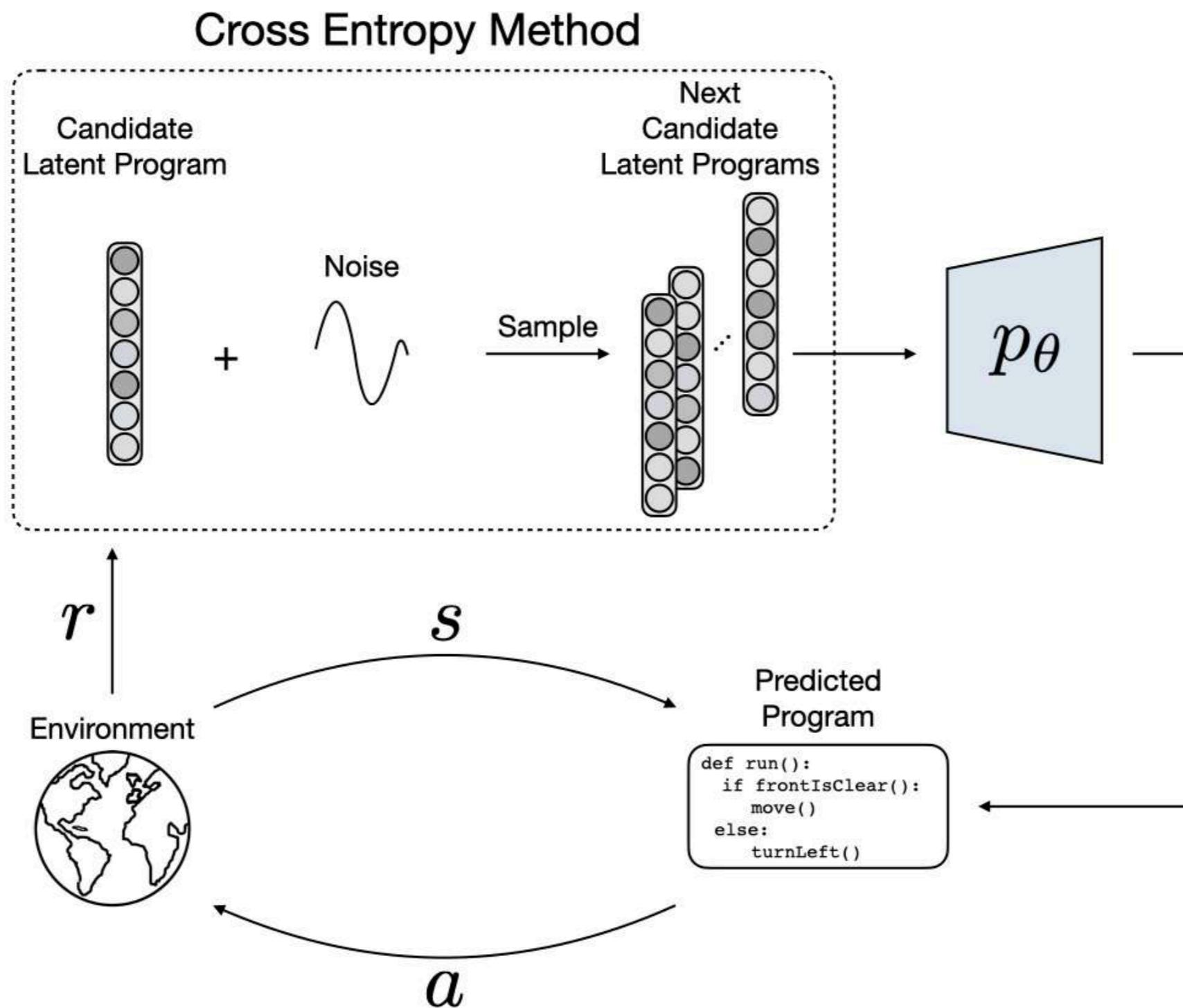
Desired
Behavior



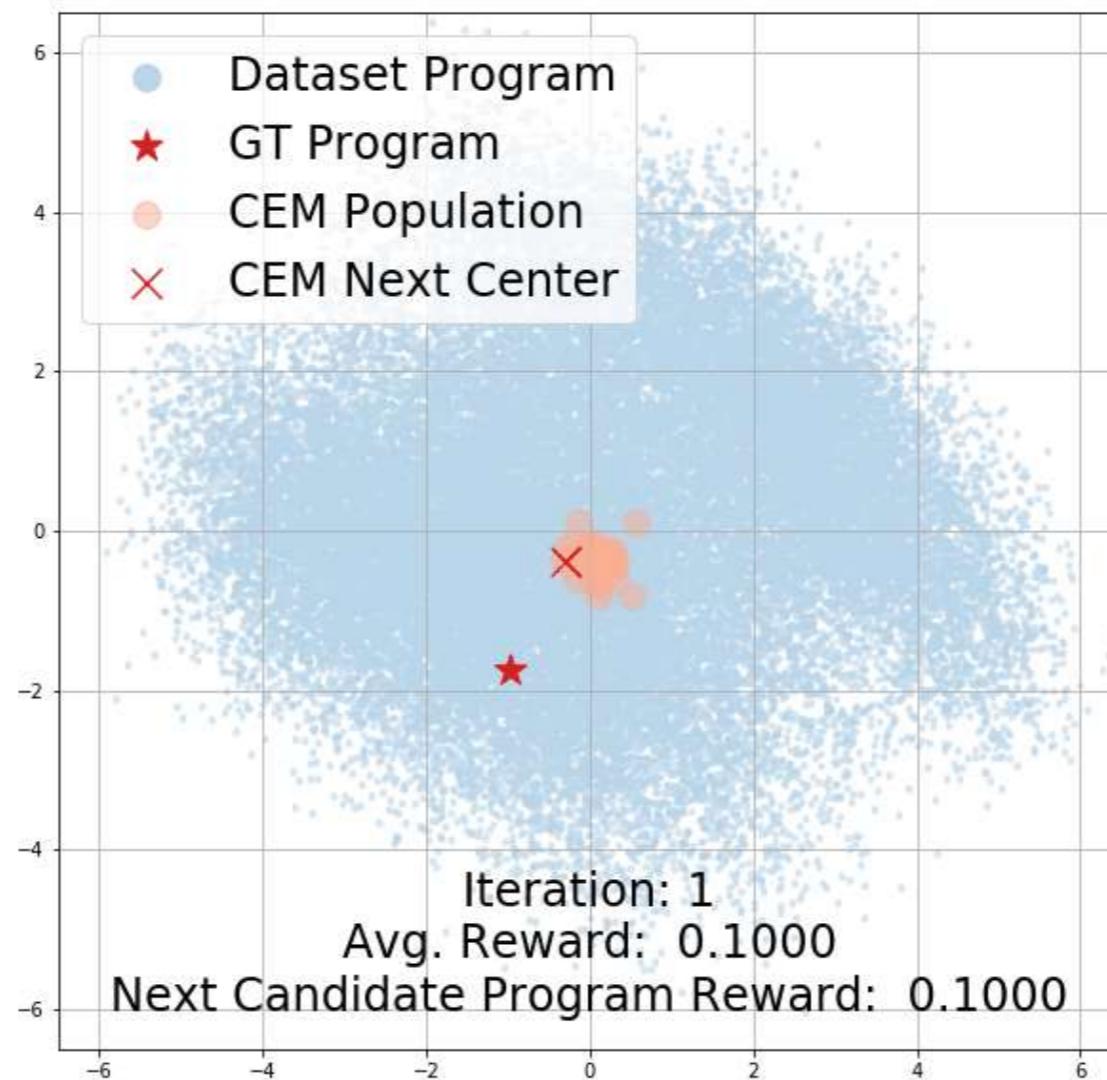
Stage 1: Learning a Program Embedding Space



Stage 2: Latent Program Search with Cross Entropy Method

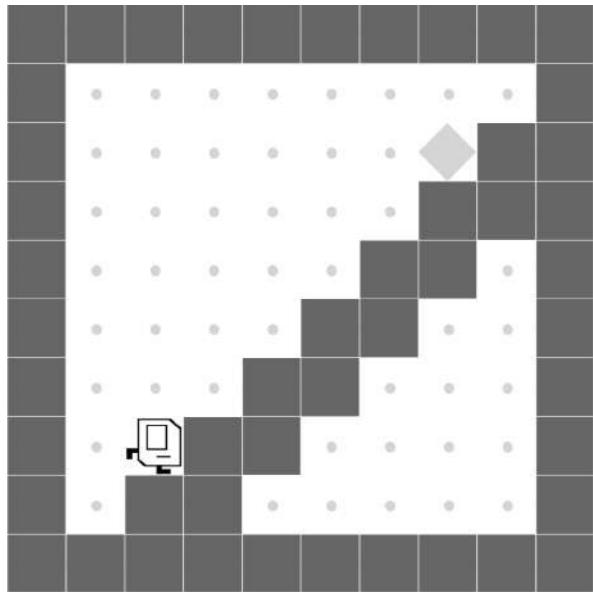


Cross Entropy Method Trajectory Visualization

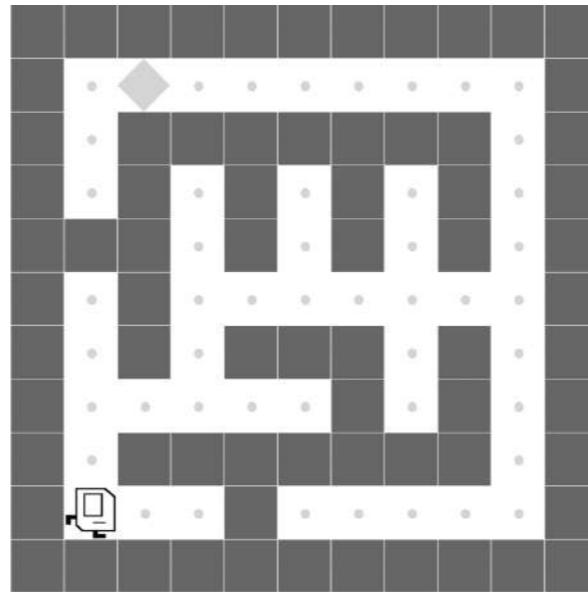


Karel Tasks

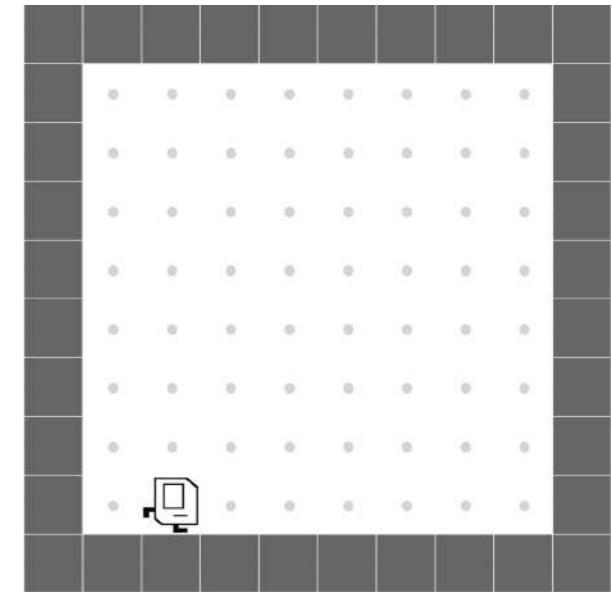
StairClimber



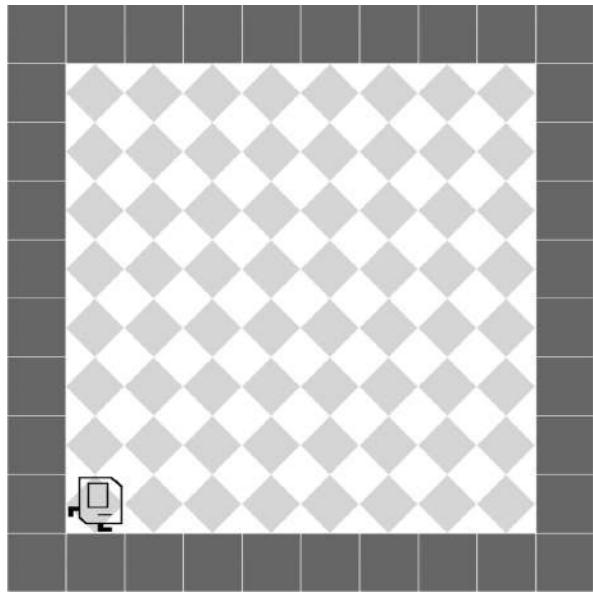
Maze



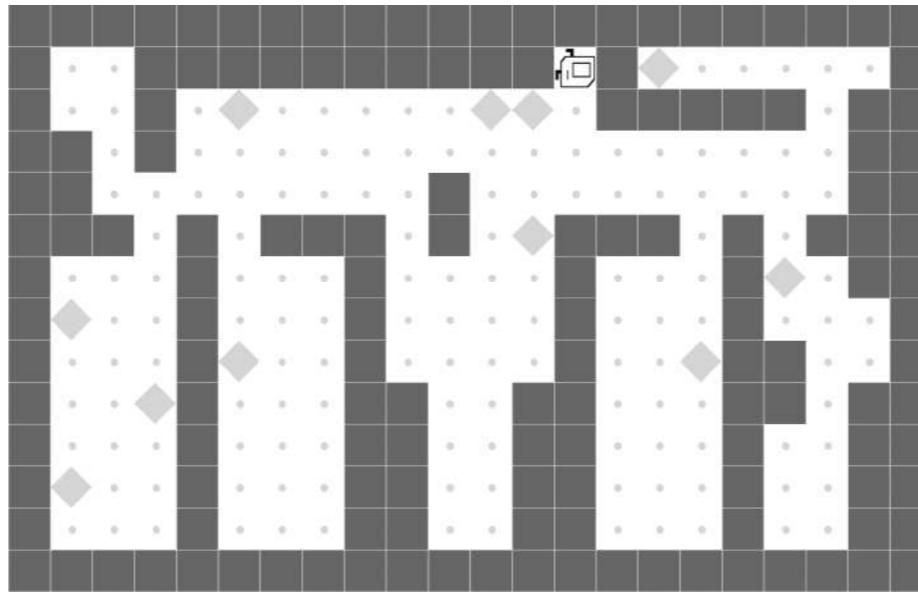
FourCorners



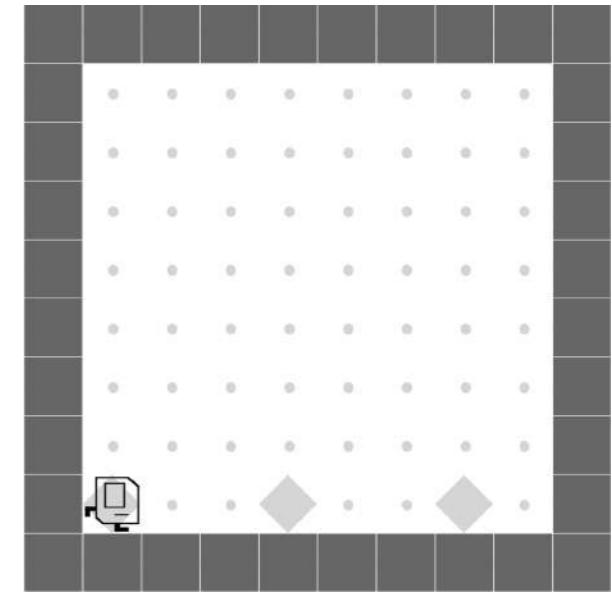
Harvester



CleanHouse

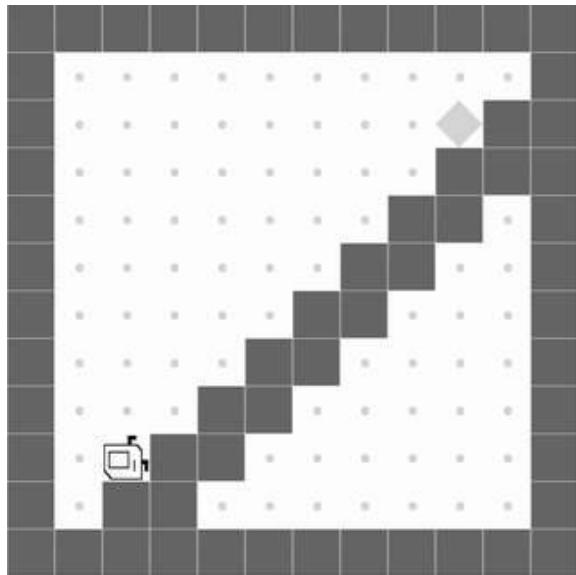


TopOff

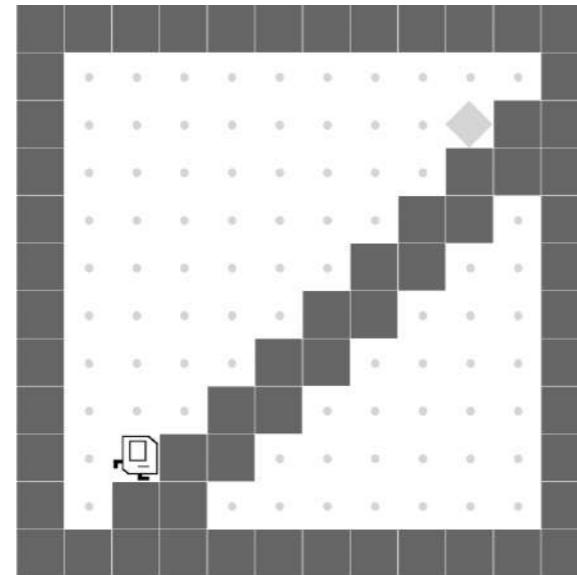


Qualitative Results

StairClimber

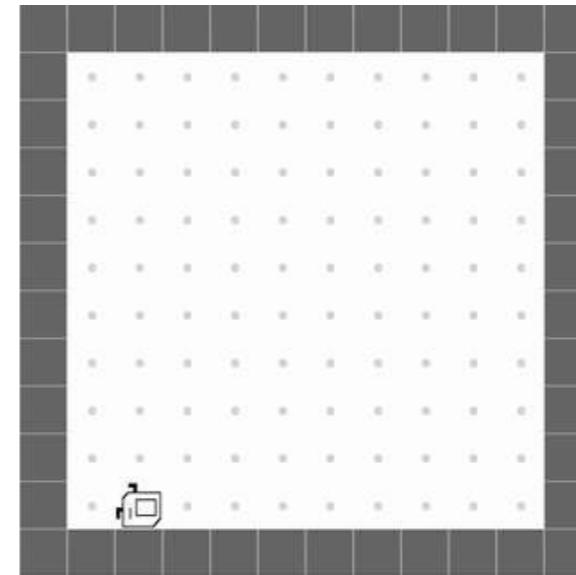


DRL

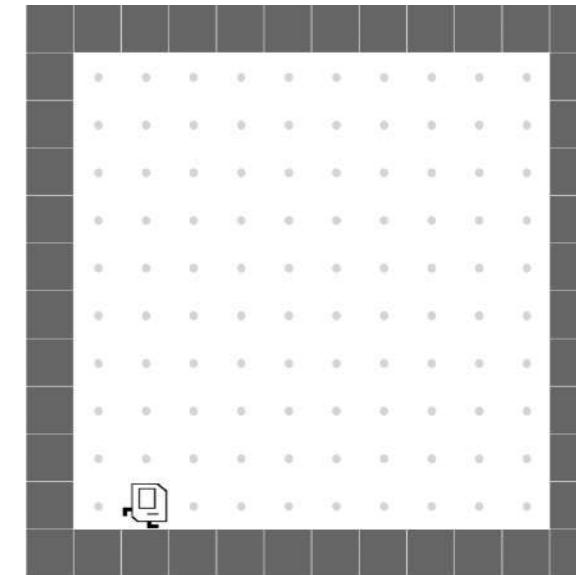


LEAPS

FourCorners

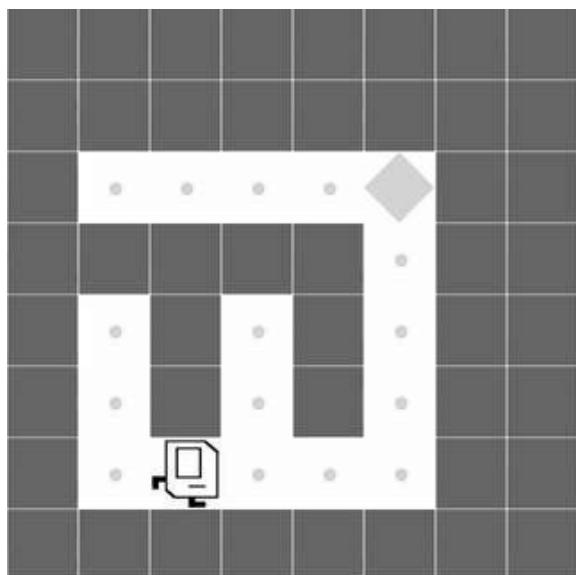


DRL

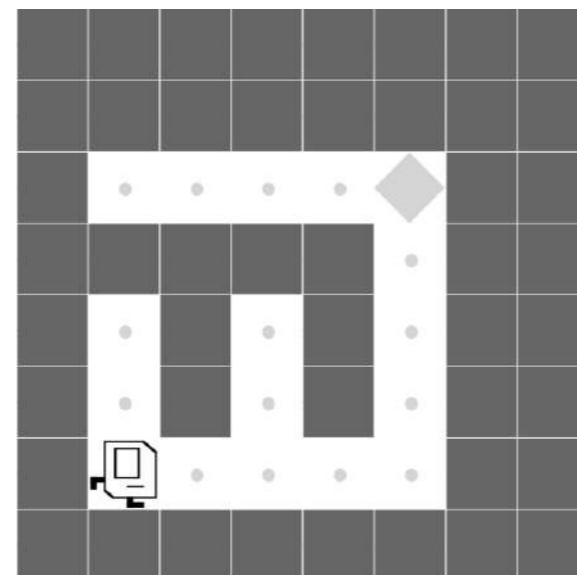


LEAPS

Maze

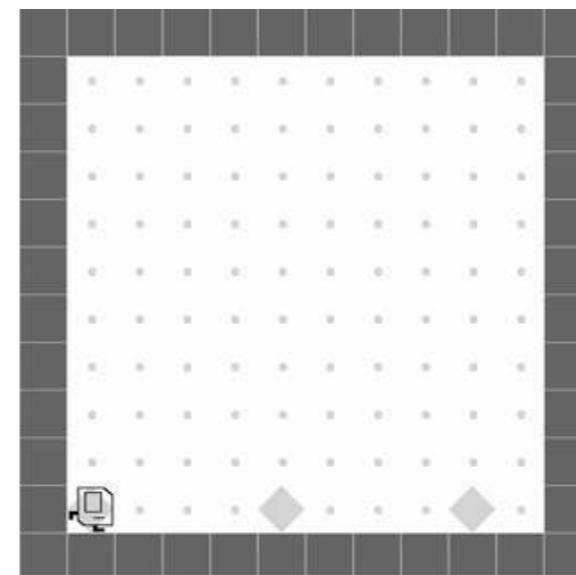


DRL

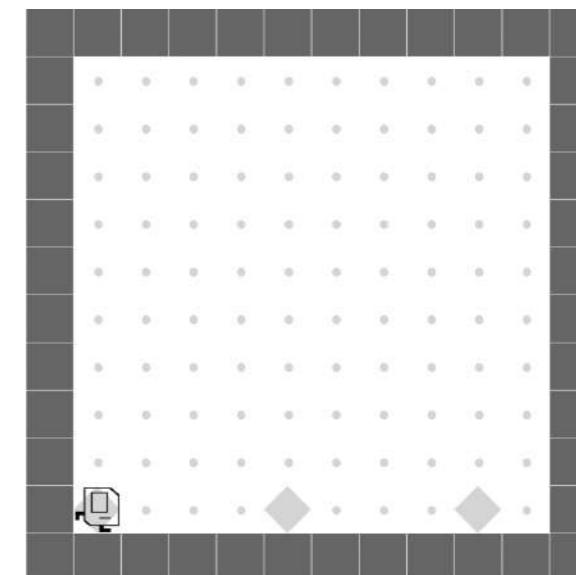


LEAPS

TopOff

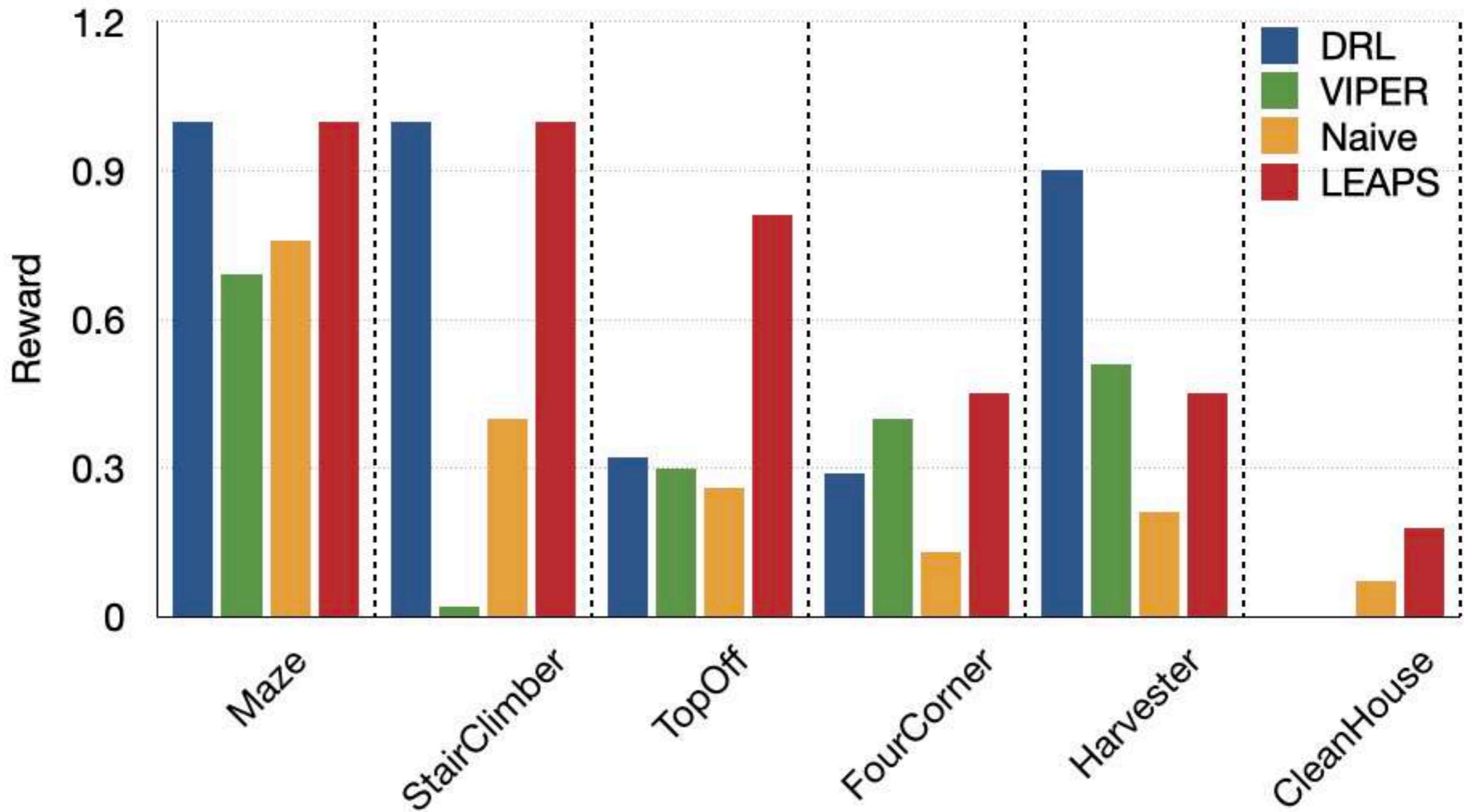


DRL



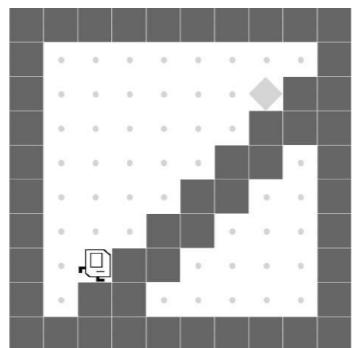
LEAPS

Quantitative Results



Zero-shot Generalization

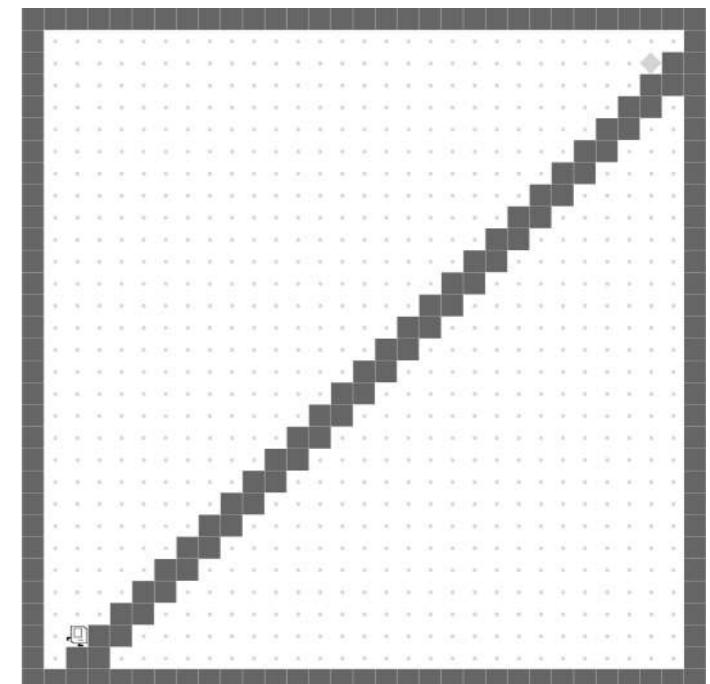
Learning on
8x8 grids



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

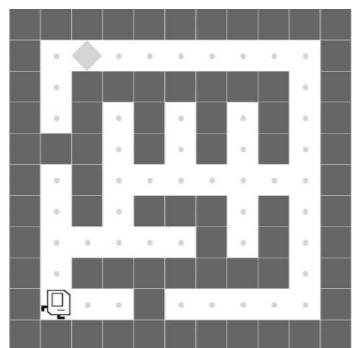


Evaluation on
100x100 grids

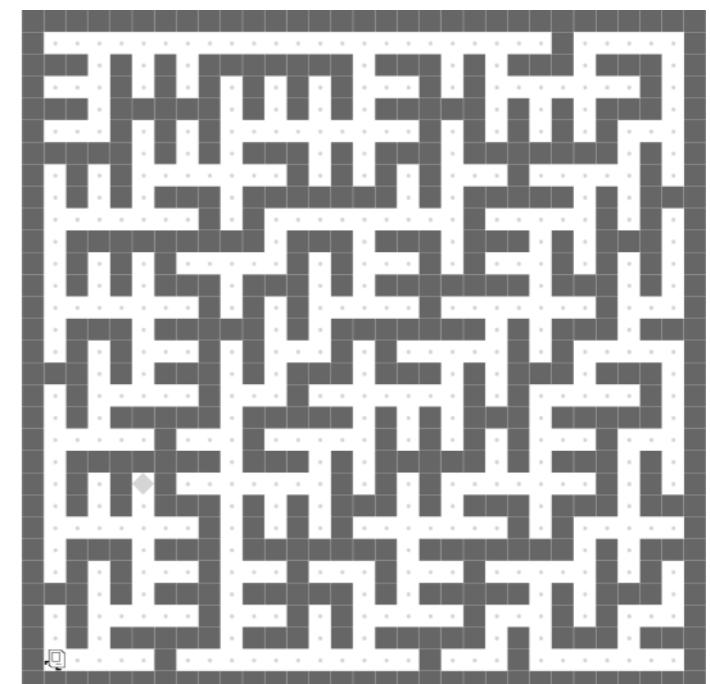


StairClimber

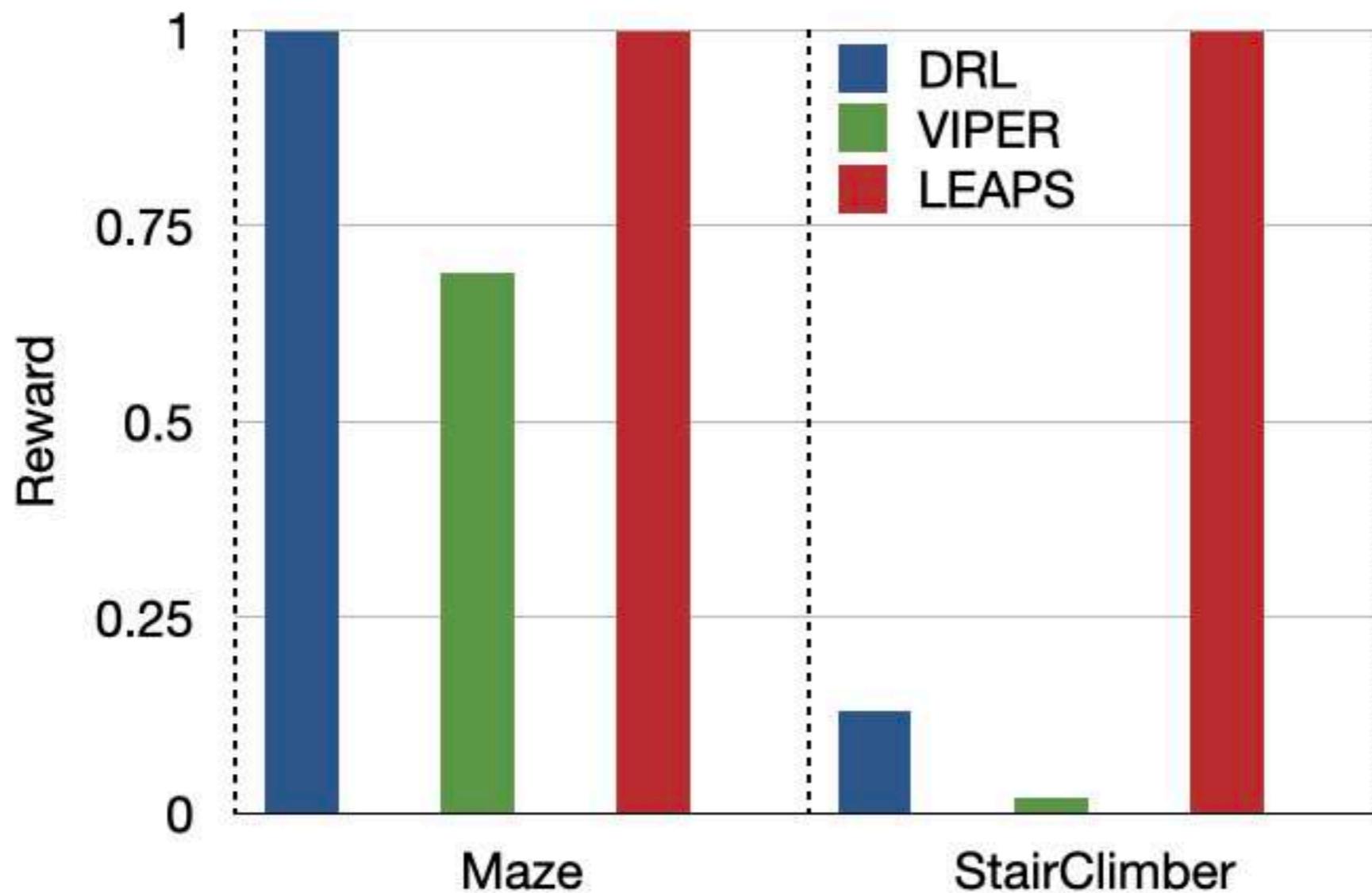
Maze



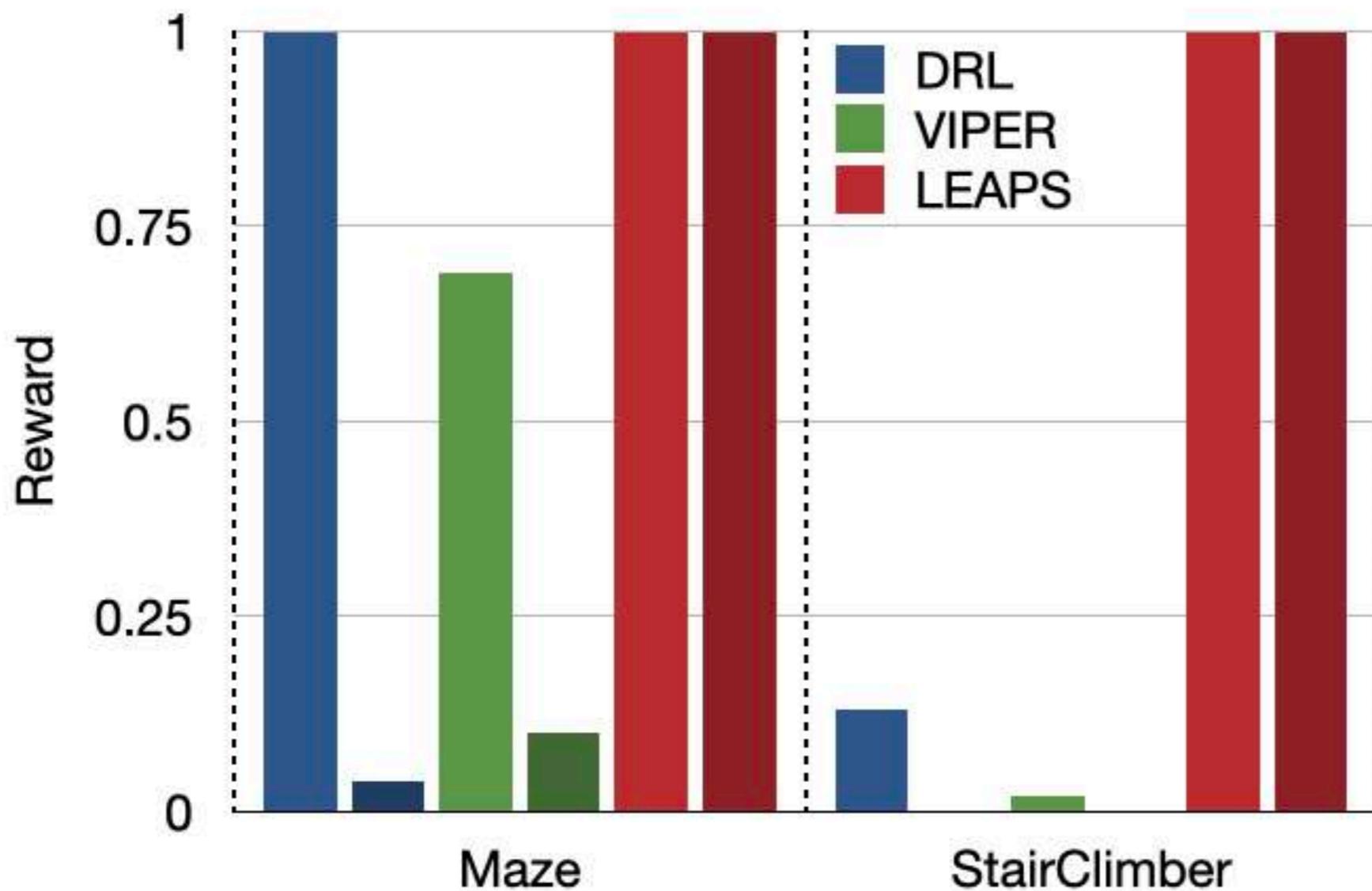
```
DEF run()
  IF frontIsClear()
    turnLeft
  WHILE noMarkersPresent()
    turnRight
    move
```



Zero-shot Generalization



Zero-shot Generalization

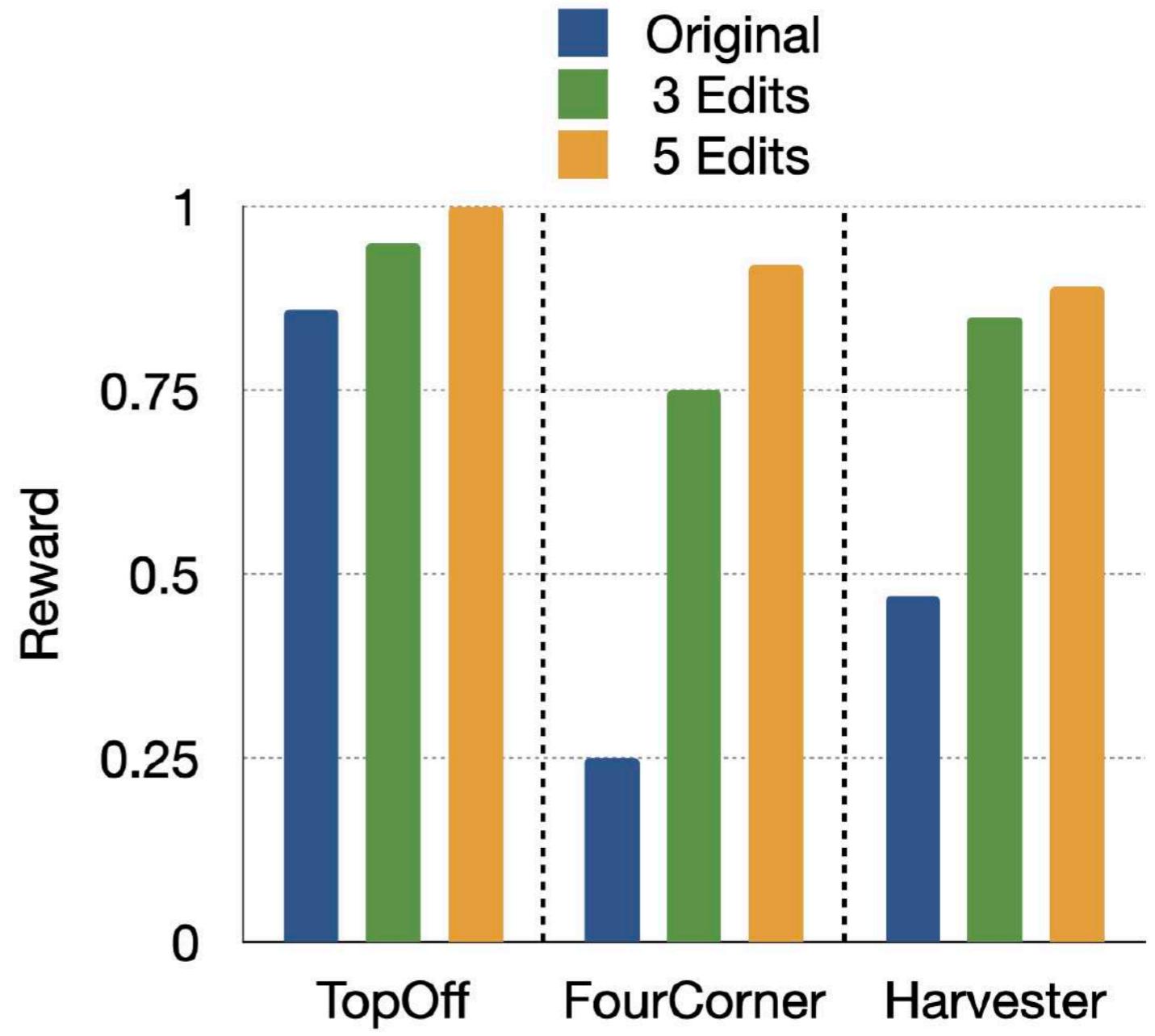


Interpretability

Human Debugging Interface

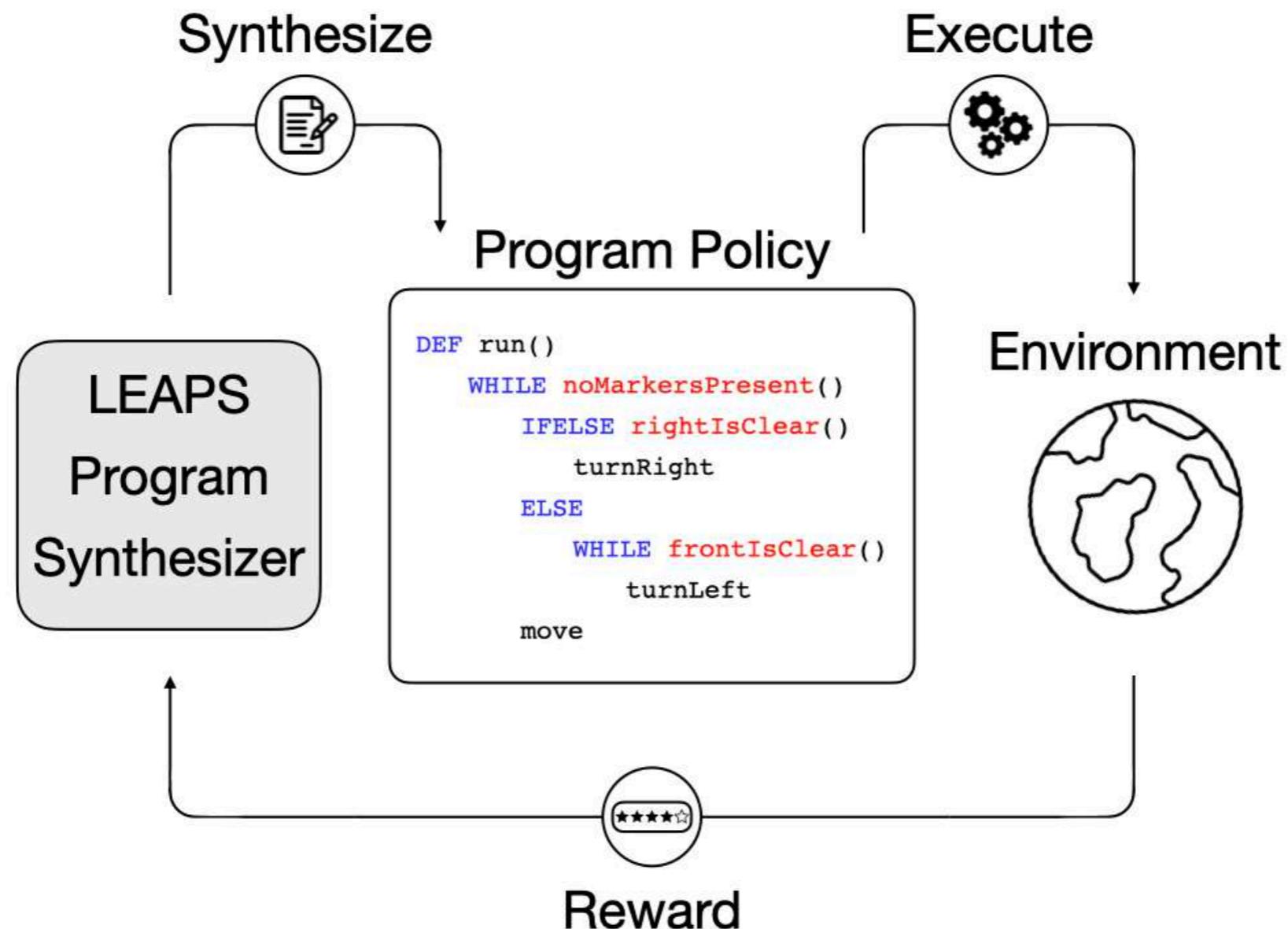


Improved Performance

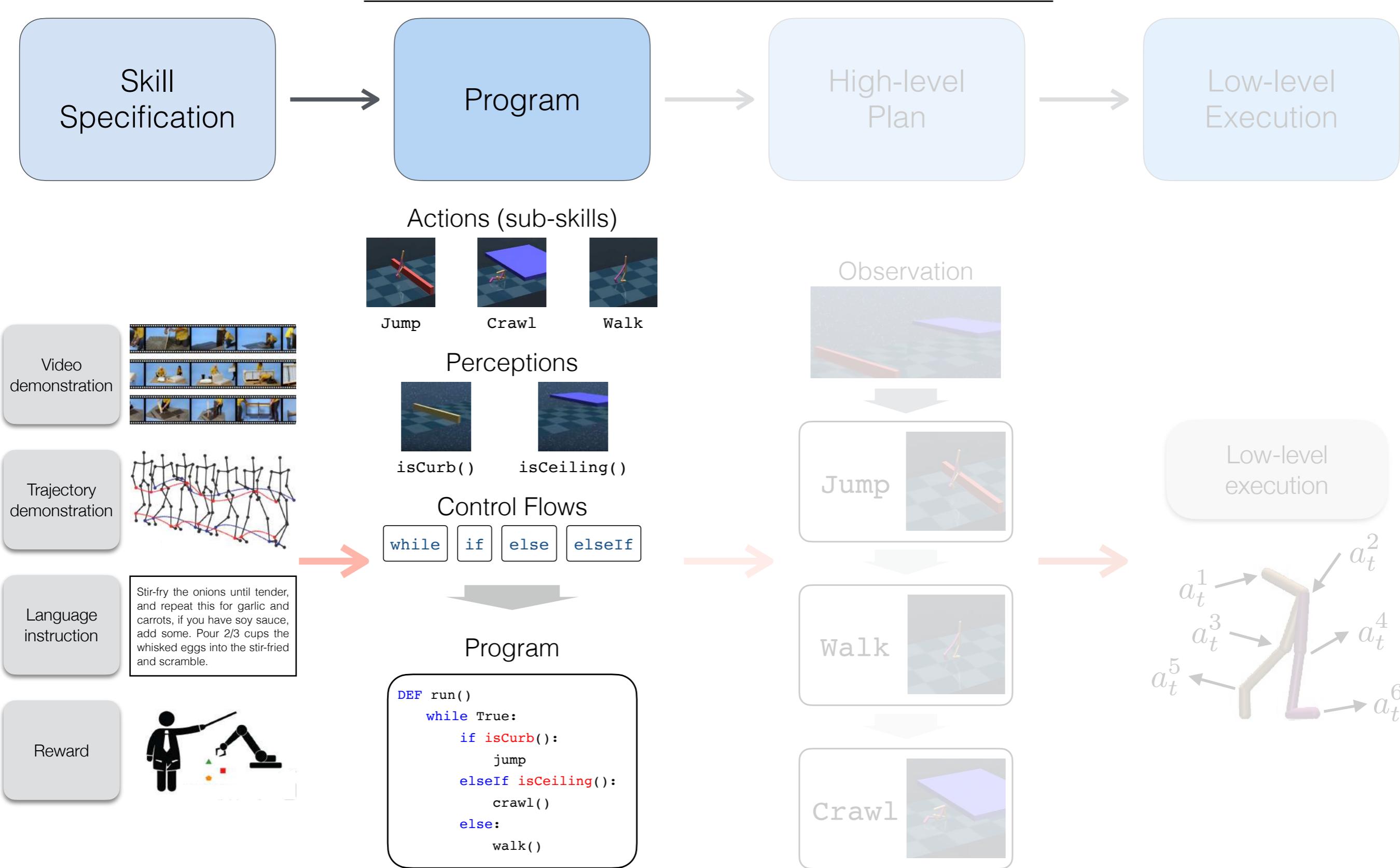


Takeaway

- Synthesize generalizable and interpretable programs from rewards



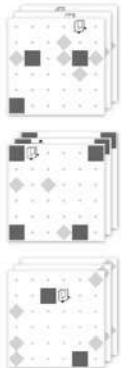
Program Inference



Program Inference

Imitation learning from demonstrations

Demonstrations



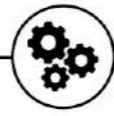
Synthesize



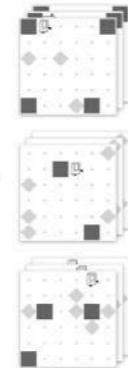
Program

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

Execute

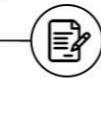


Execution



Reinforcement learning from rewards

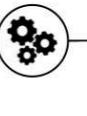
Synthesize



Program Policy

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

Execute



Reward

LEAPS
Program
Synthesizer

Program Inference

Imitation learning from demonstrations

Demonstrations



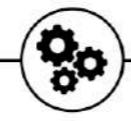
Synthesize



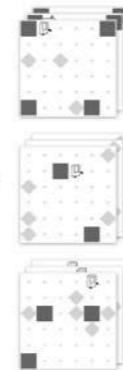
Program

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

Execute



Execution



Reinforcement learning from rewards

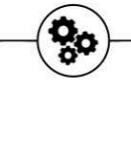
Synthesize



Program Policy

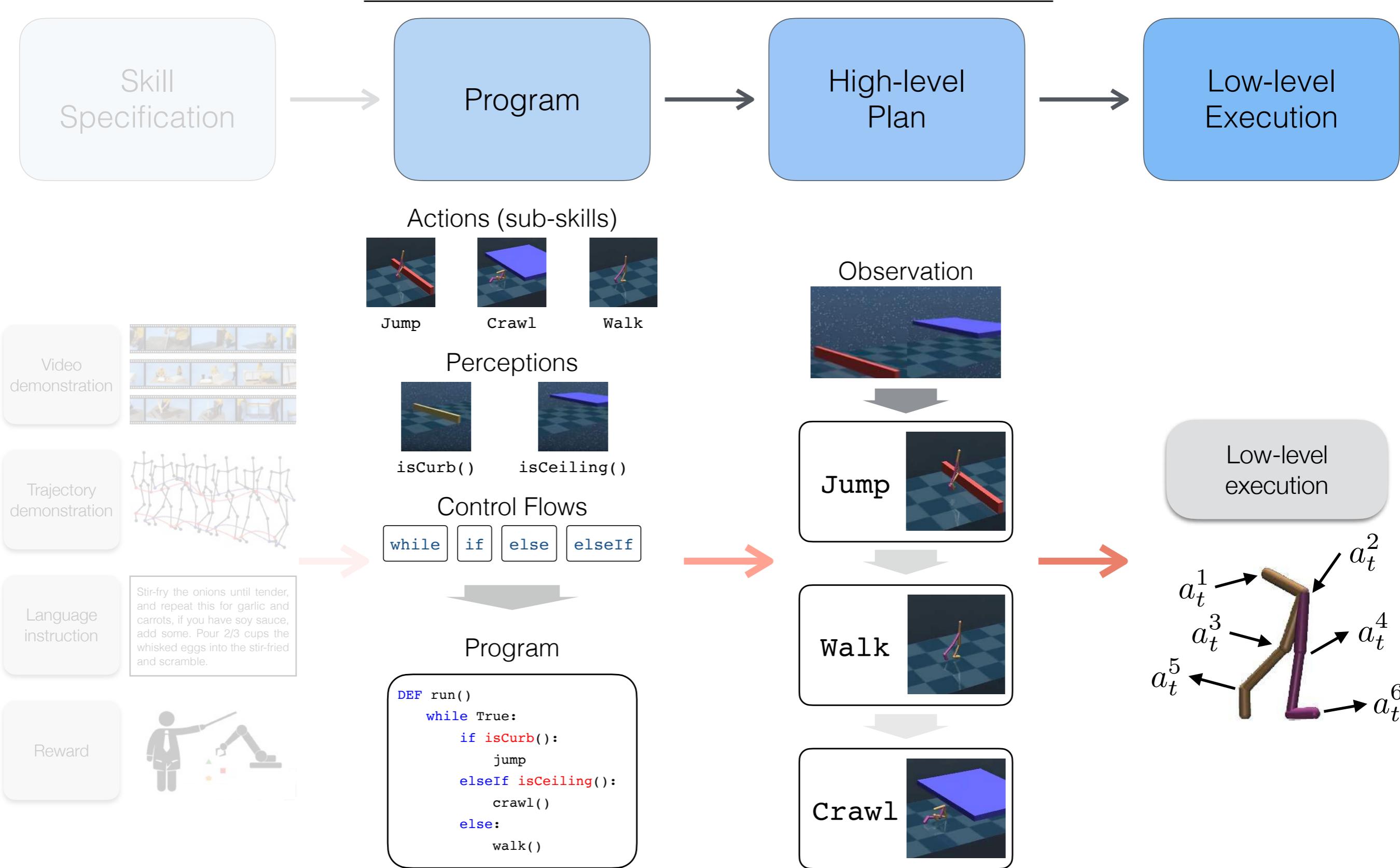
```
DEF run()
  WHILE noMarkersPresent()
    IFELSE rightIsClear()
      turnRight
    ELSE
      WHILE frontIsClear()
        turnLeft
      move
```

Execute



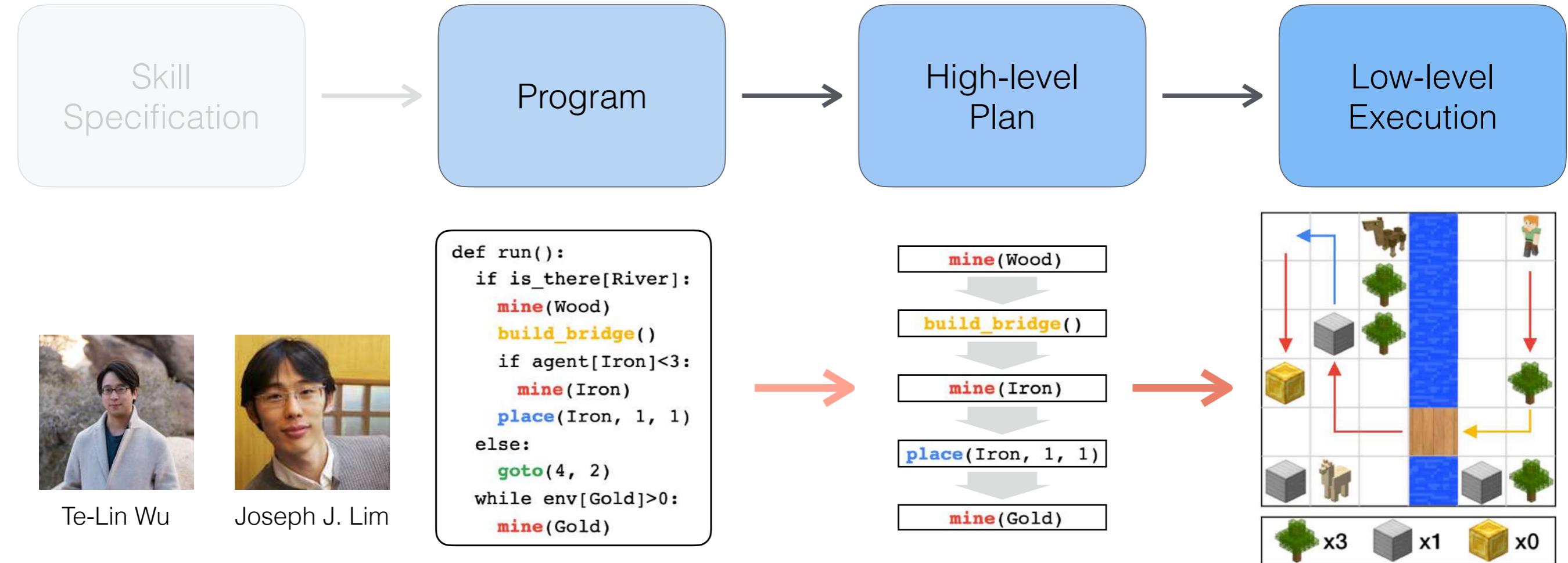
Reward

Task Execution



Program Guided Agent

ICLR 2020 (Spotlight)



Te-Lin Wu



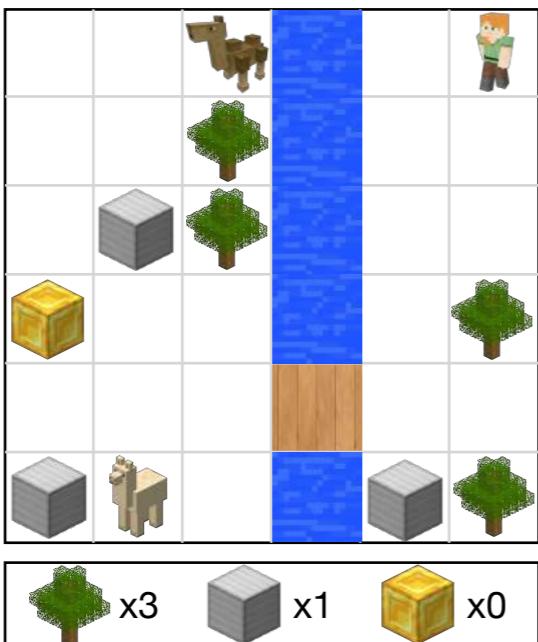
Joseph J. Lim

Problem Formulation

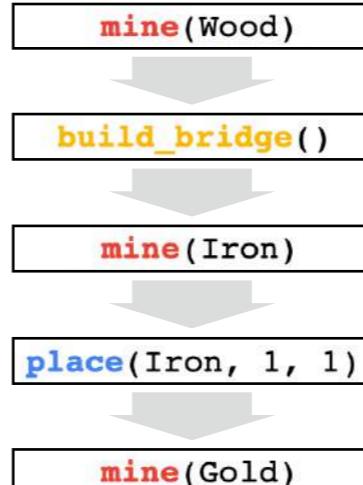
Program
(task)

```
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)
```

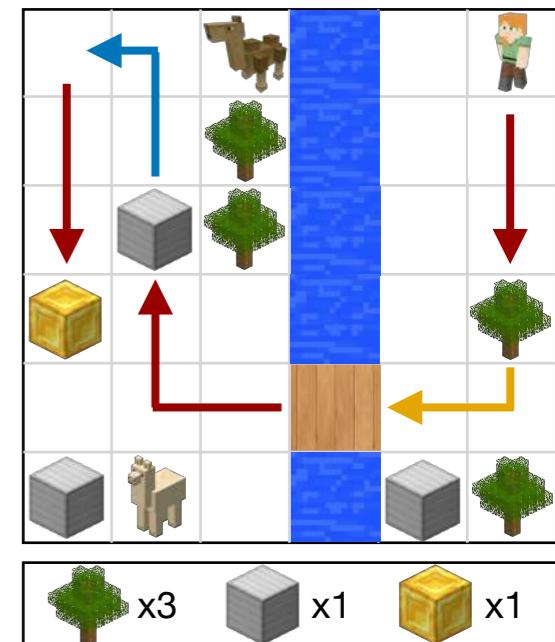
Observation



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)
```

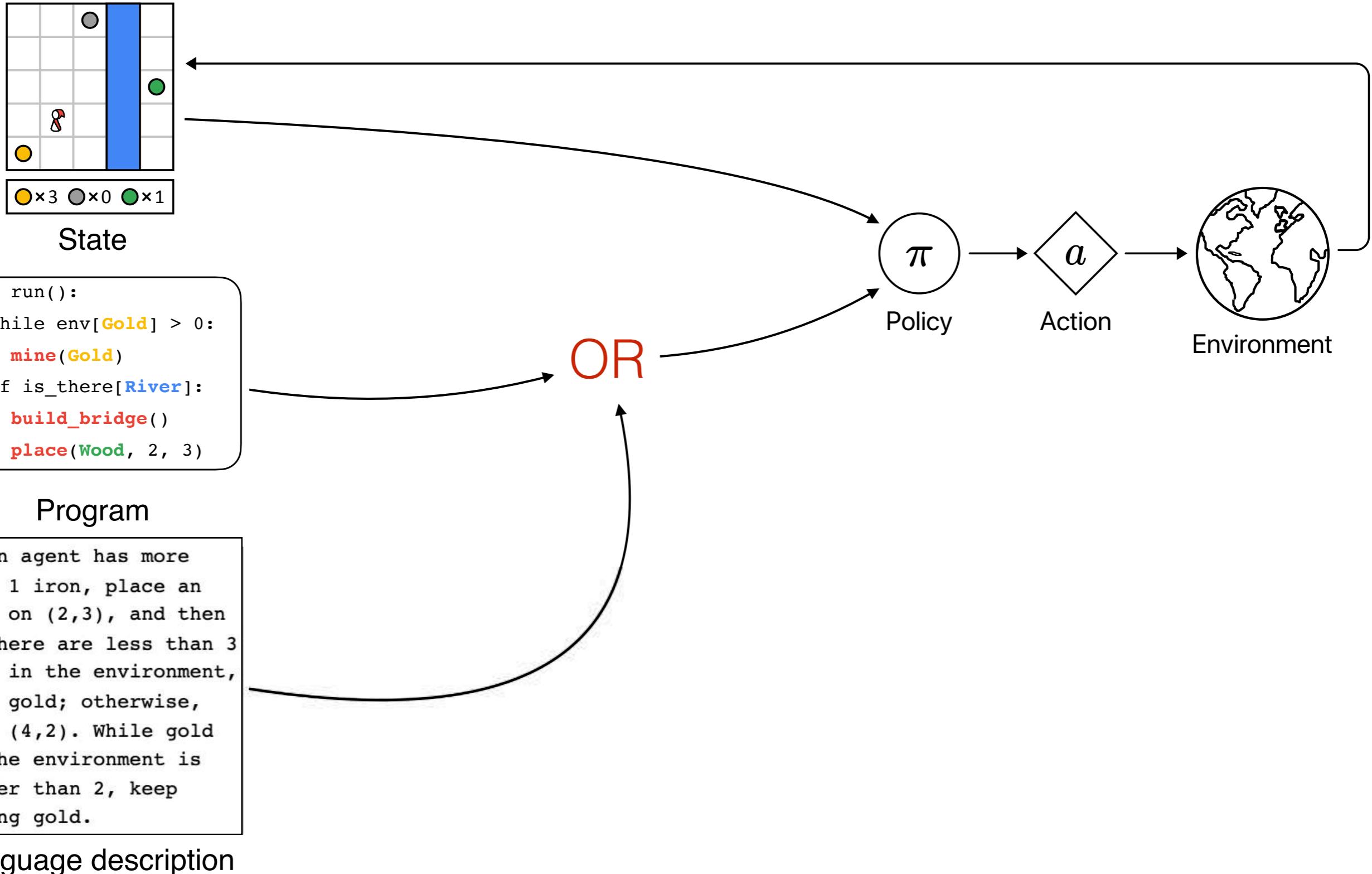
```
def run()
    while agent[Wood] <= 11:
        place(Wood, 2, 4)
        place(Iron, 1, 1)
        place(Iron, 8, 5)
        mine(Gold)
        mine(Gold)
        mine(Gold)
        repeat(4):
            sell(Gold)
            sell(Iron)
```

Natural Language Descriptions

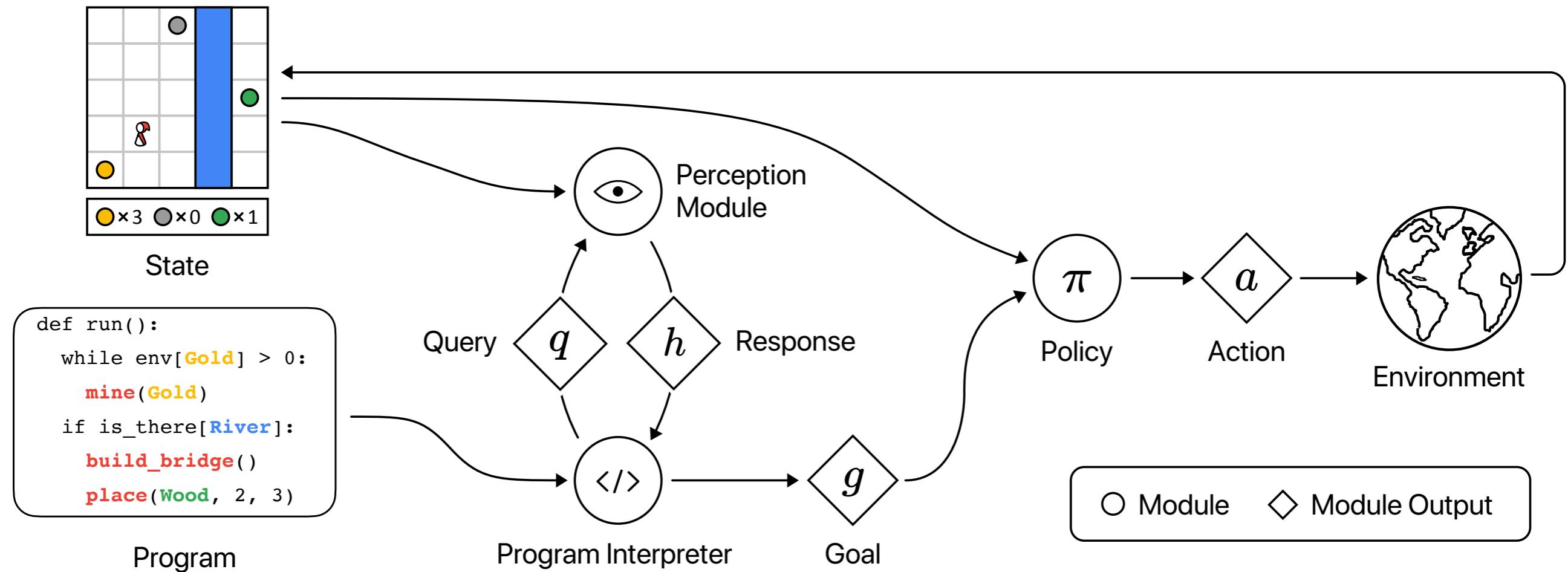
If a river is in the environment, mine a wood and then use it to build a bridge. And then if agent has less than there iron, place an iron at (2,3). Otherwise if no river, goto location (4,2). Finally, whenever there's still gold in the environment, mine a gold.

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



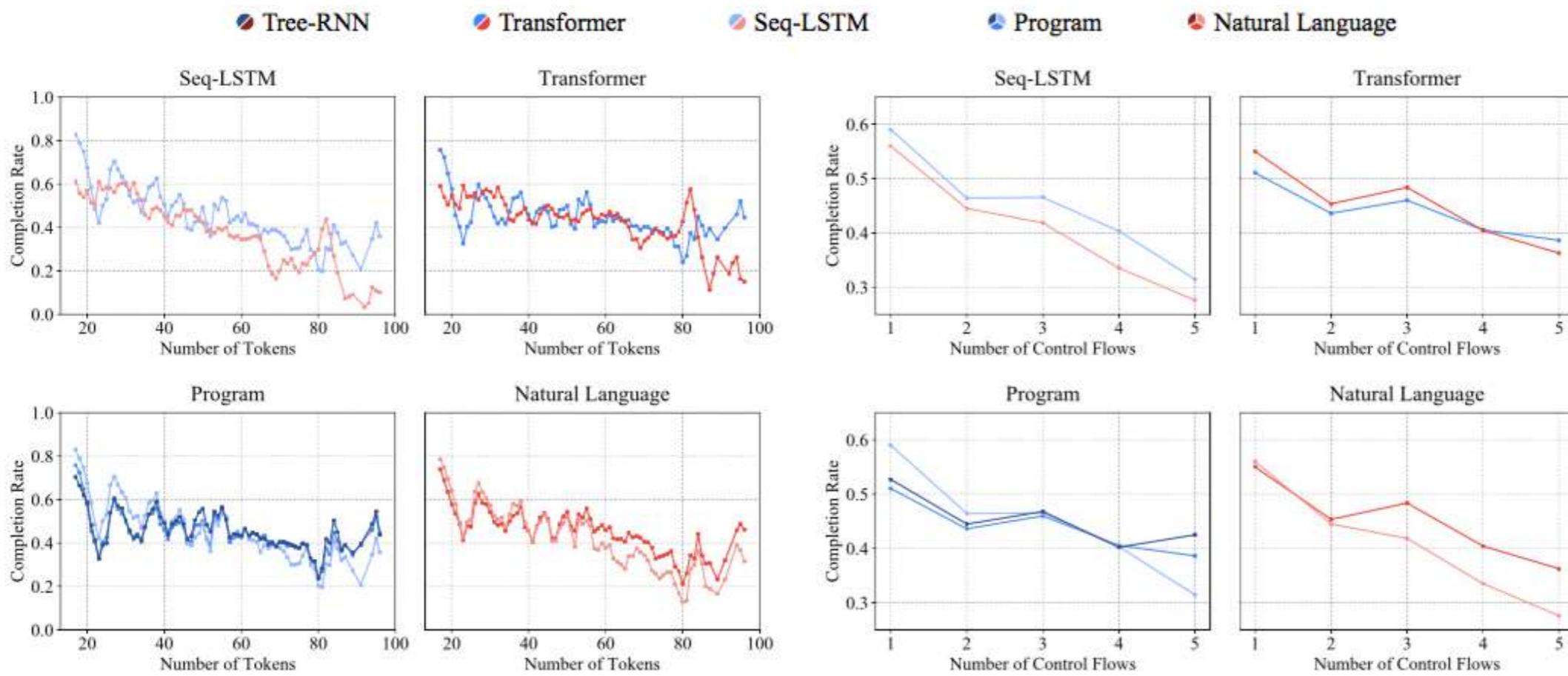
Program Guided Agent



Quantitative Results

Generalization

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



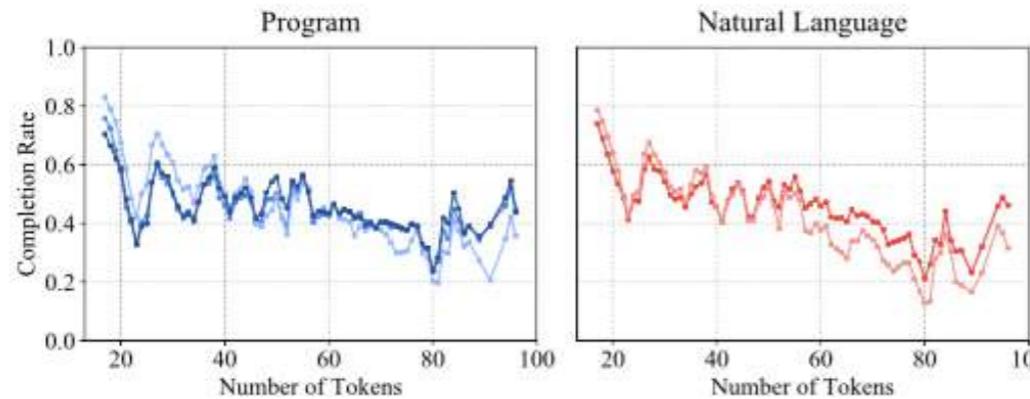
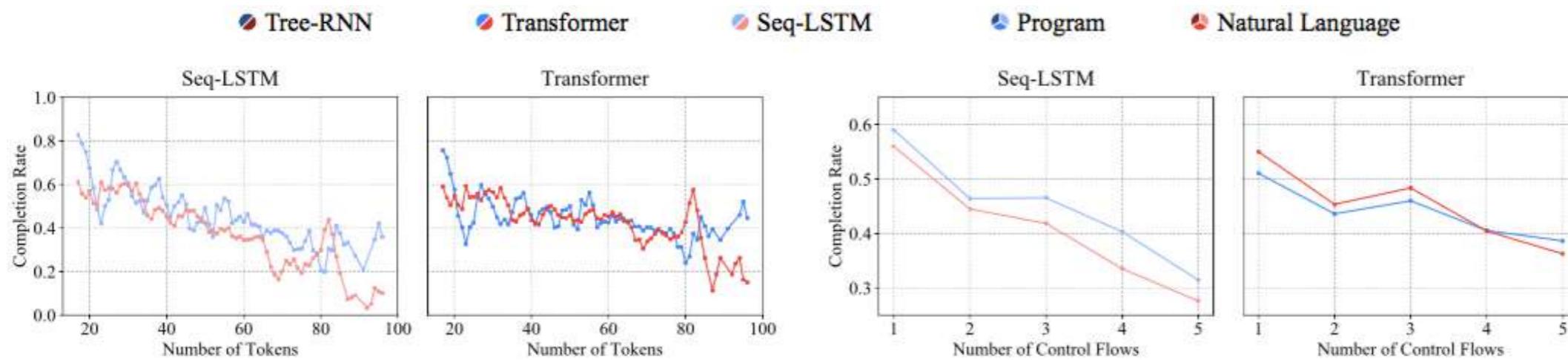
(a) Instruction Length

(b) Instruction Complexity

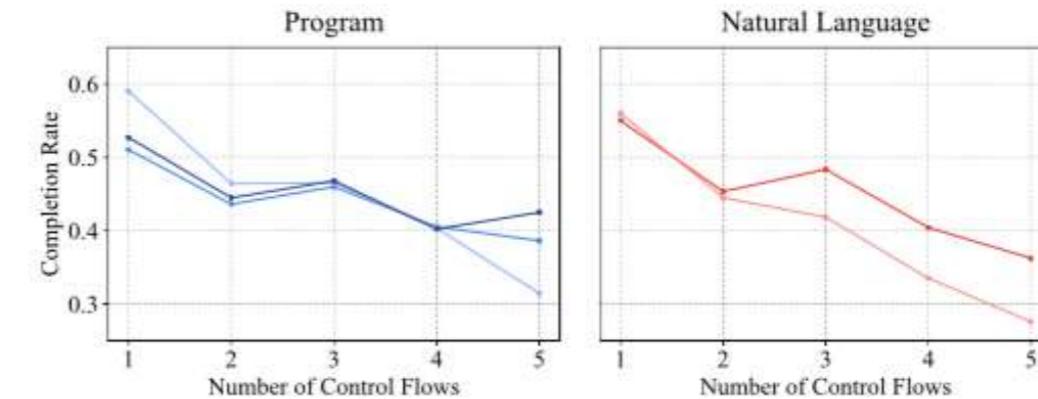
Quantitative Results

Natural Languages < Programs

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



(a) Instruction Length

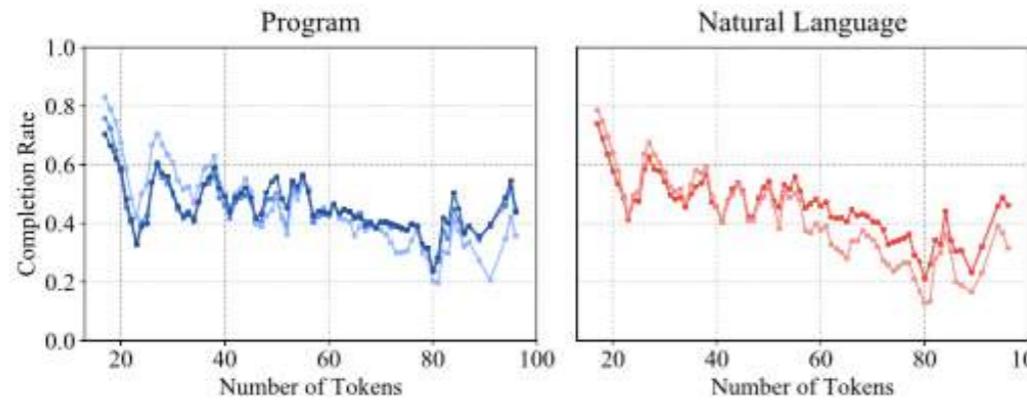
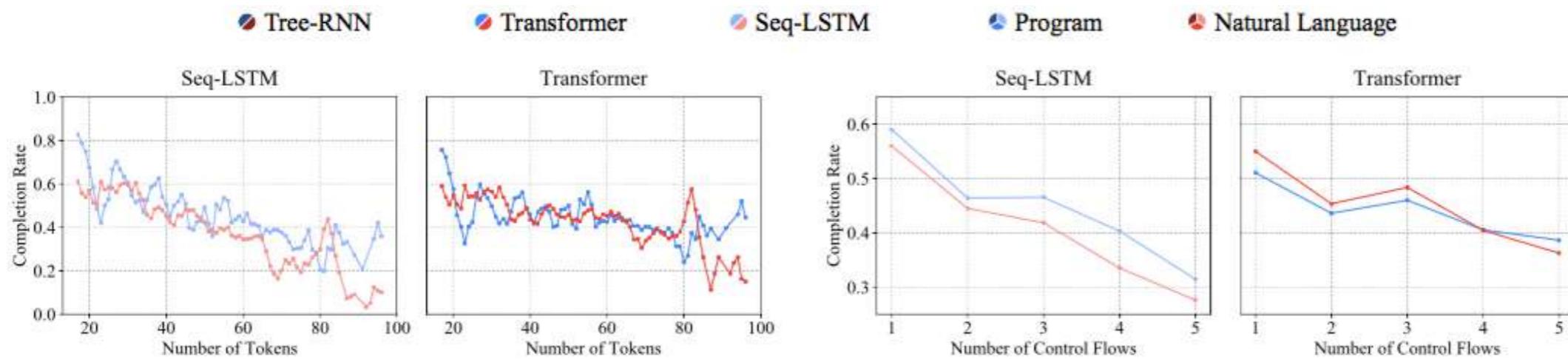


(b) Instruction Complexity

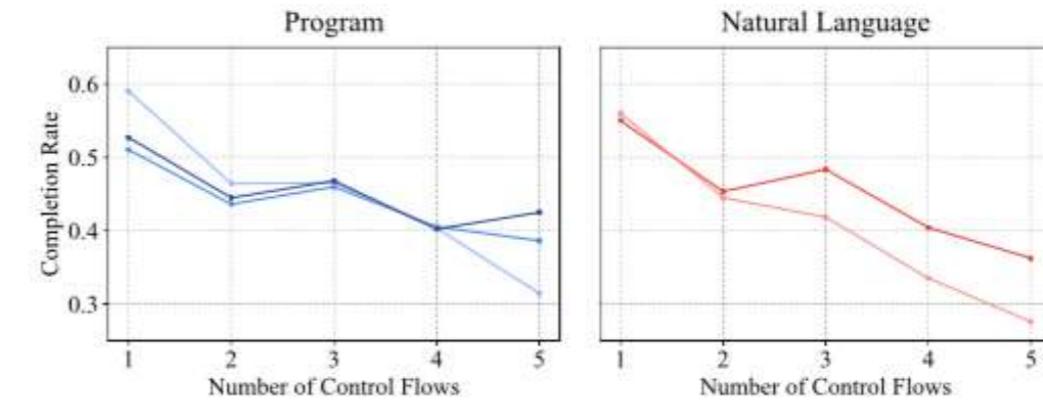
Quantitative Results

End-to-end < Ours (modular)

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



(a) Instruction Length



(b) Instruction Complexity

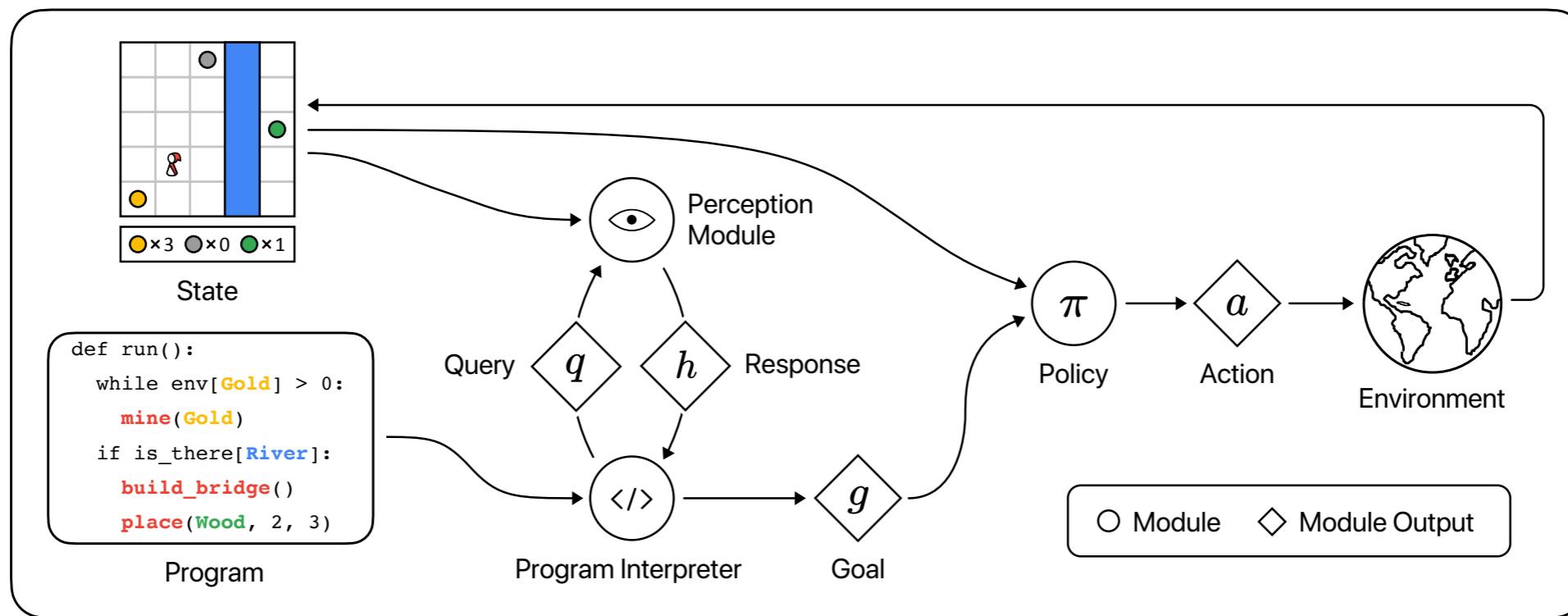
Takeaway

- Specific tasks using programs

Program

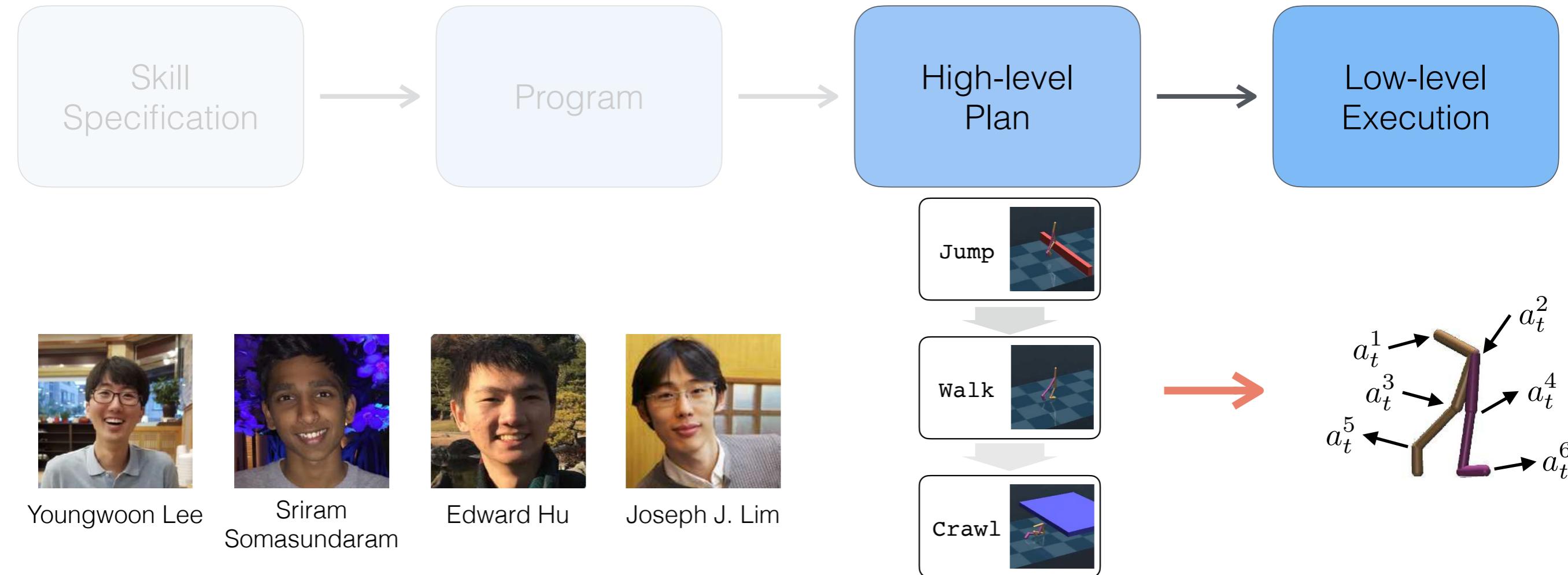
```
def run():
    if is_there[River]:
        mine(Wood)
        build_bridge()
    if agent[Iron]<3:
        mine(Iron)
        place(Iron, 1, 1)
    else:
        goto(4, 2)
    while env[Gold]>0:
        mine(Gold)
```

- Leverage the structure of programs with a modular framework

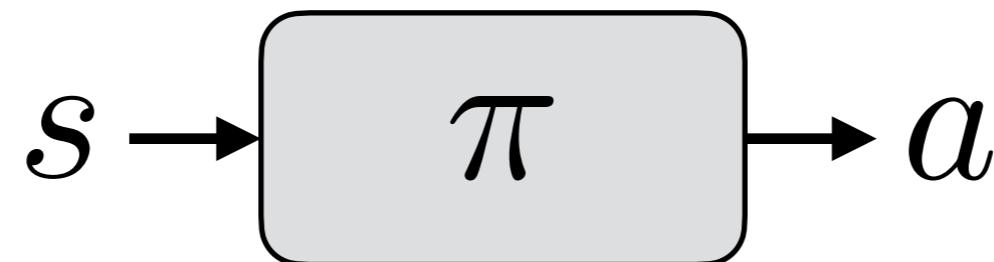


Composing Complex Skills by Learning Transition Policies

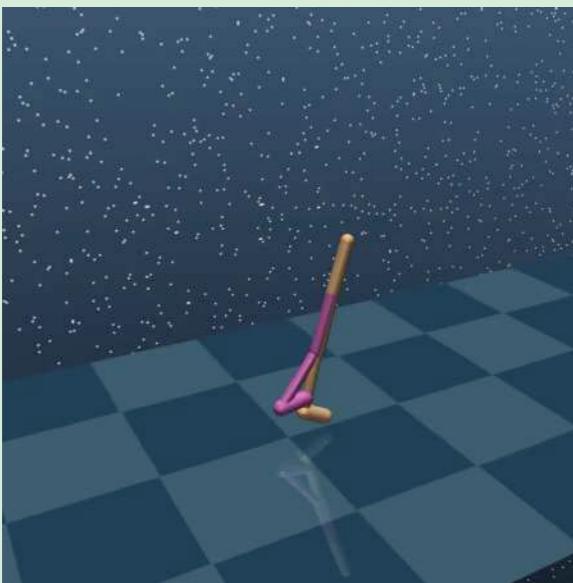
ICLR 2019



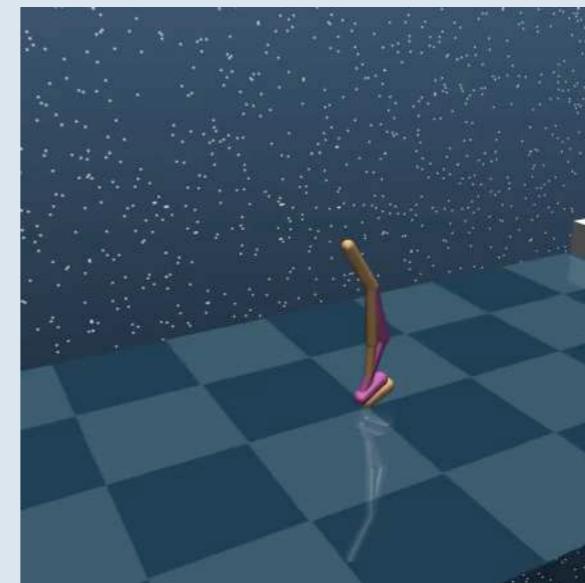
Learned Skills



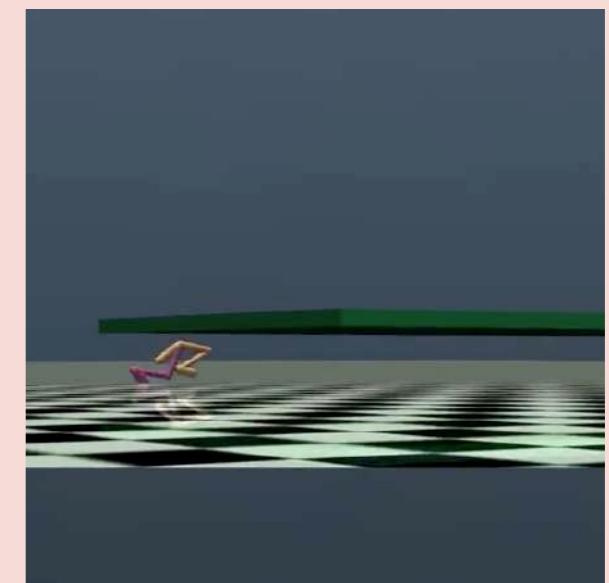
Walk



Jump



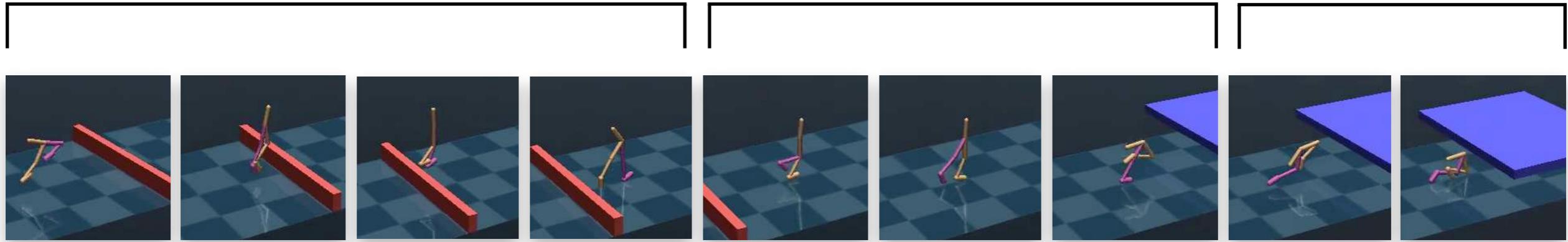
Crawl



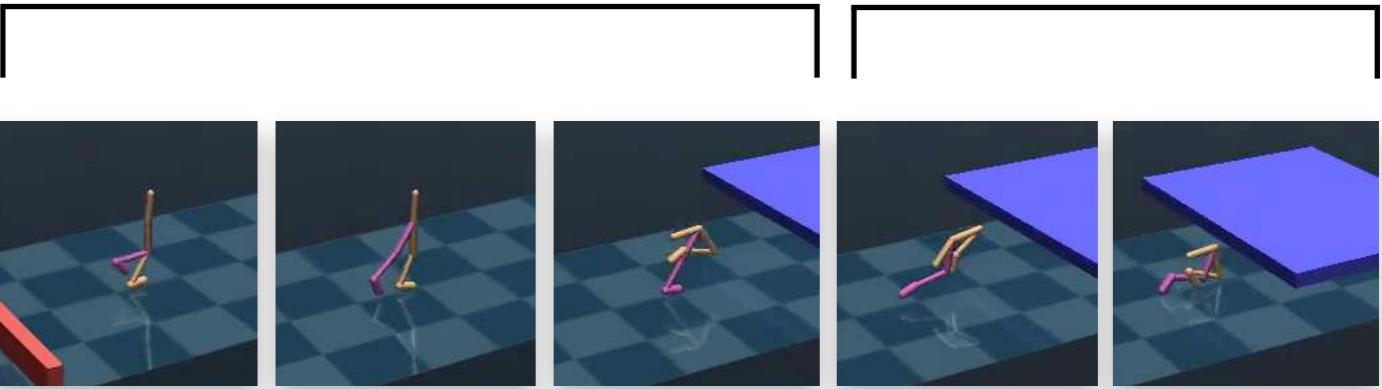
Compose Complex Skills

High-level plan

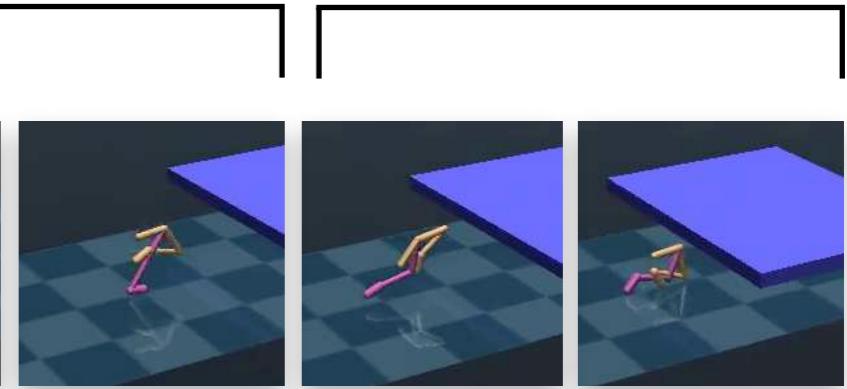
Jump



Walk



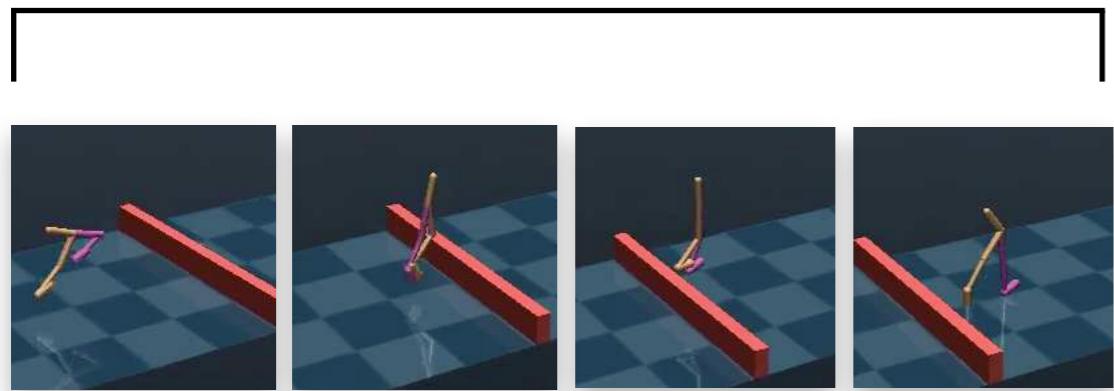
Crawl



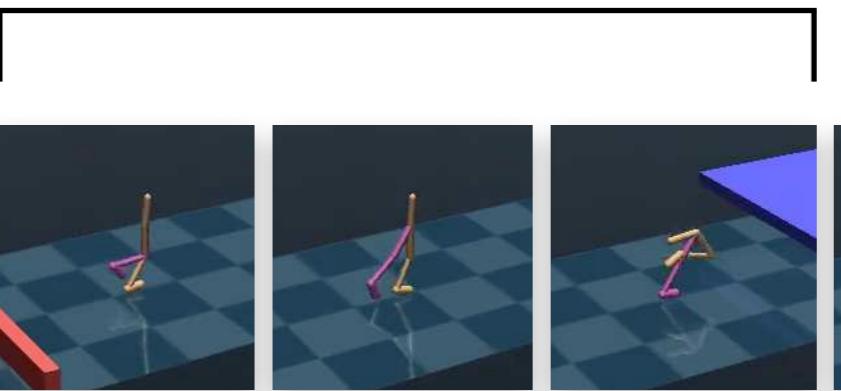
Compose Complex Skills

High-level plan

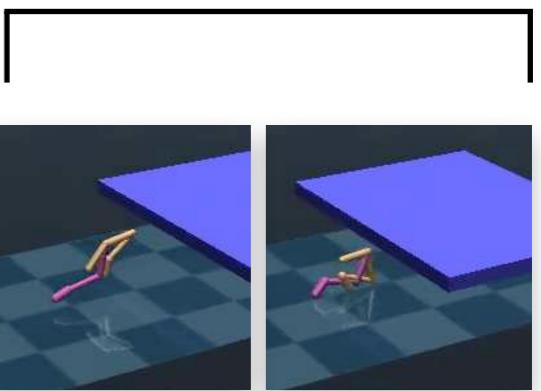
Jump



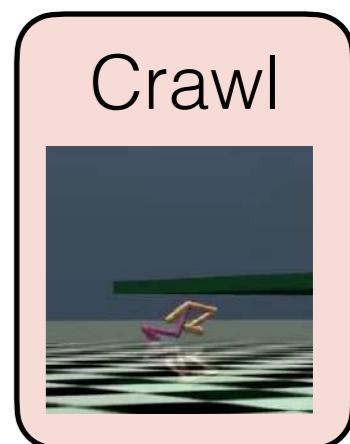
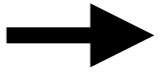
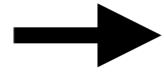
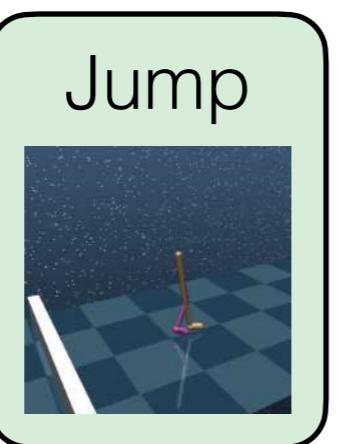
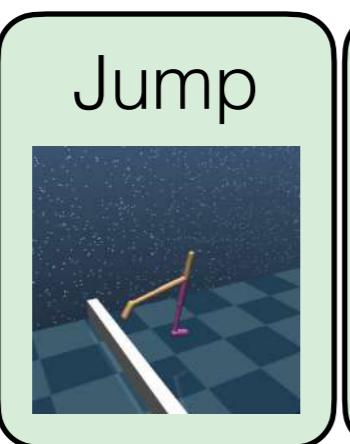
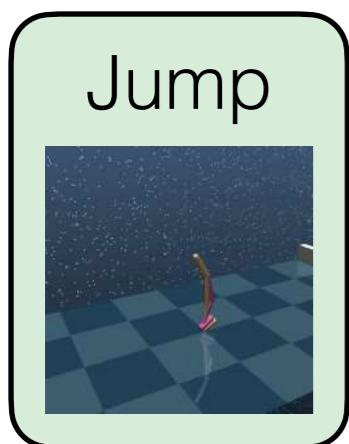
Walk



Crawl



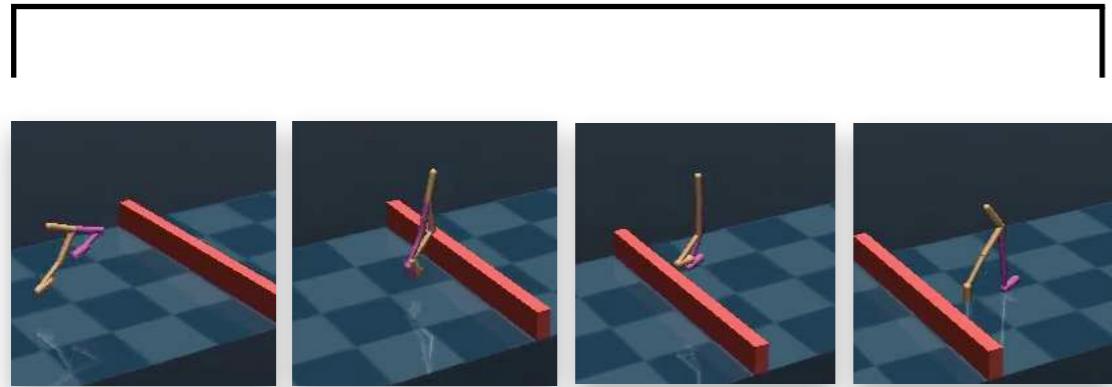
Sequentially execute corresponding policies



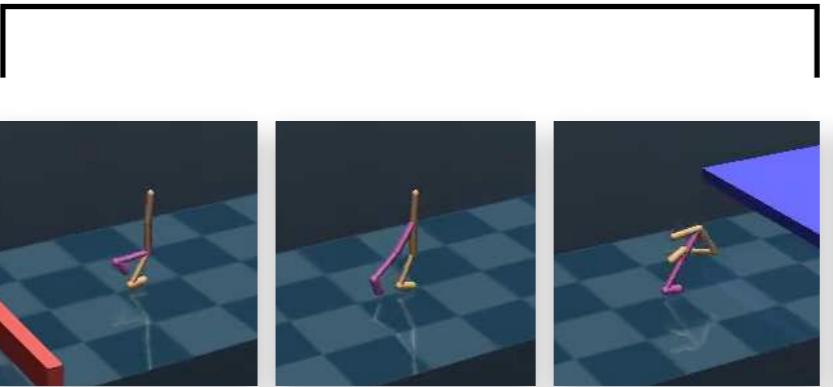
Compose Complex Skills

High-level plan

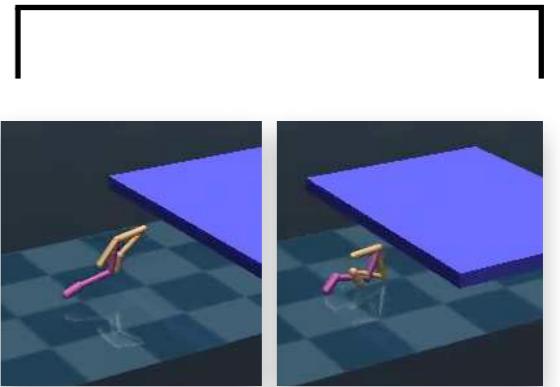
Jump



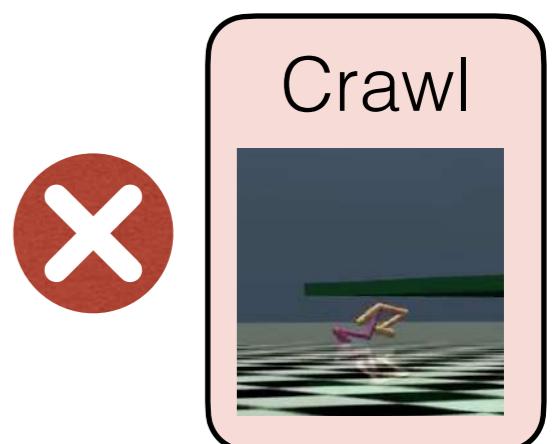
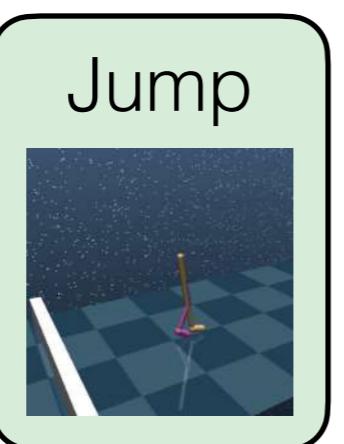
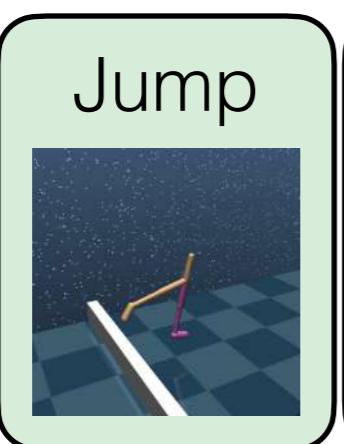
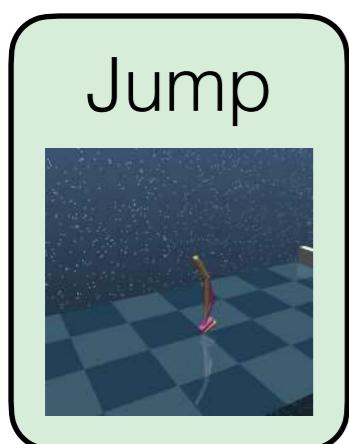
Walk



Crawl



Sequentially execute corresponding policies

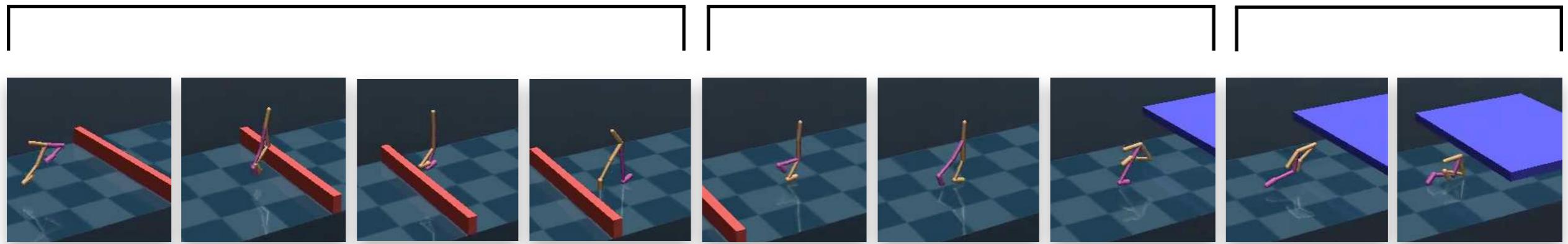


An **end state** of a previous policy might not be a good **initial state** of the following policy

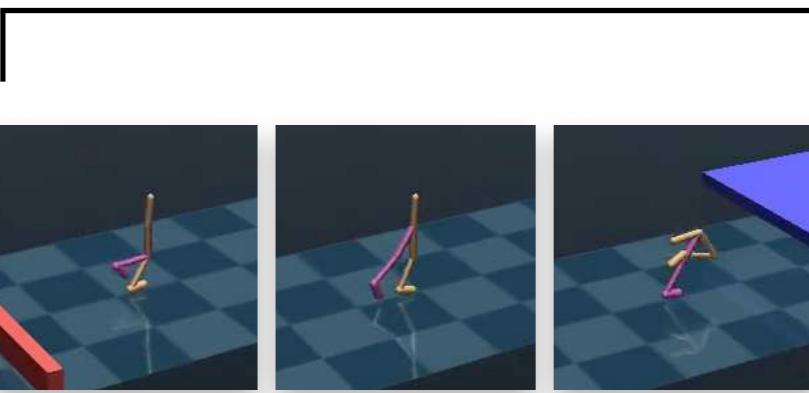
Compose Complex Skills

High-level plan

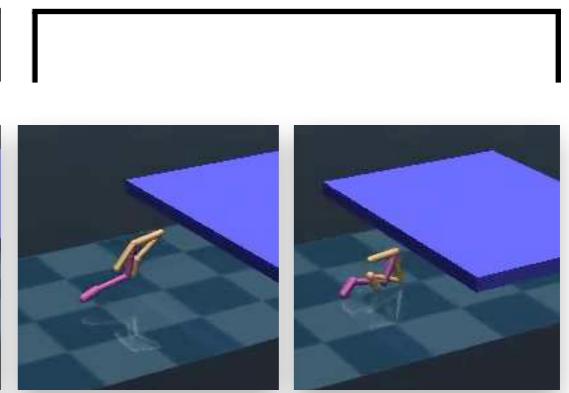
Jump



Walk



Crawl

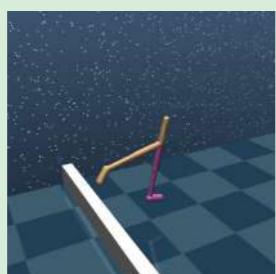


Sequentially execute corresponding policies

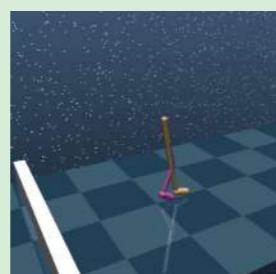
Jump



Jump



Jump



Trans
 π

Walk



Walk



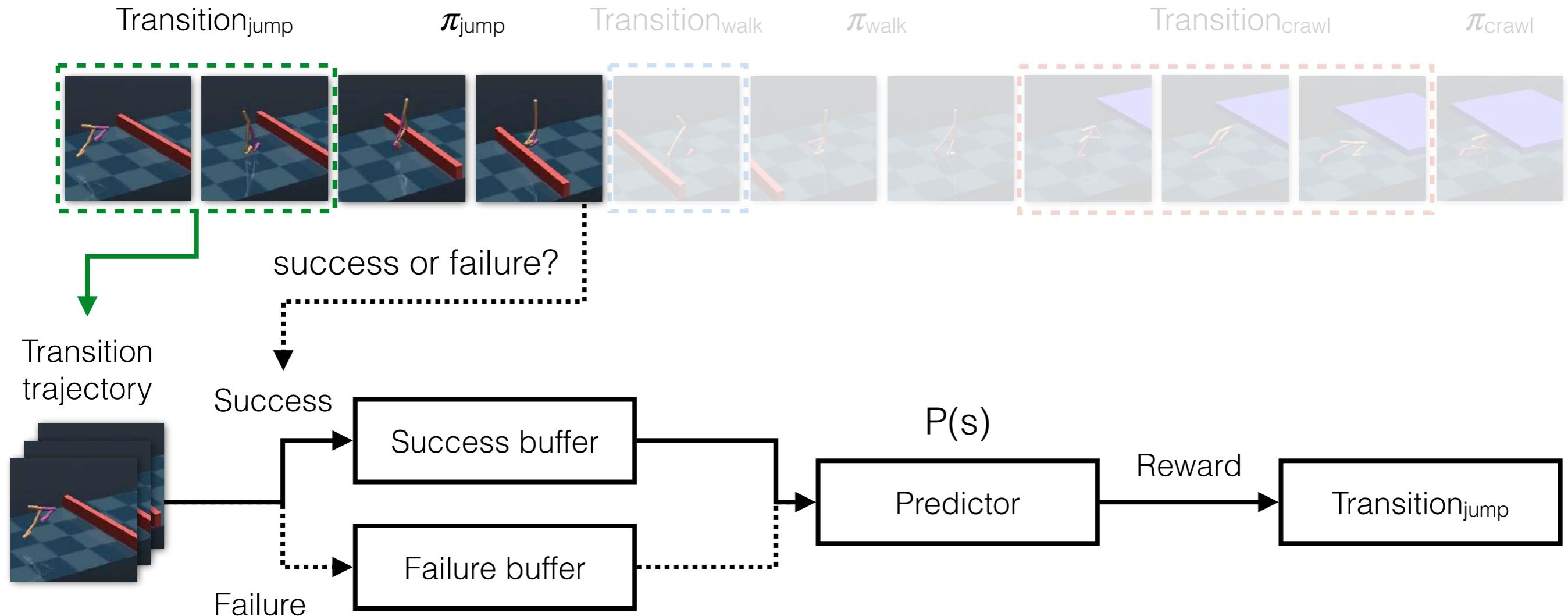
Trans
 π

Crawl



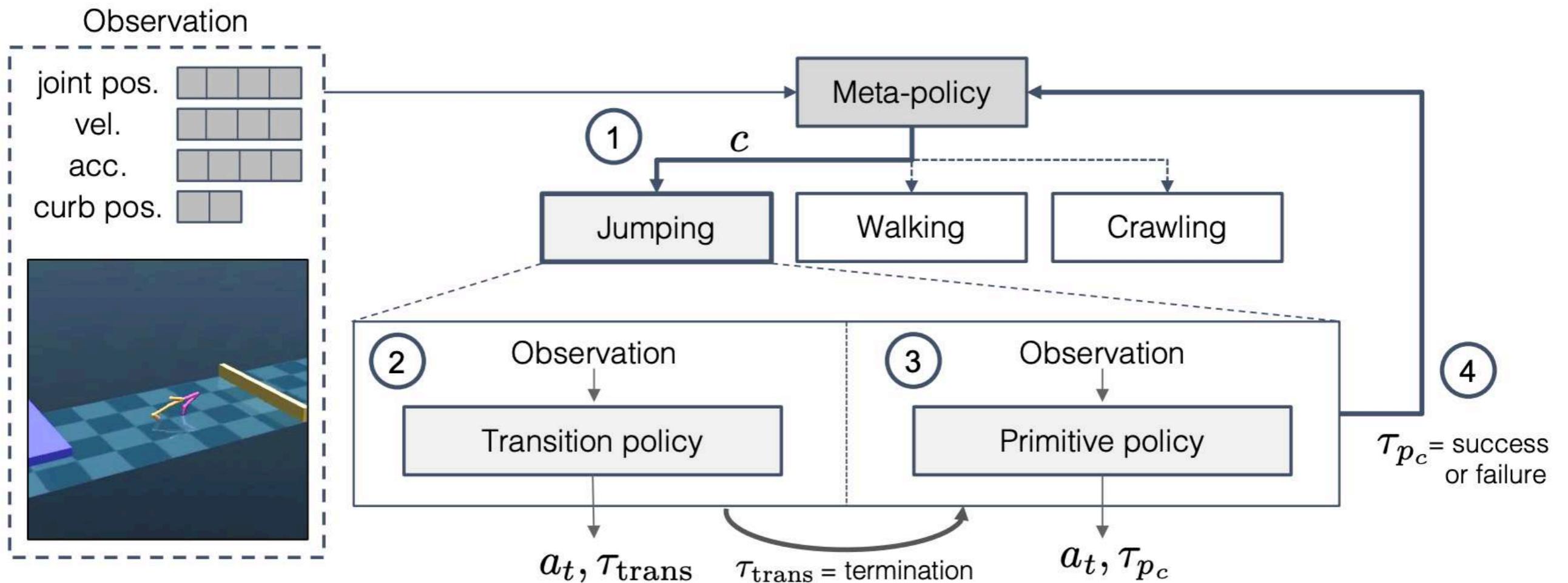
Transition policies

Learning Transition Policies



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

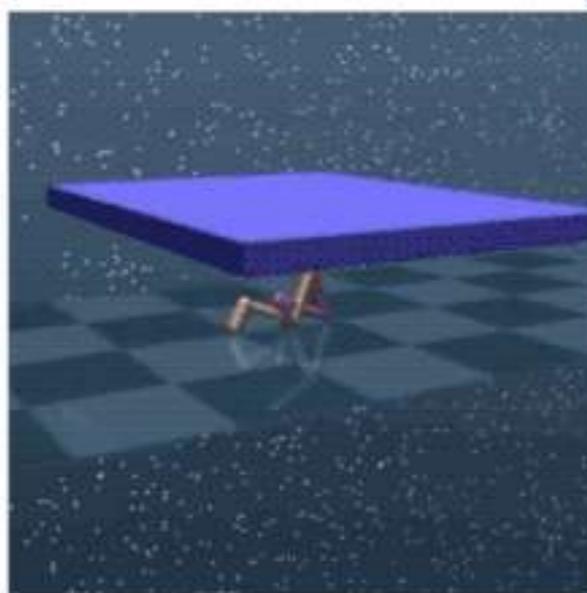
Modular Framework



Qualitative Results

Locomotion

Crawl



Transition

Walk

Walk Forward

Transition

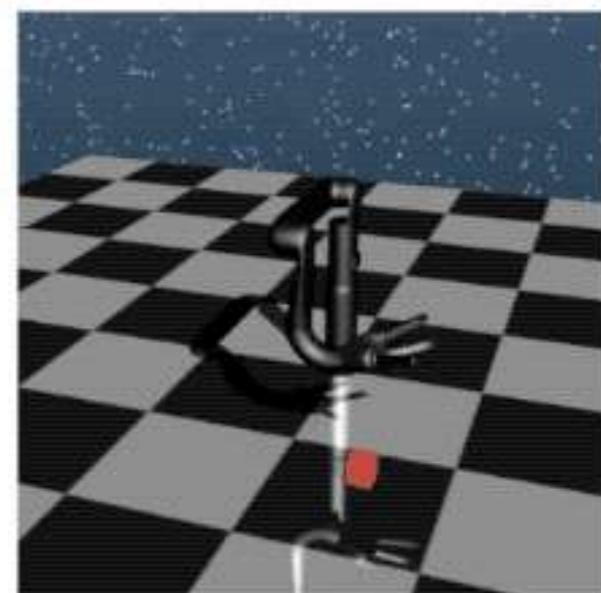
Walk Backward

Manipulation

Pick

Transition

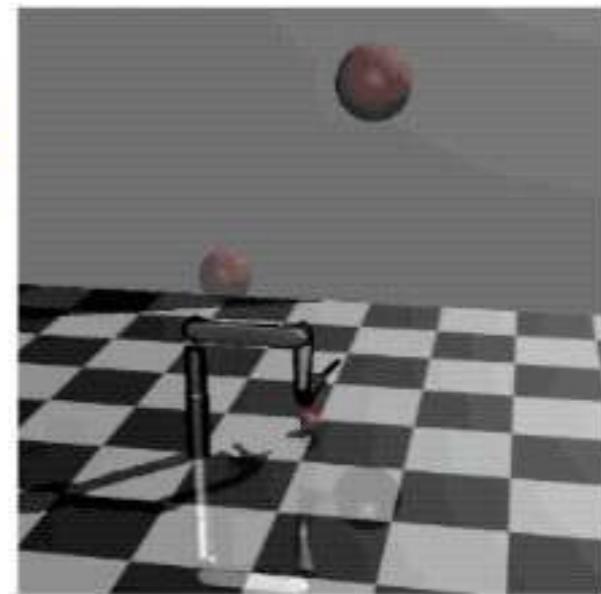
Pick



Toss

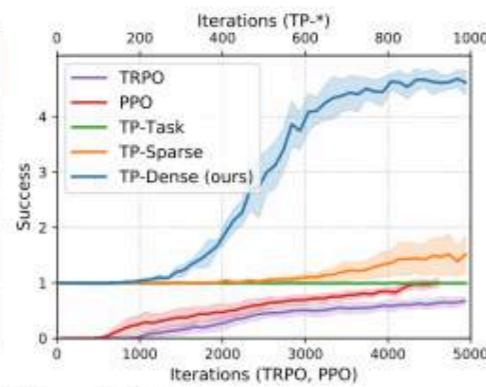
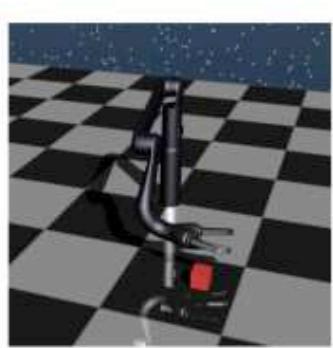
Transition

Hit

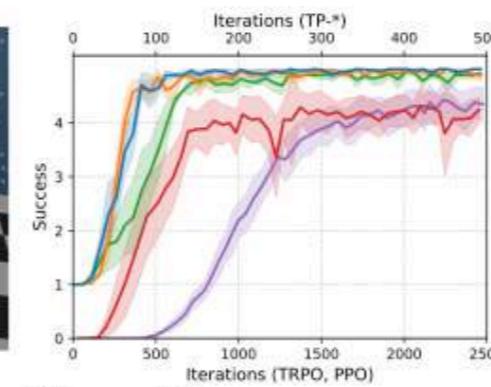


Quantitative Results - Sample Efficiency

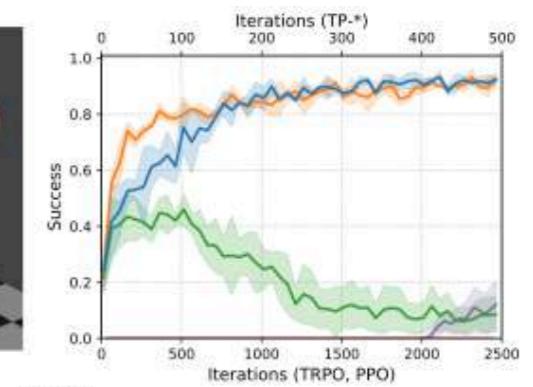
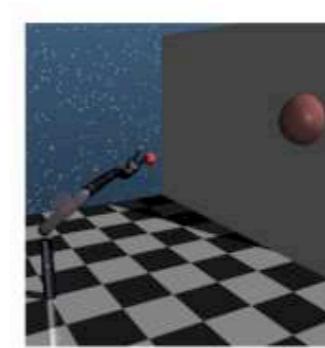
Manipulation



(a) Repetitive picking up

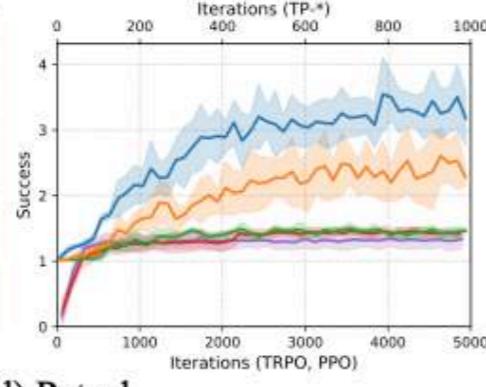
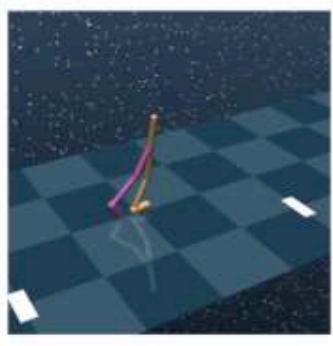


(b) Repetitive catching

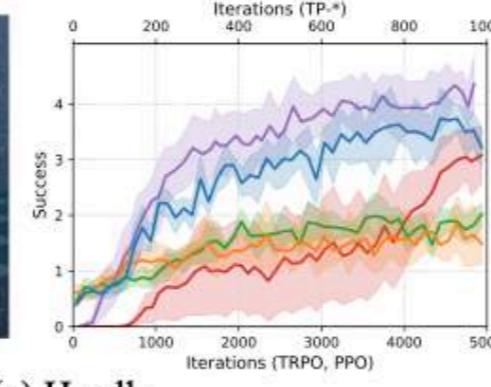
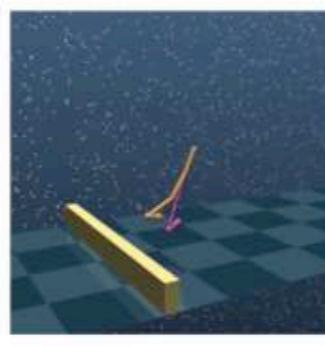


(c) Serve

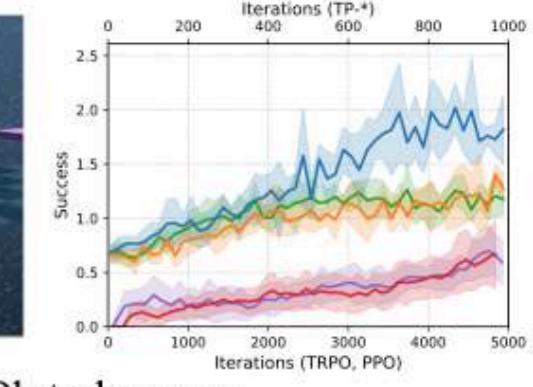
Locomotion



(d) Patrol



(e) Hurdle

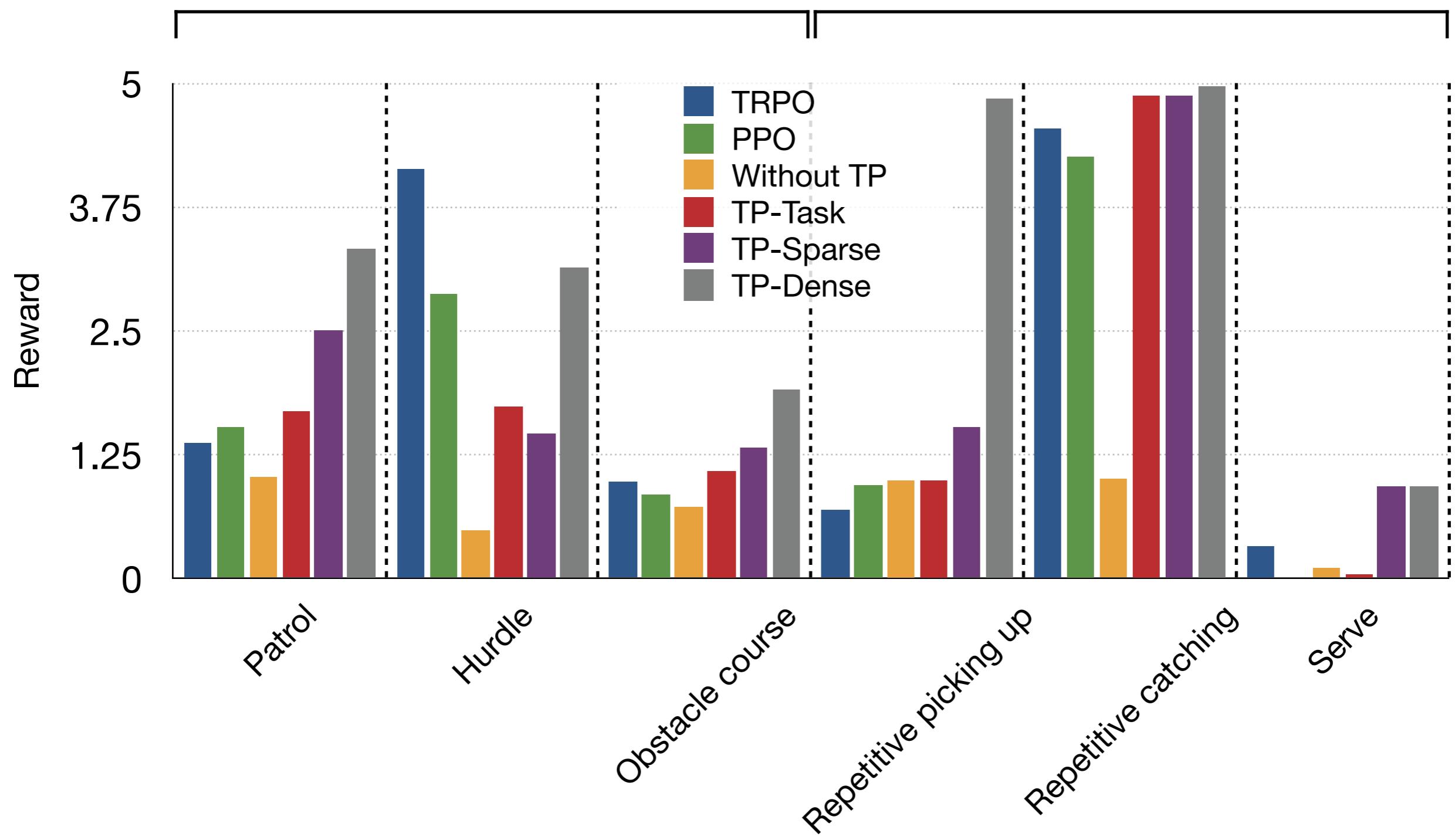


(f) Obstacle course

Quantitative Results - Reward

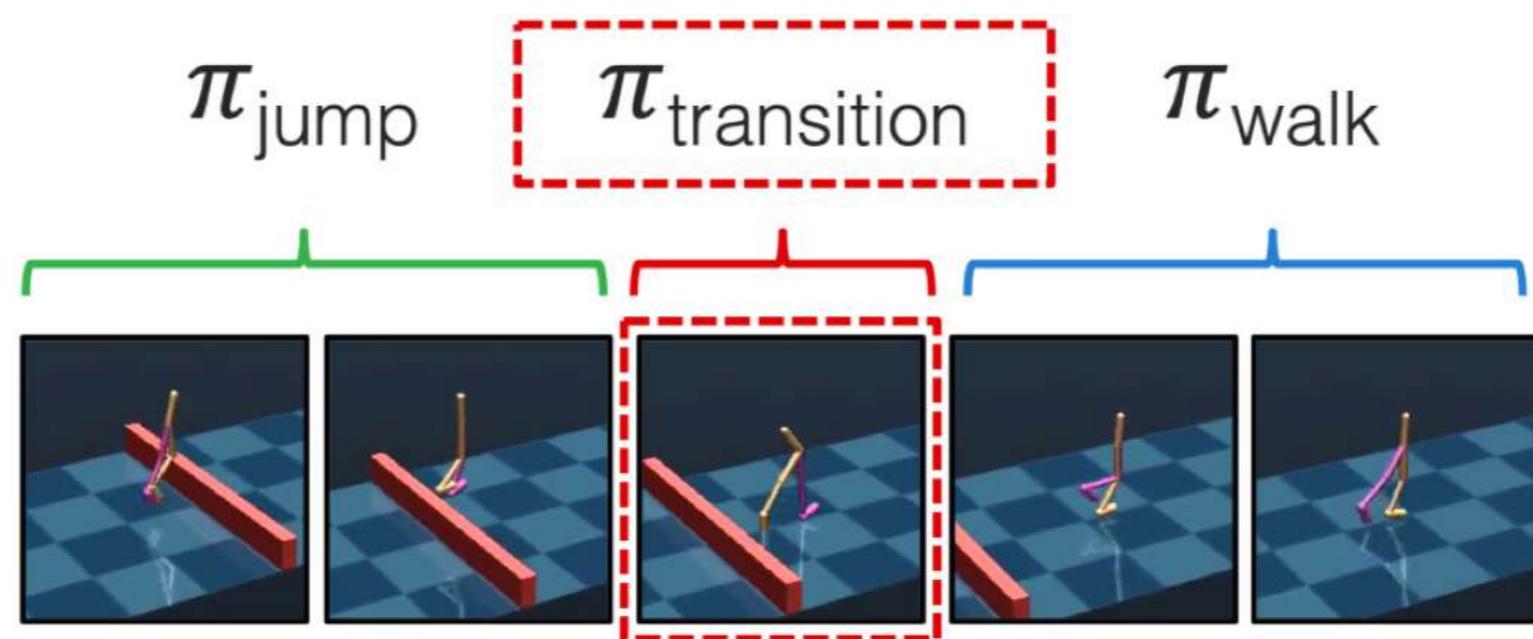
Locomotion

Manipulation

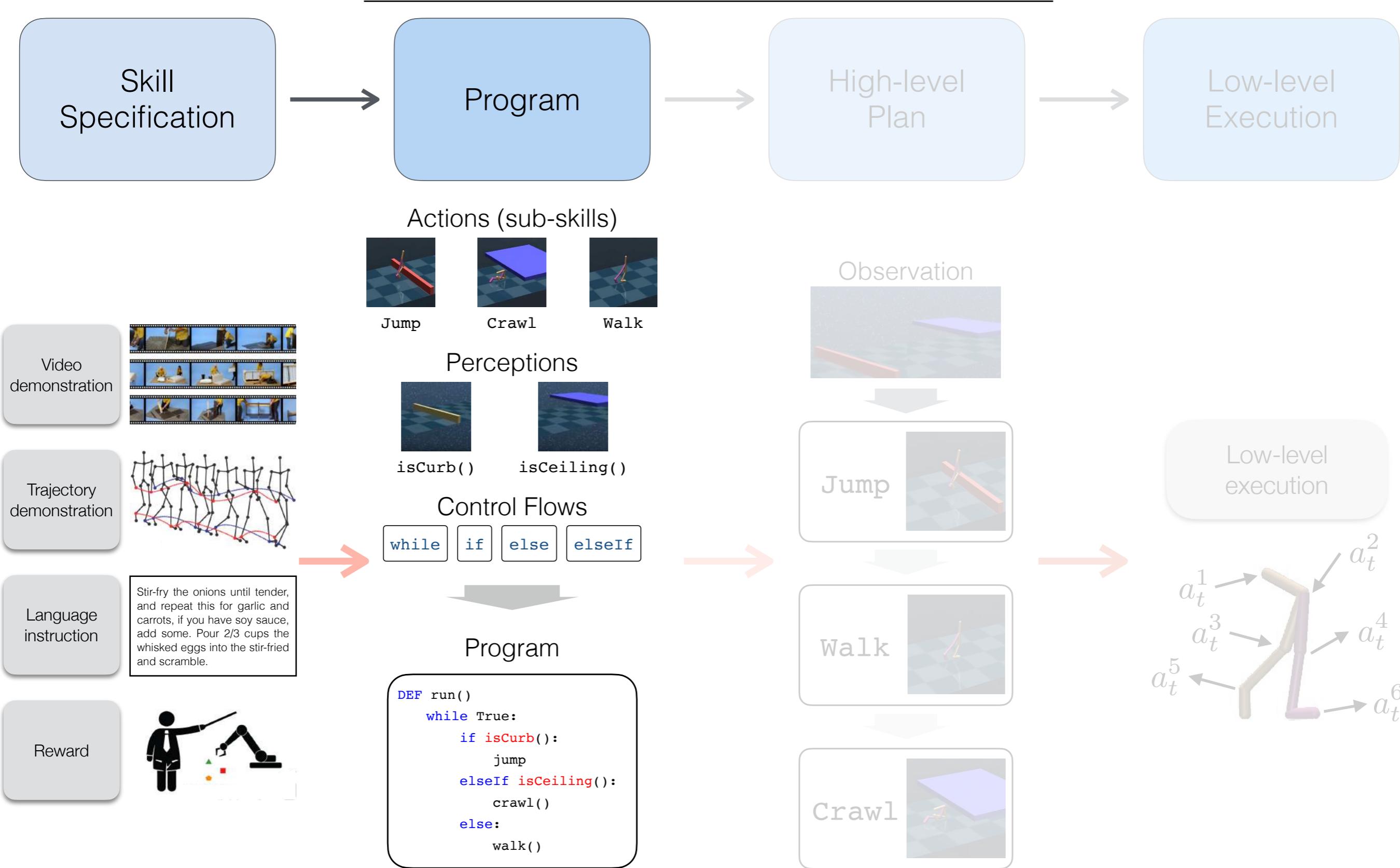


Takeaway

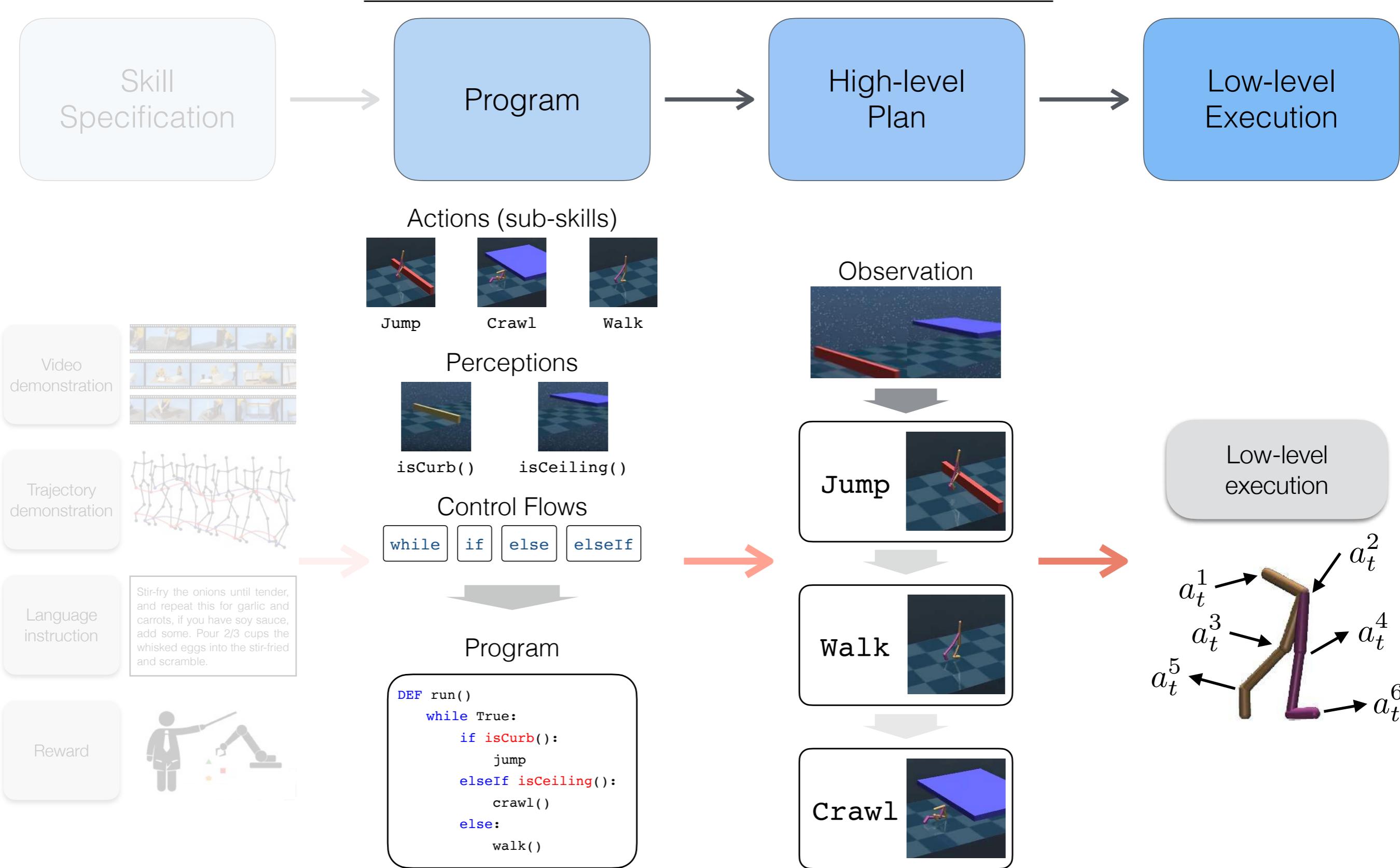
- Learning transition policies to smoothly compose learned skills



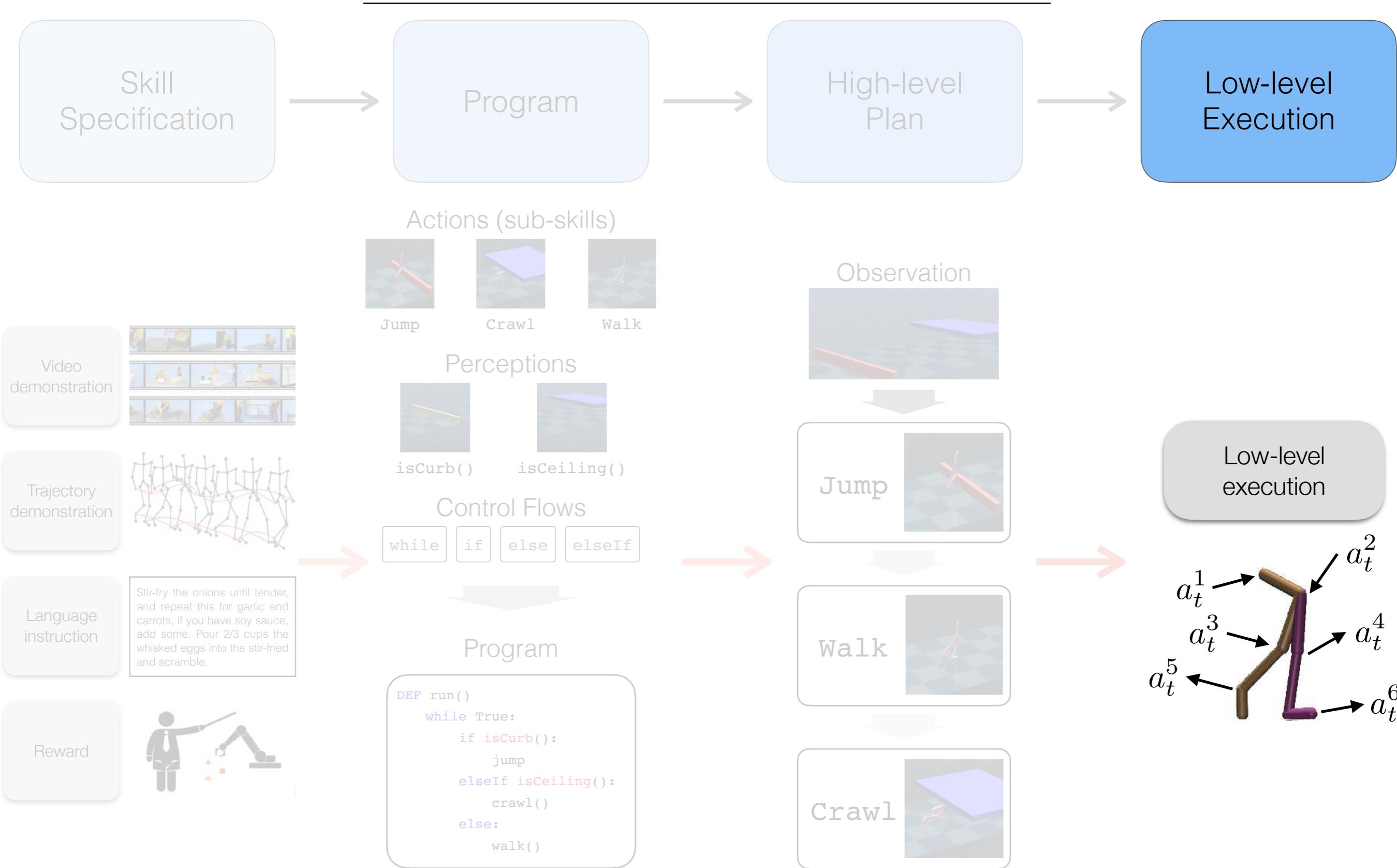
Program Inference



Task Execution

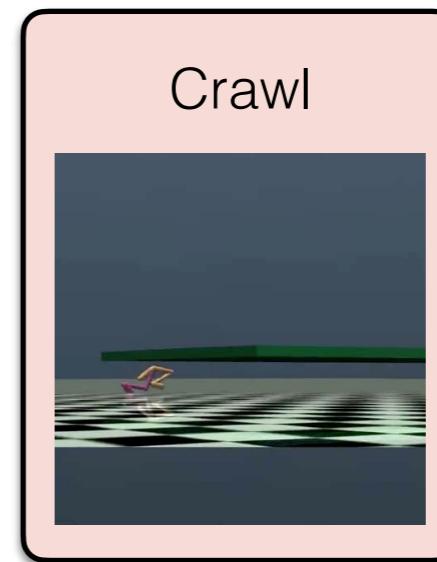


Primitive Skill Acquisition



Primitive Skill Acquisition

- Goal: acquire a diverse set of primitive skills efficiently



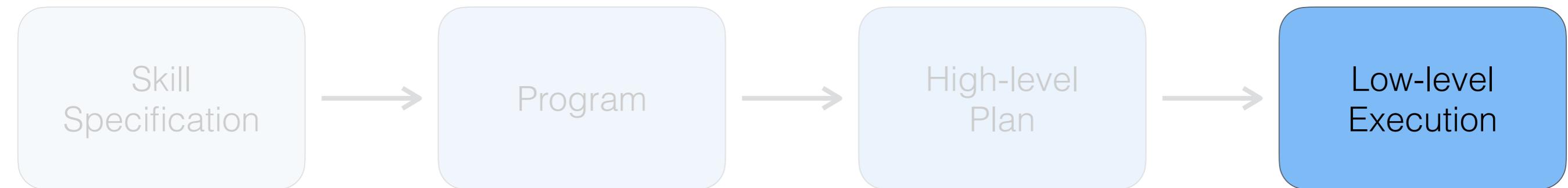
- Key directions

Meta-learning
Meta-RL

Learning from
experts

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

NeurIPS 2019 (Spotlight)



Risto Vuorio

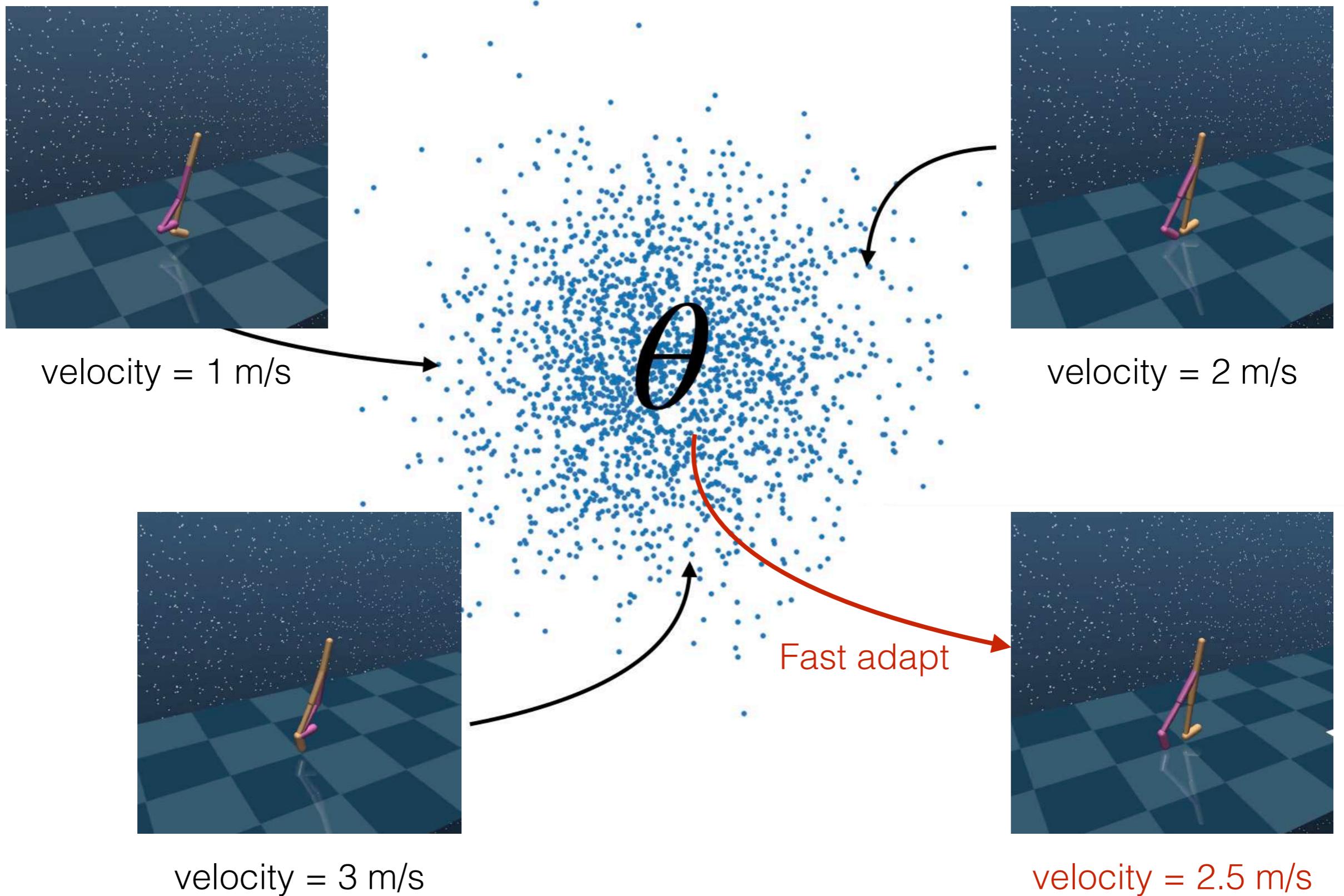


Hexiang Hu

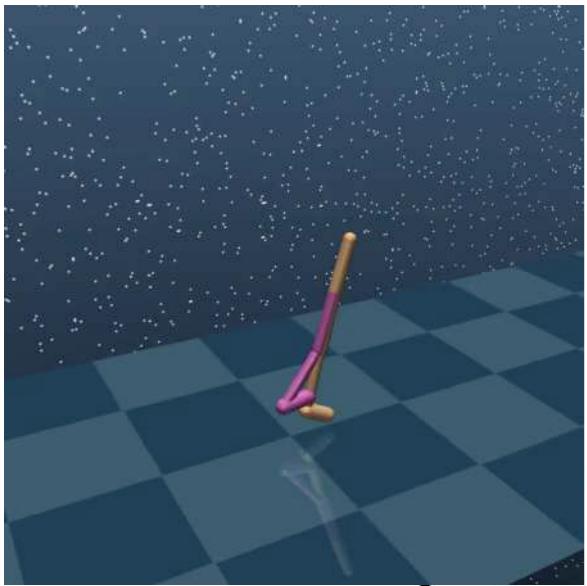


Joseph J. Lim

Model-Agnostic Meta-Learning (MAML)

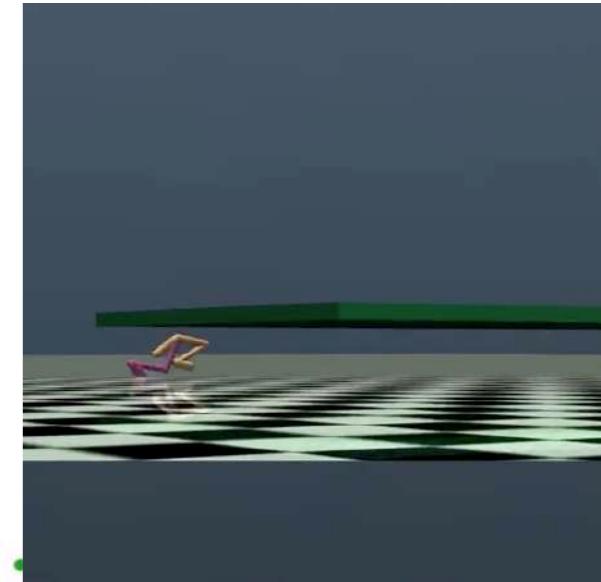


Walk

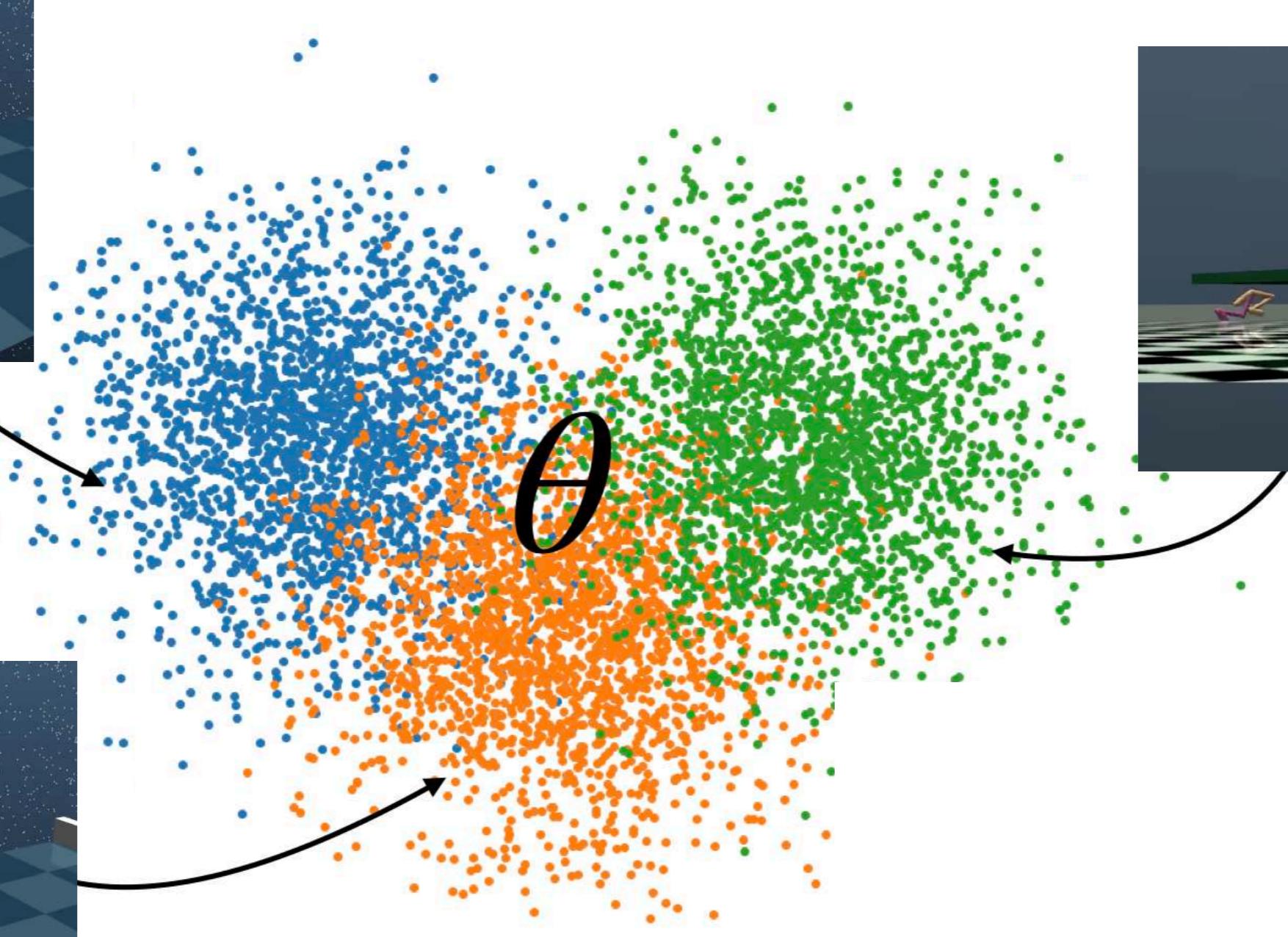
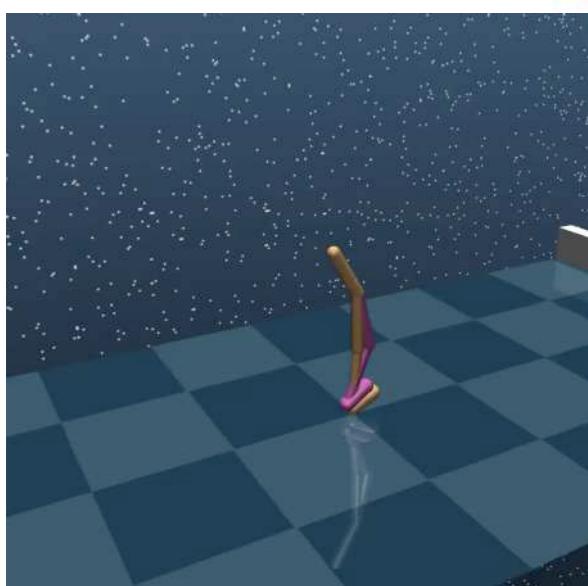


Multimodal Task Distribution

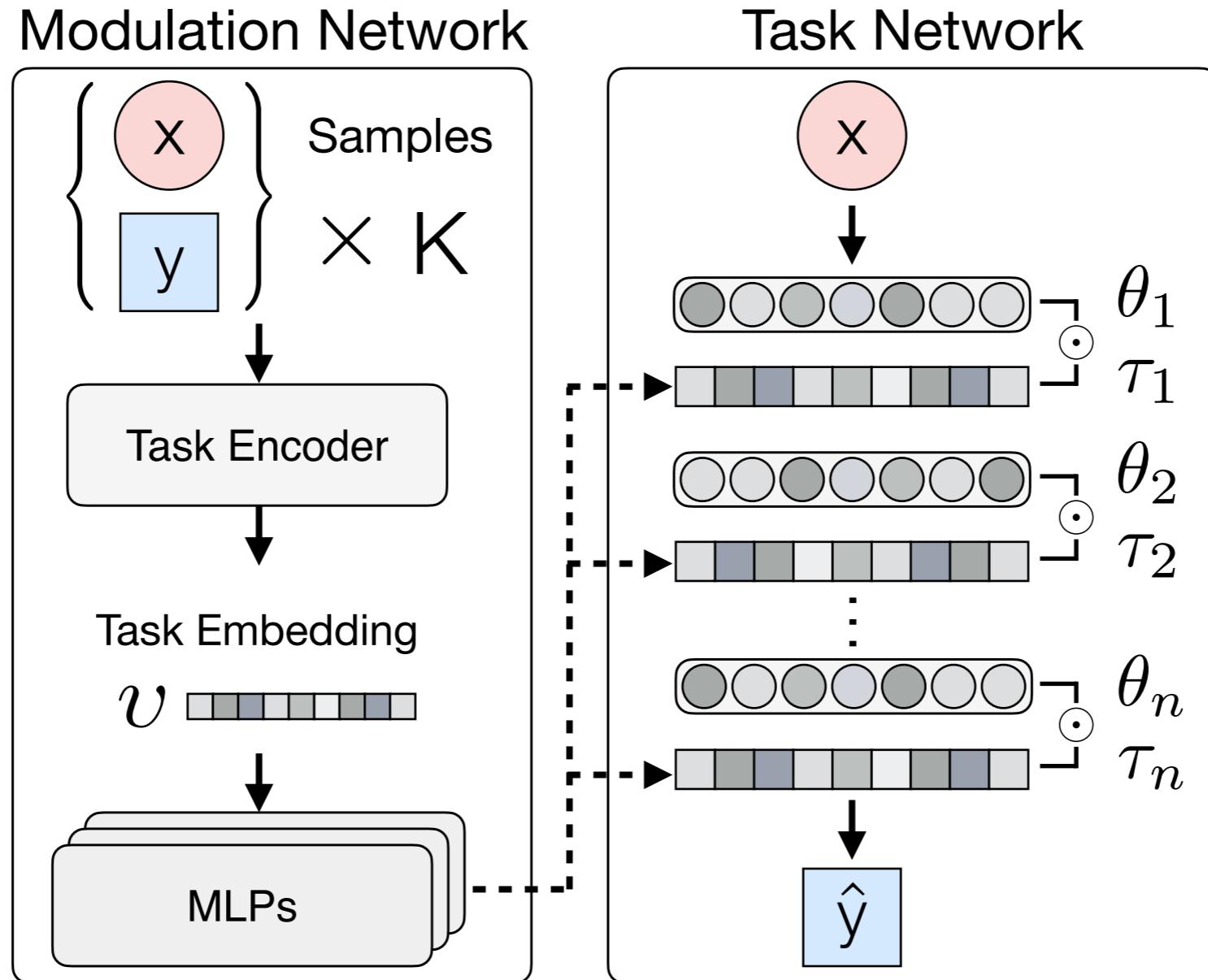
Crawl



Jump



Multimodal Model-Agnostic Meta-Learning (MMAML)



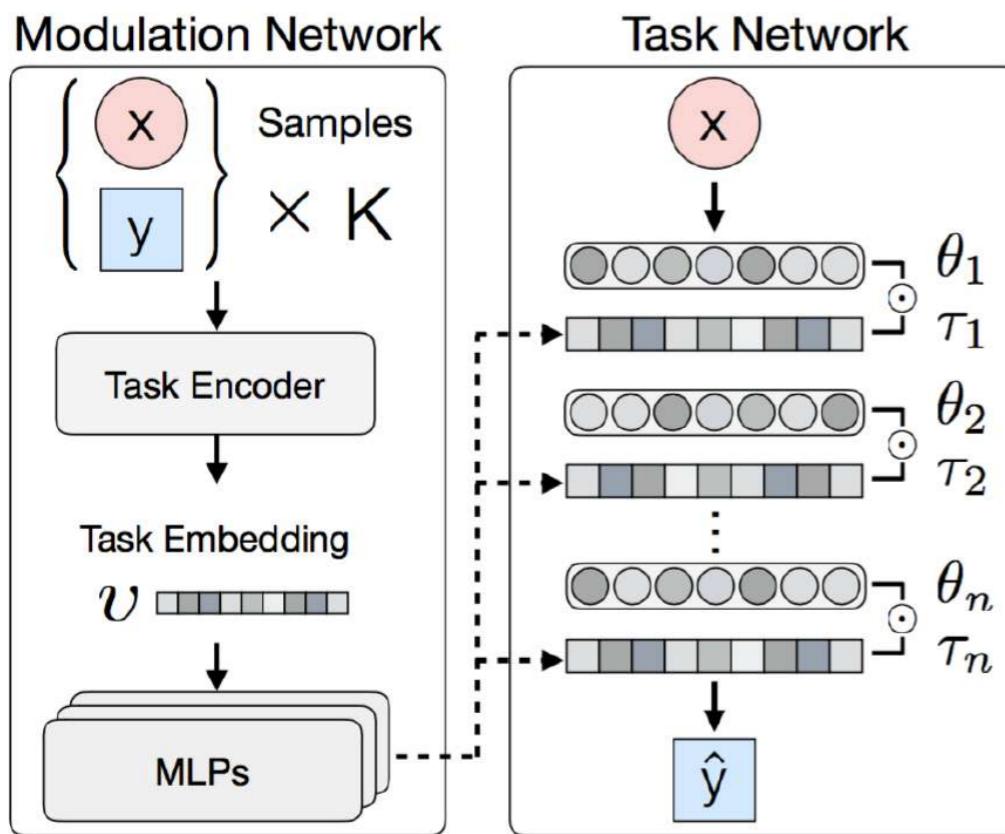
Training Algorithm

Outer loop

- Task Encoder: produce the task embedding ω_g
- MLPs: modulate the task network blocks ω_h

Inner loop

- Task network: fast adapt through gradient updates θ



Algorithm 1 MMAML META-TRAINING PROCEDURE.

```

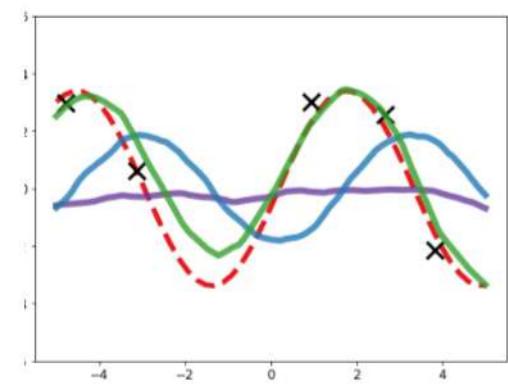
1: Input: Task distribution  $P(\mathcal{T})$ , Hyper-parameters  $\alpha$  and  $\beta$ 
2: Randomly initialize  $\theta$  and  $\omega$ .
3: while not DONE do
4:   Sample batches of tasks  $\mathcal{T}_j \sim P(\mathcal{T})$ 
5:   for all j do
6:     Infer  $v = h(\{x, y\}_K; \omega_h)$  with K samples 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

```

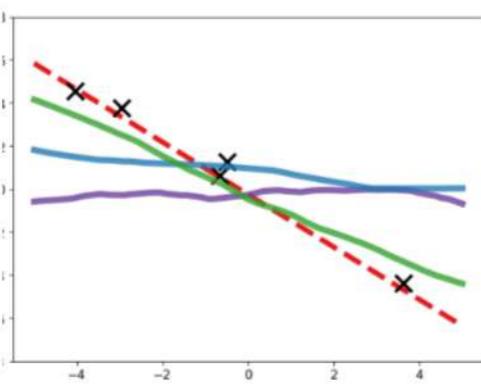
Regression



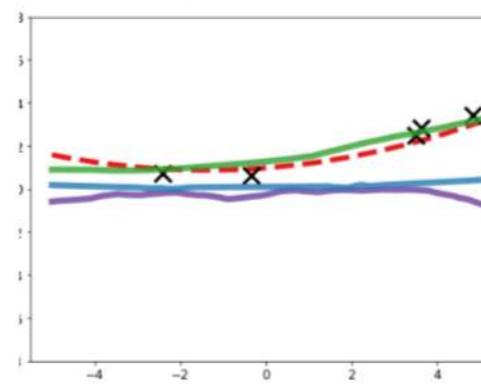
Sinusoidal



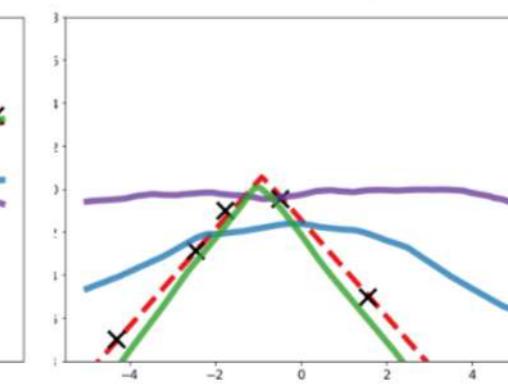
Linear



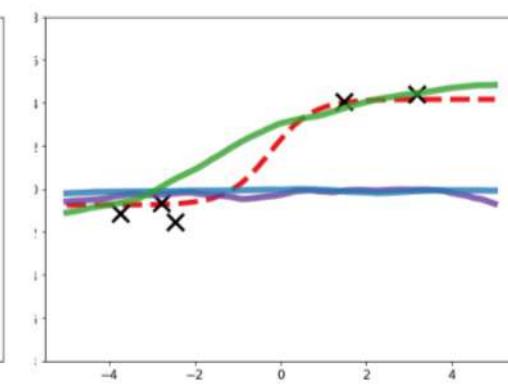
Quadratic



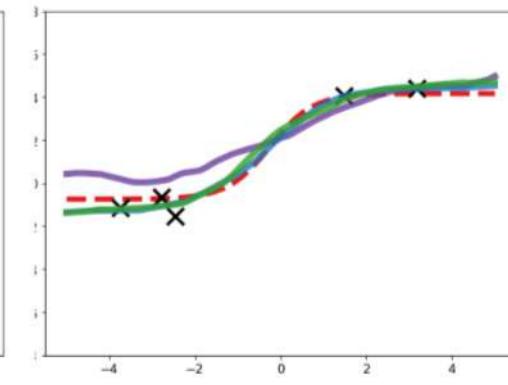
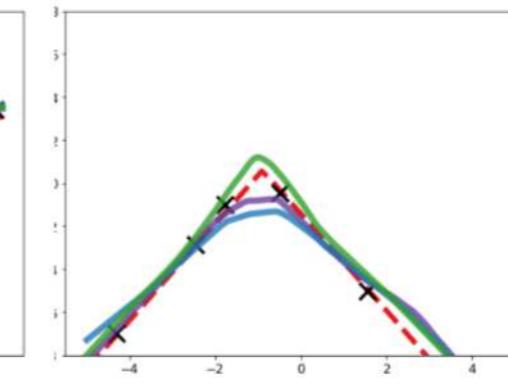
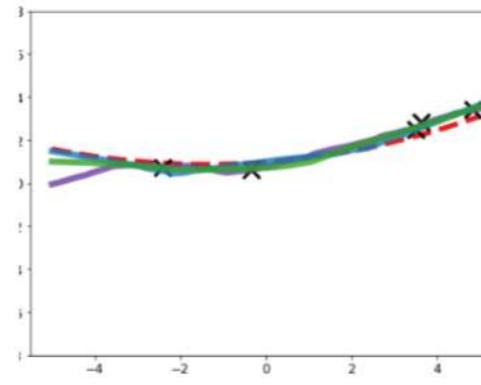
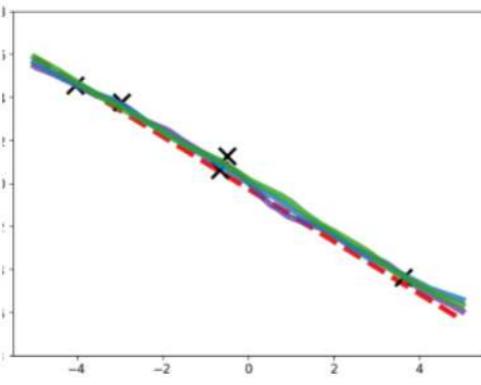
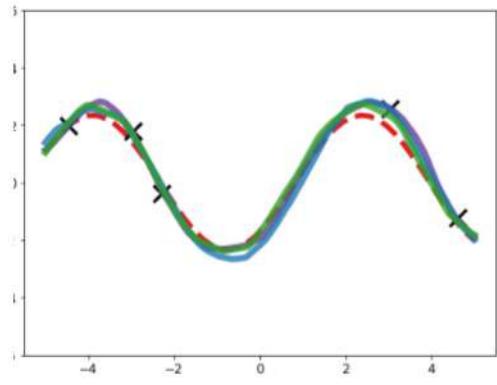
Transformed ℓ_1 Norm



Tanh



(a) MMAML post modulation vs. other prior models

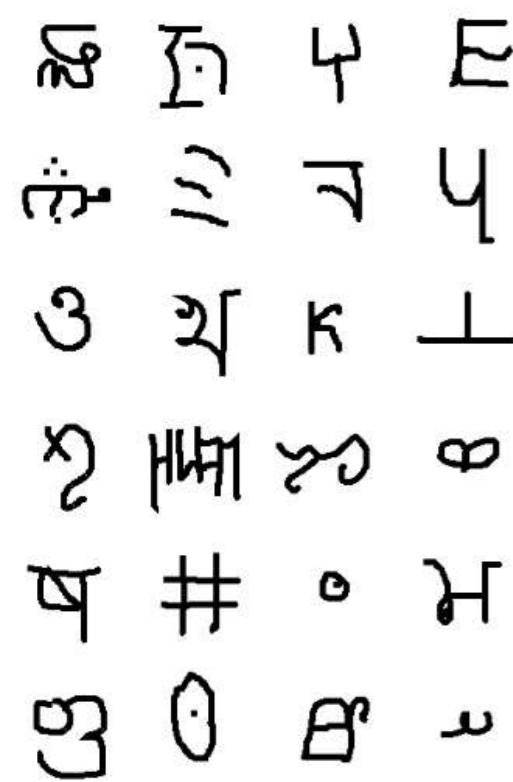


(b) MMAML post adaptation vs. other posterior models

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	-
Ours: MMAML (Softmax)	1.548	0.361	2.213	0.444	2.421	0.939
Ours: MMAML (FiLM)	2.421	0.336	1.923	0.444	2.166	0.868

Image Classification

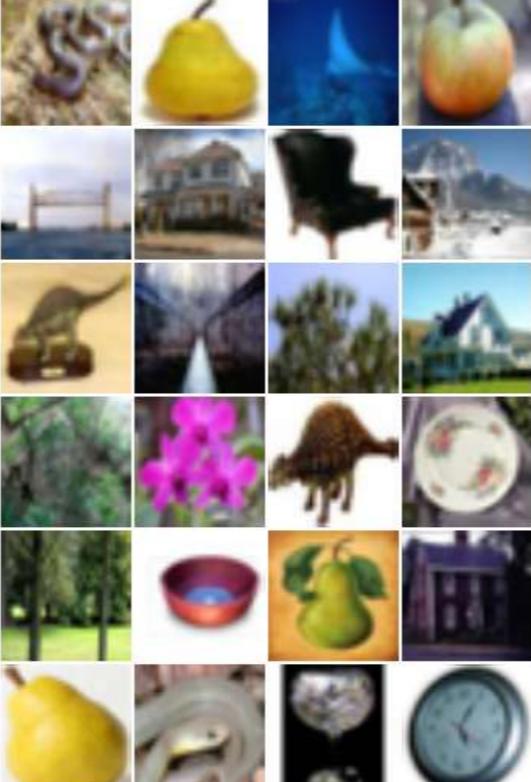
(a) Omniglot



(b) Mini-ImageNet



(c) FC100



(d) CUB



(e) Aircraft

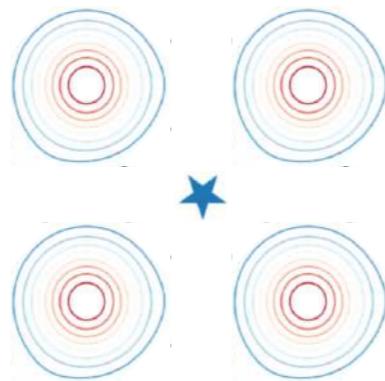


Method & Setup	2 Modes			3 Modes			5 Modes			
	Way	5-way	20-way	Way	5-way	20-way	Way	5-way	20-way	
Shot	1-shot	5-shot	1-shot	1-shot	5-shot	1-shot	1-shot	5-shot	1-shot	3-shot
MAML [1]	66.80%	77.79%	44.69%	54.55%	67.97%	28.22%	66.00%	70.87%	44.69%	56.57%
Multi-MAML	66.85%	73.07%	53.15%	55.90%	62.20%	39.77%	71.94%	76.94%	58.66%	61.11%
MMAML (ours)	69.93%	78.73%	47.80%	57.47%	70.15%	36.27%	72.02%	78.13%	60.13%	63.52%

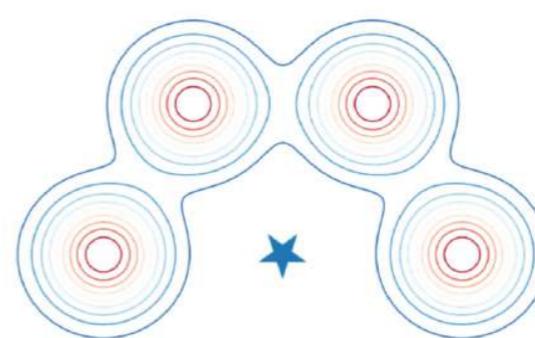
Reinforcement Learning

Goal modes

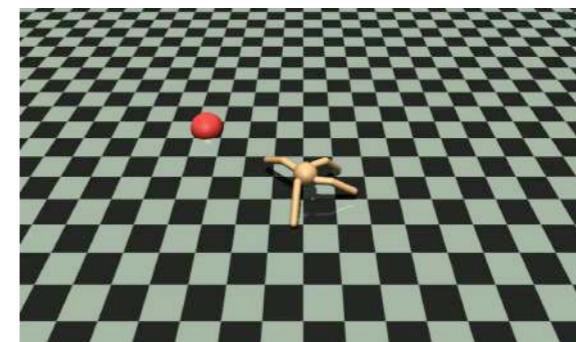
Reacher



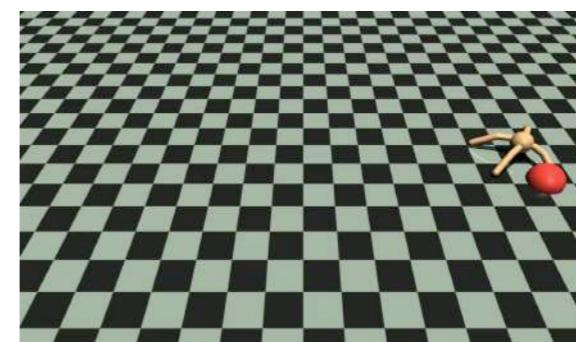
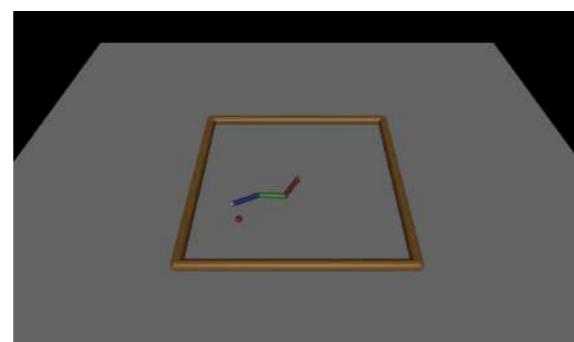
Ant



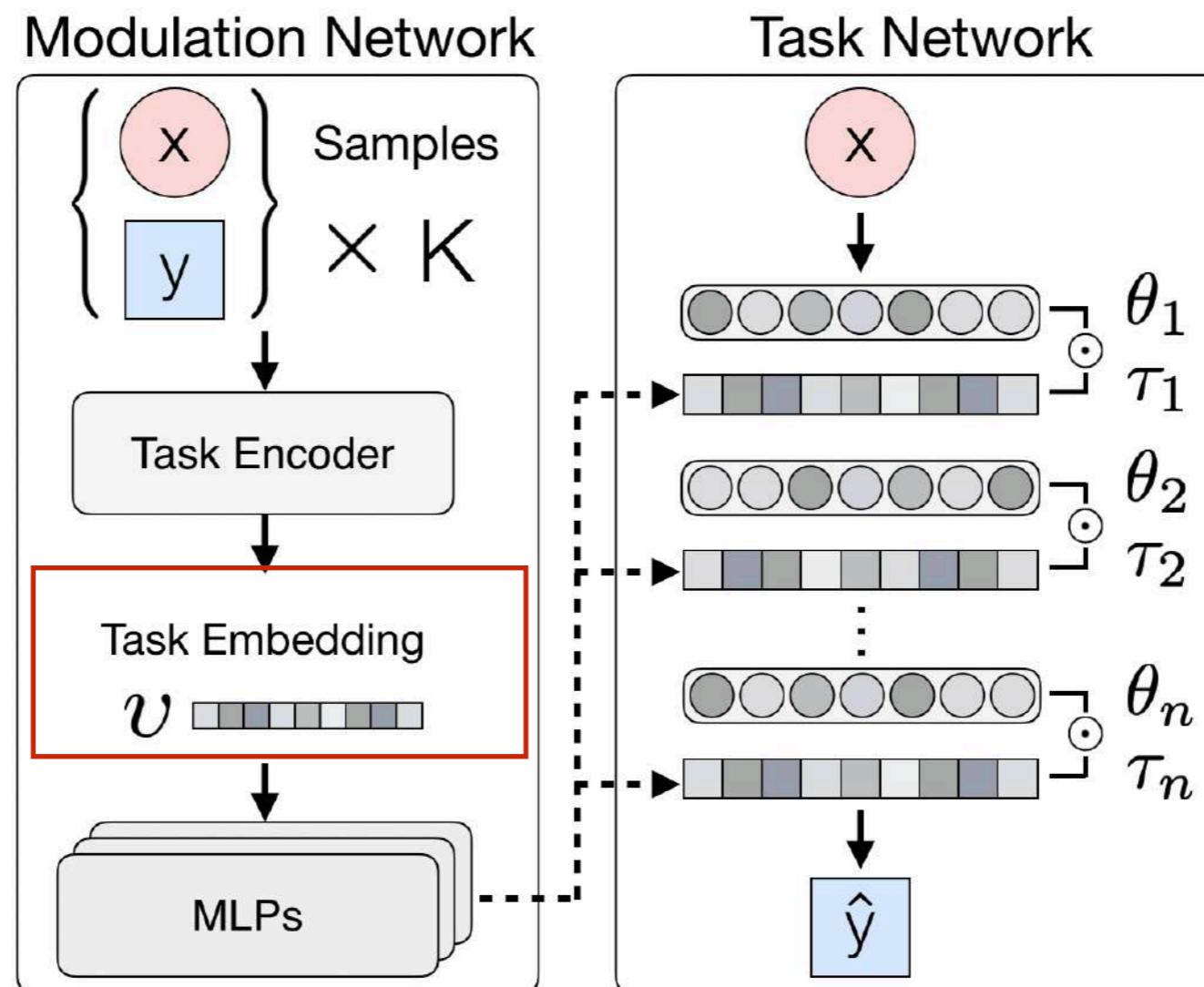
ProMP



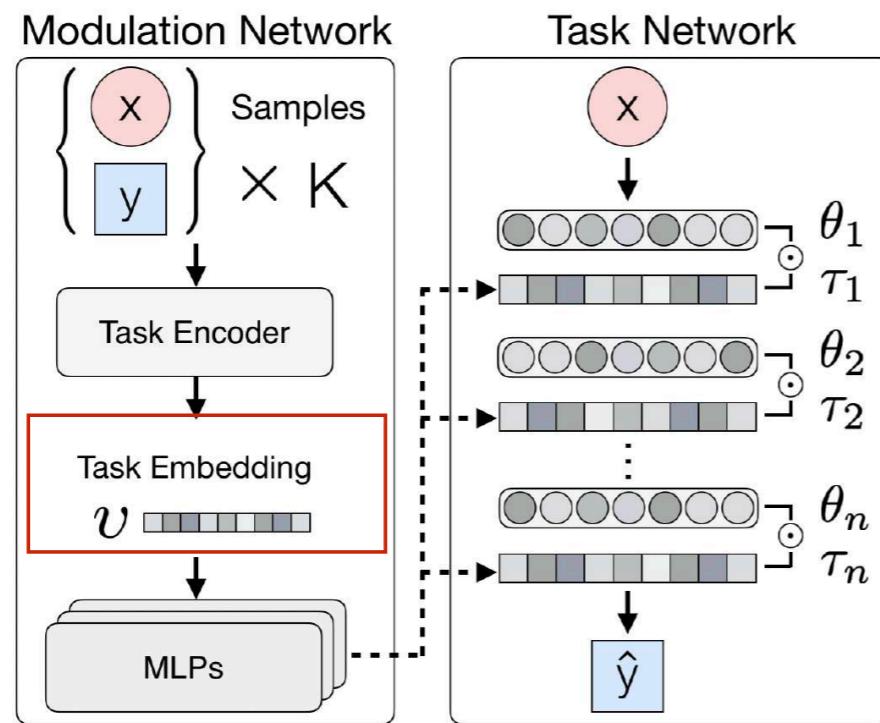
Ours



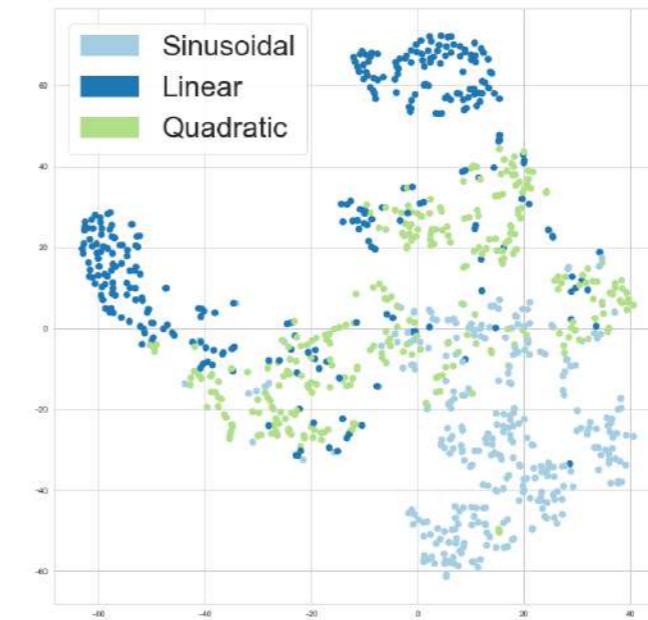
Learned Task Embedding (tSNE plot)



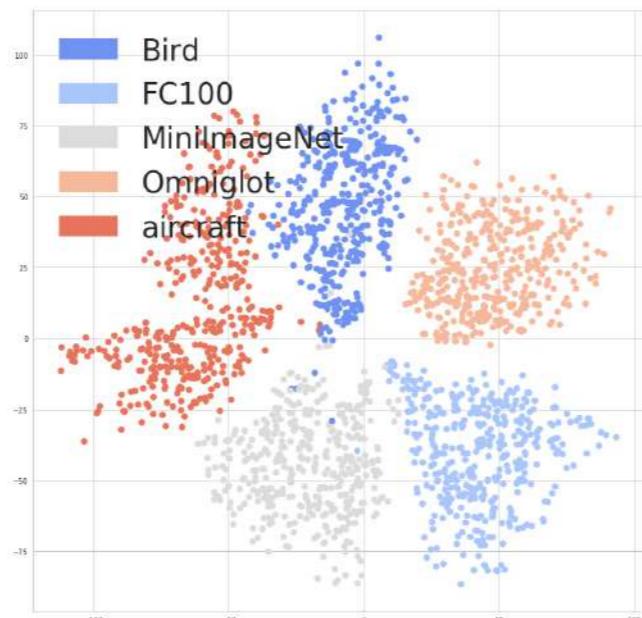
Learned Task Embedding (tSNE plot)



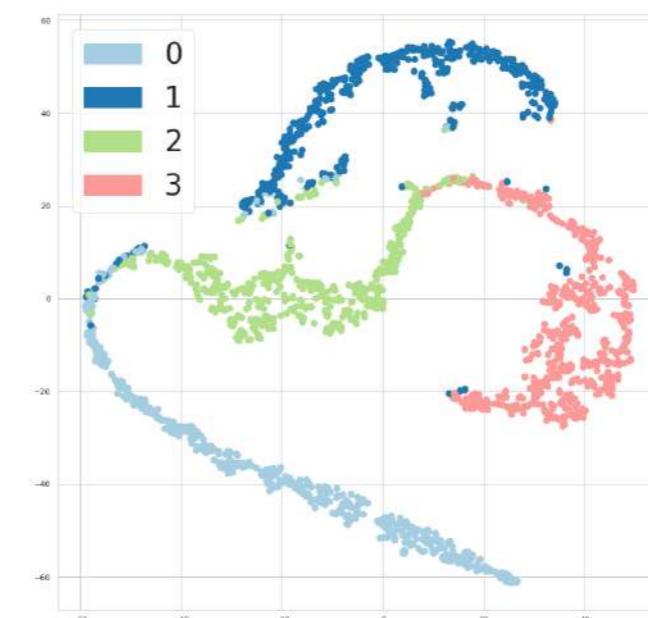
3-mode Regression



5-mode Classification

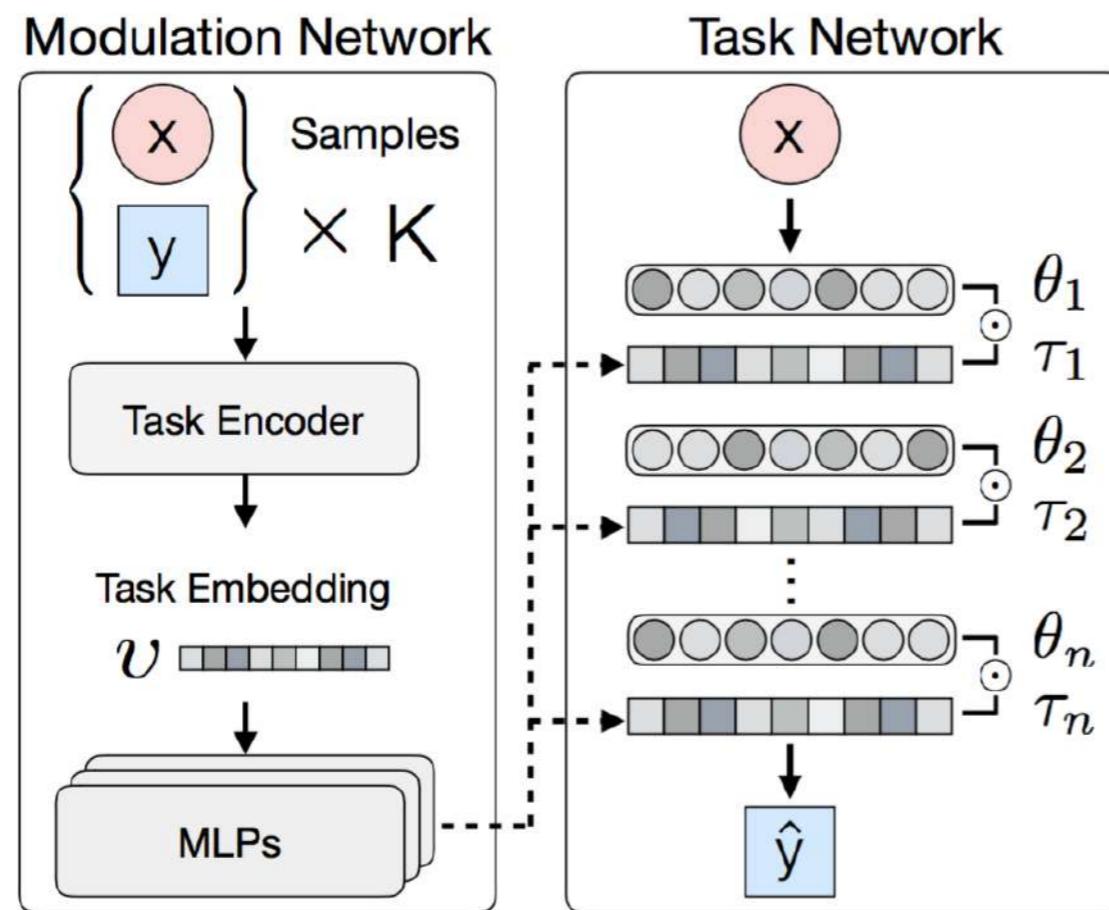


4-mode Reacher



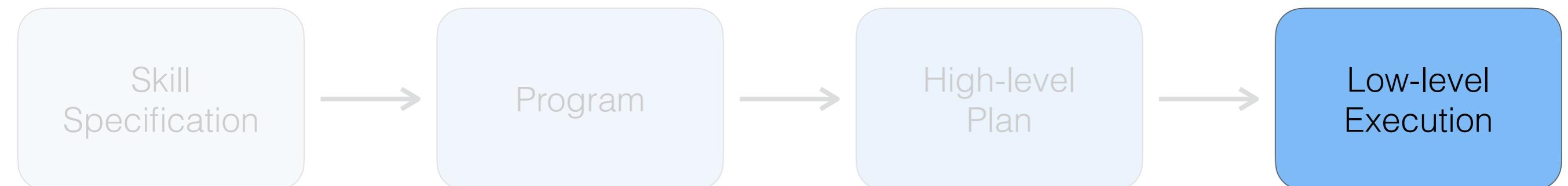
Takeaway

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



Skill-based Meta-Reinforcement Learning

ICLR 2022



Taewook Nam



Karl Pertsch

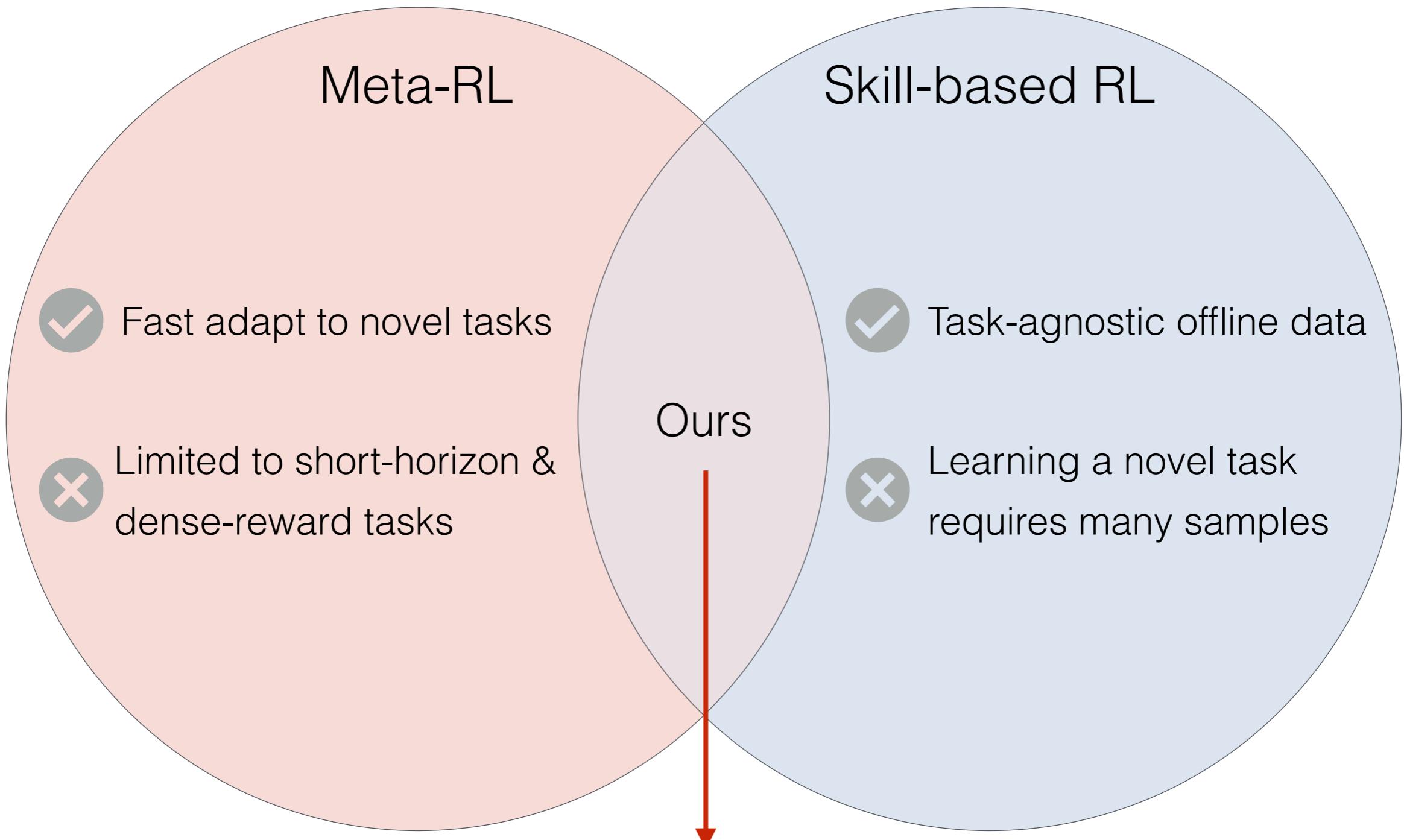


Sung Ju Hwang



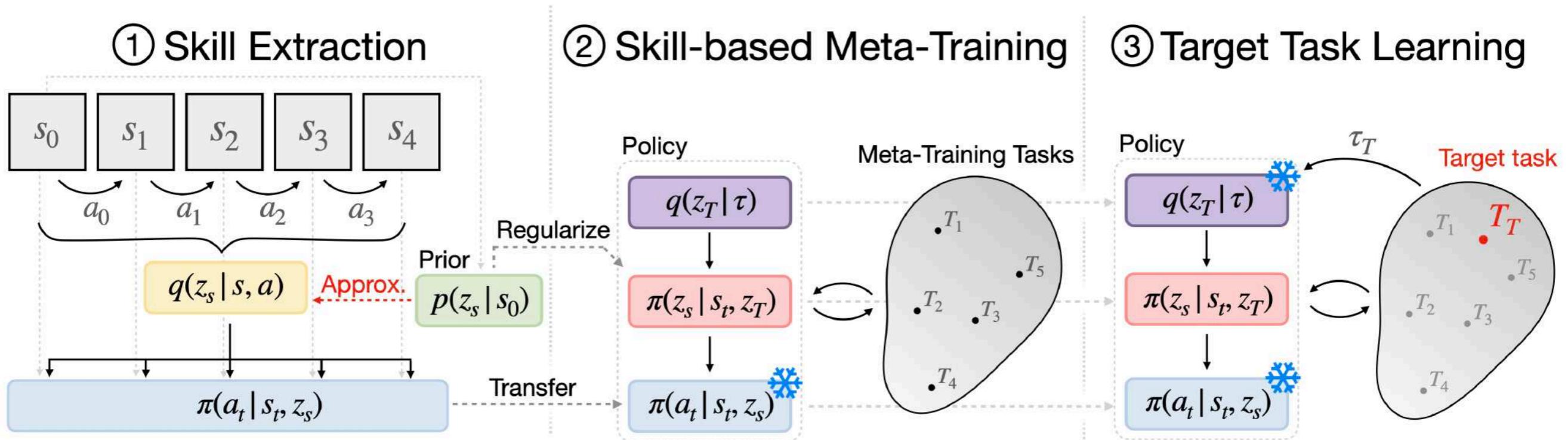
Joseph J. Lim

Meta-RL with Skills

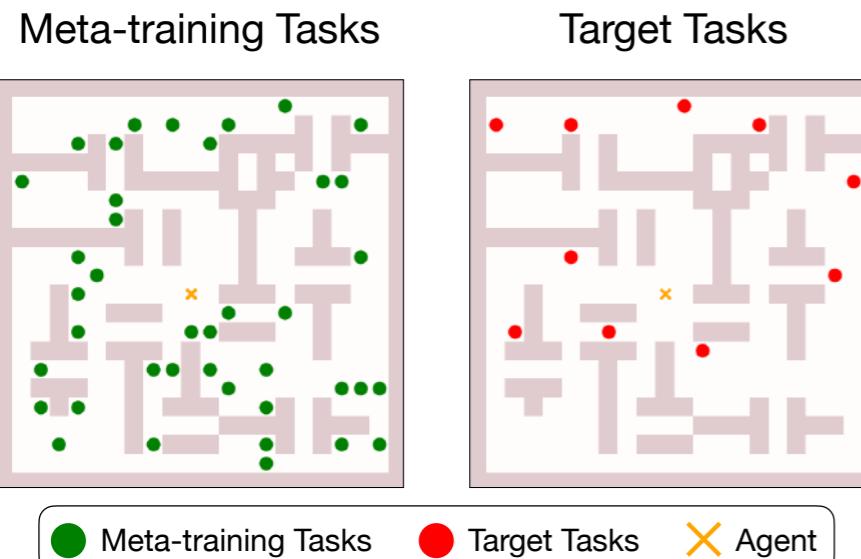


- Learn skills from task-agnostic offline data
- Meta-learn on long-horizon, sparse-reward tasks

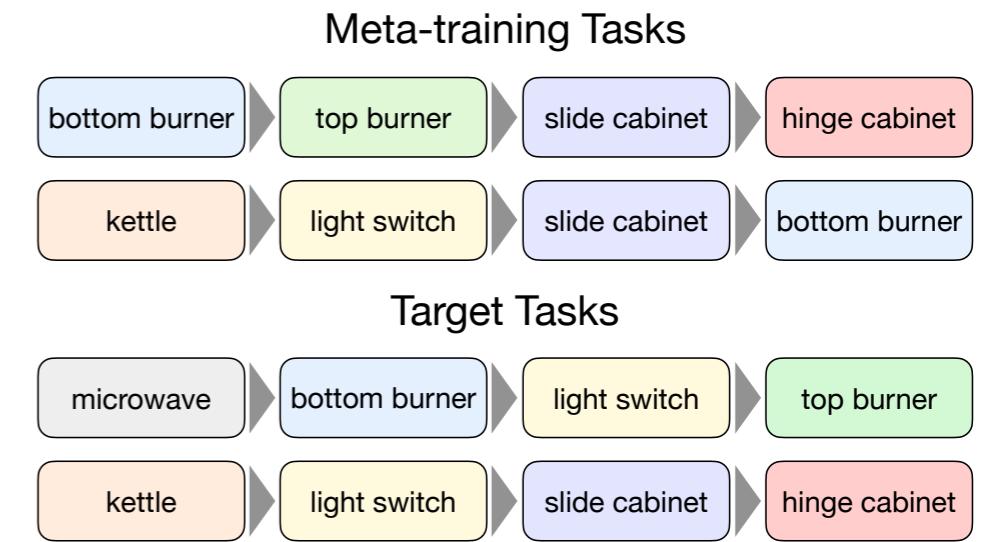
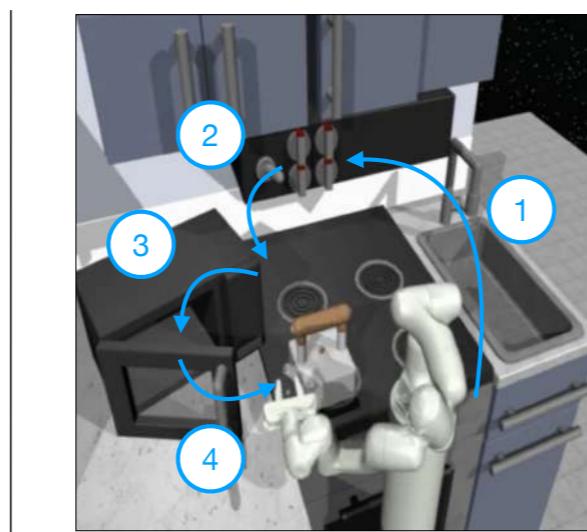
SiMPL: Skill-based Meta Policy Learning



Environments

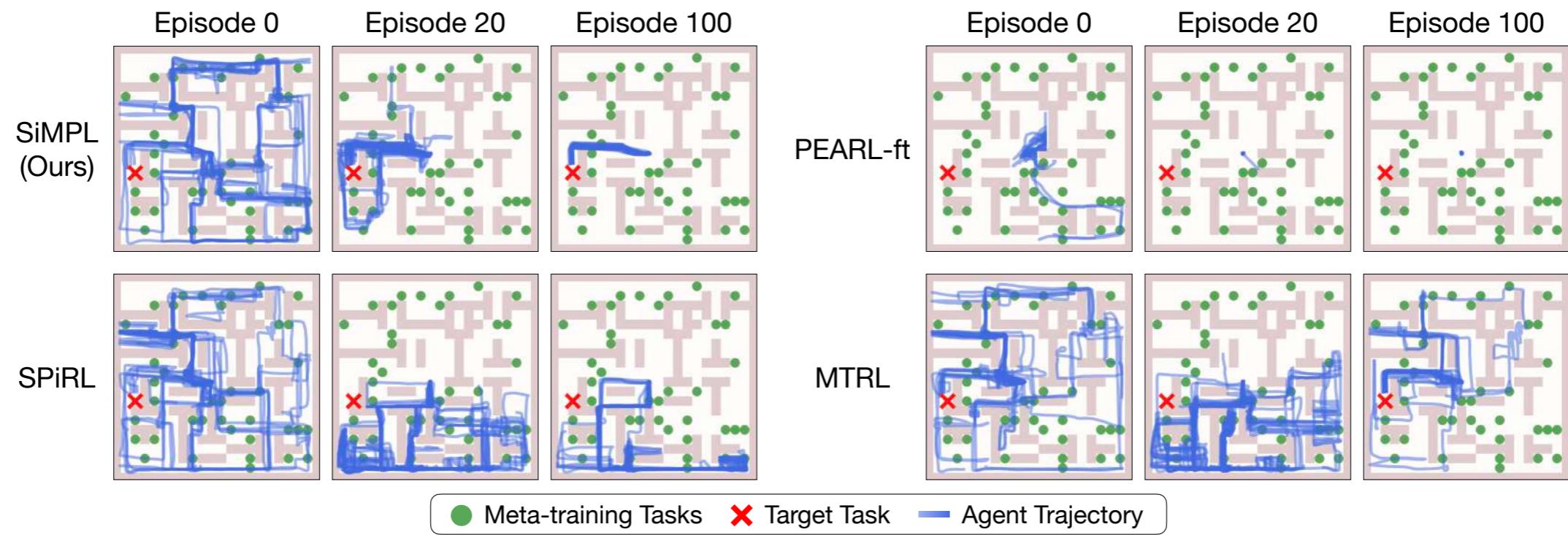
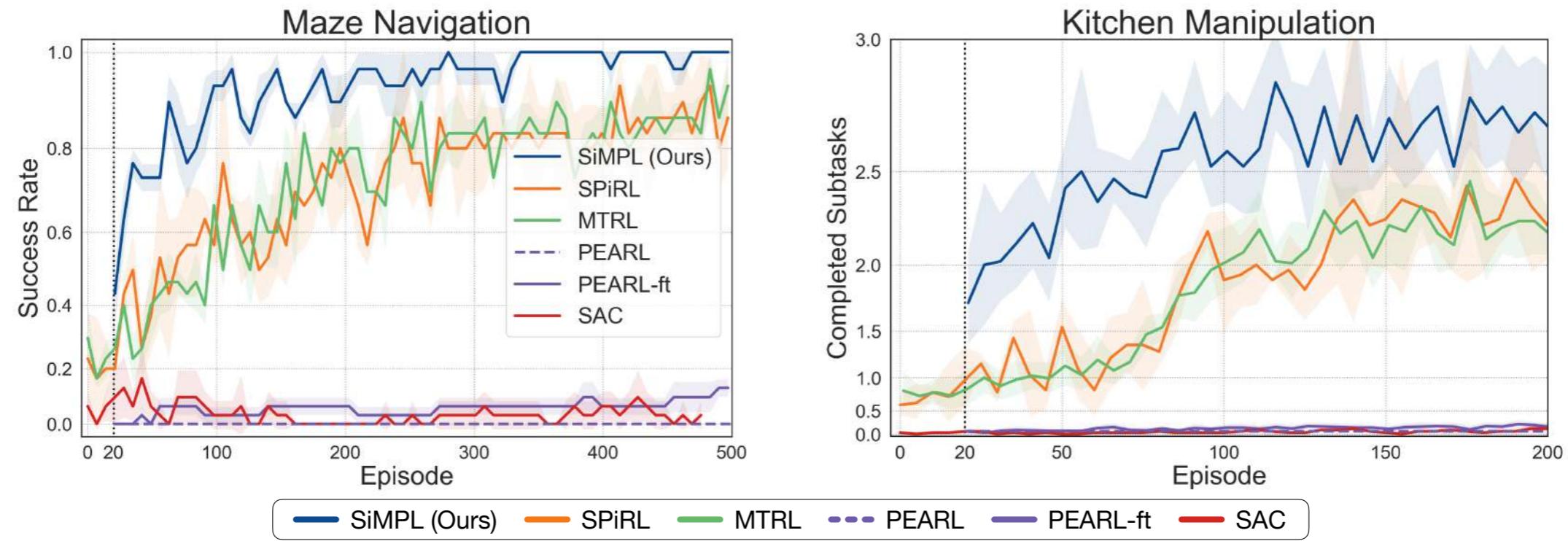


(a) Maze Navigation



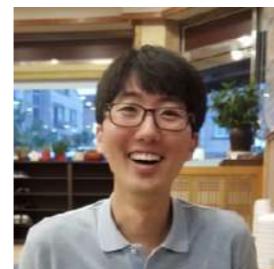
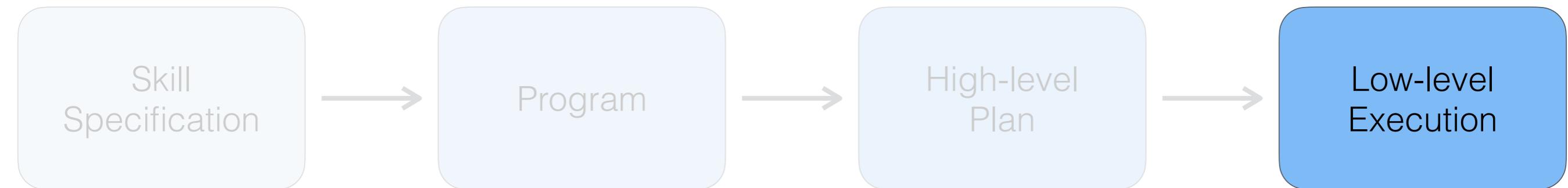
(b) Kitchen Manipulation

Results



Generalizable Imitation Learning from Observation via Inferring Goal Proximity

NeurIPS 2021



Youngwoon Lee

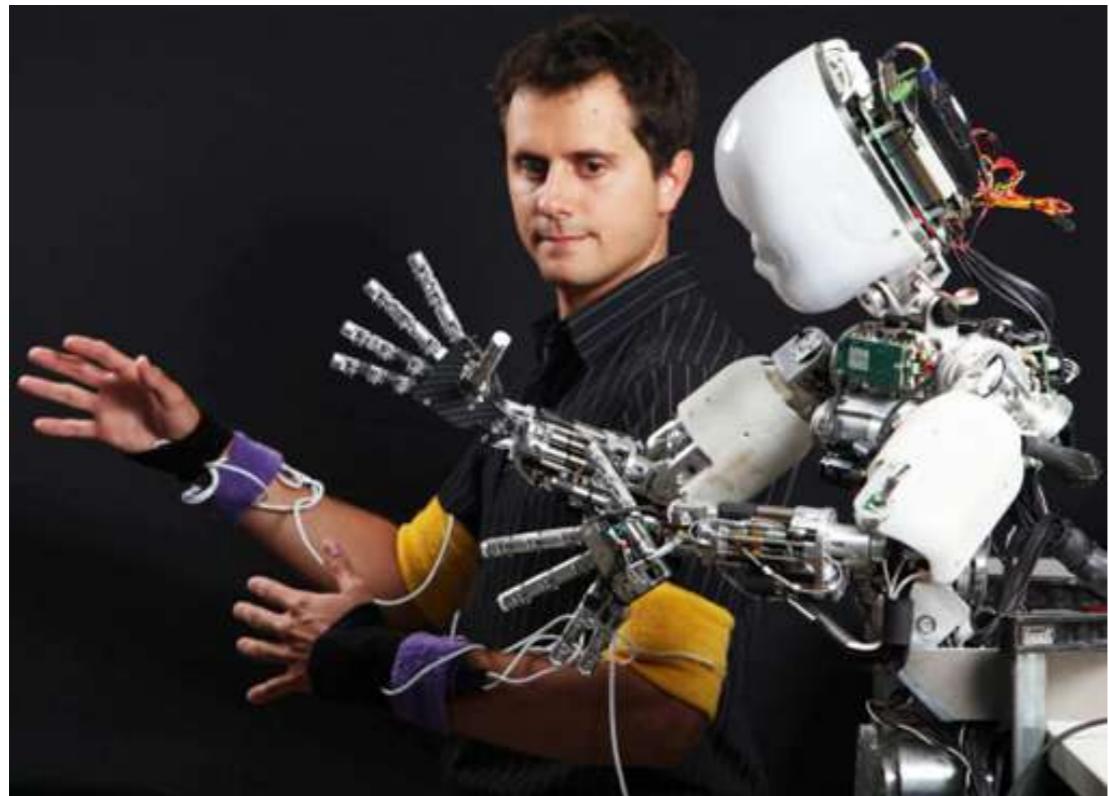


Andrew Szot



Joseph J. Lim

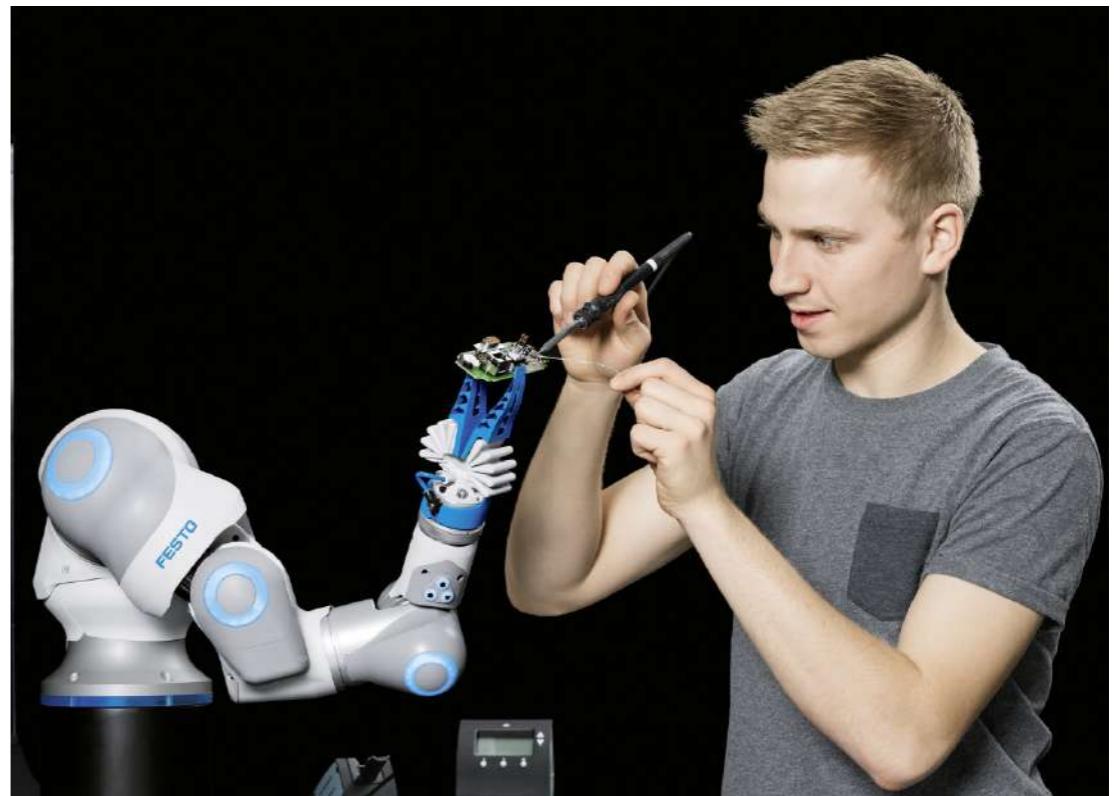
Learning from Demonstration



with expert's actions

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

Learning from Observation

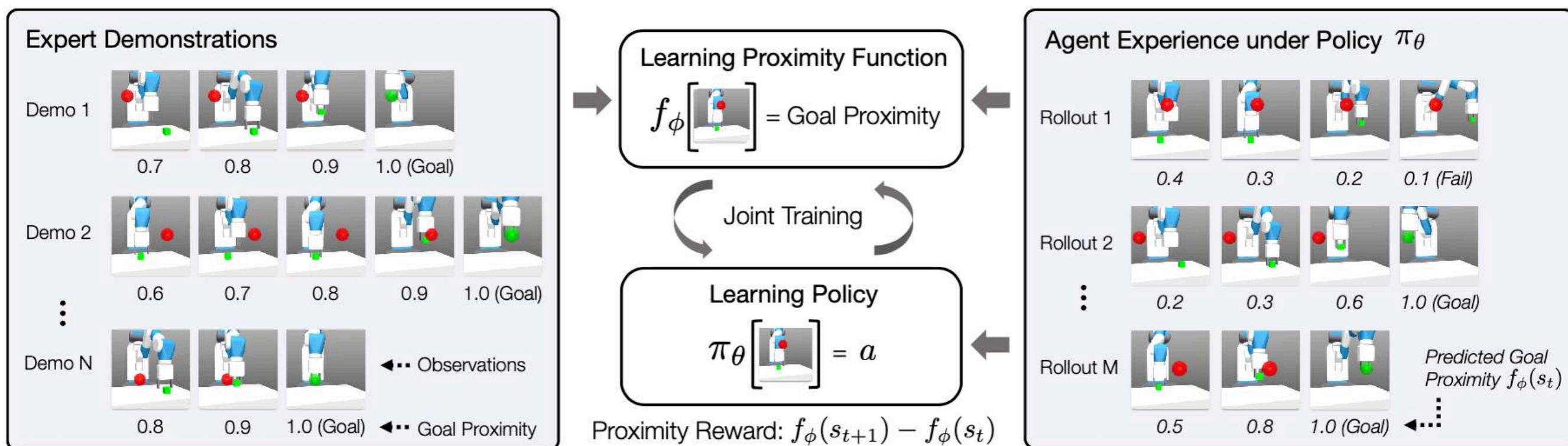


vs.

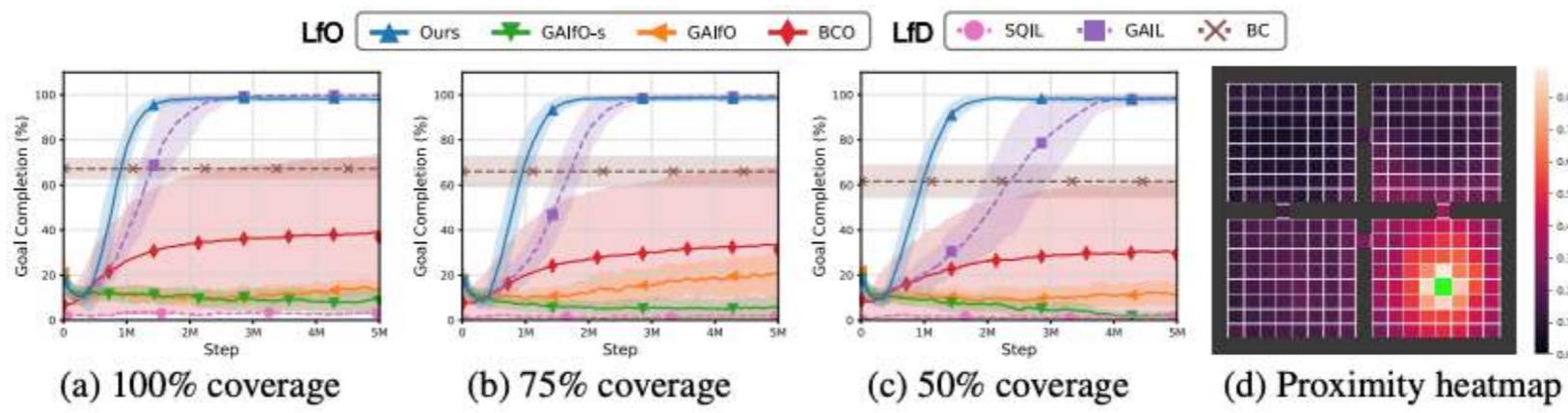
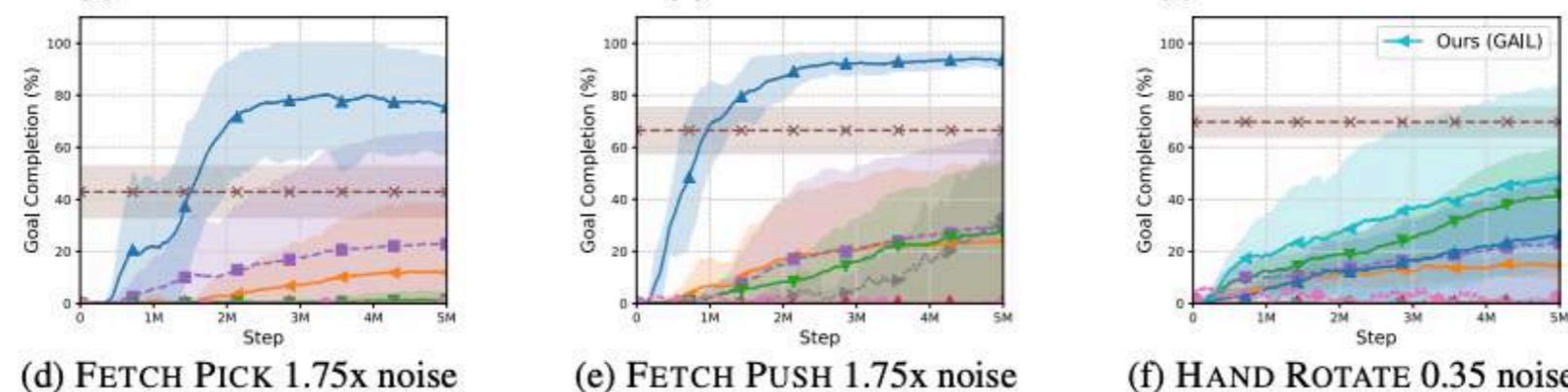
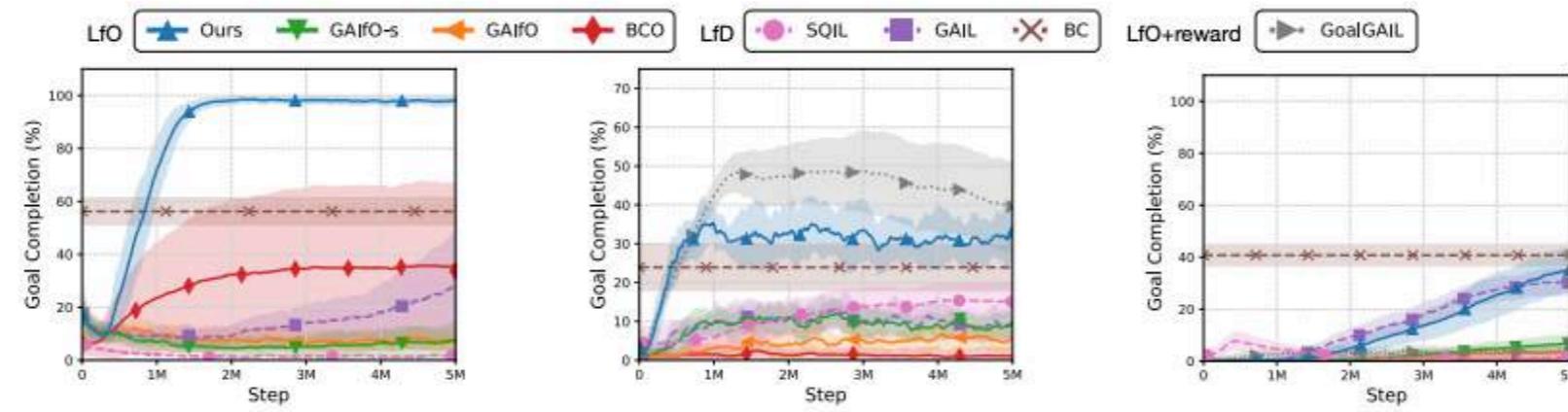
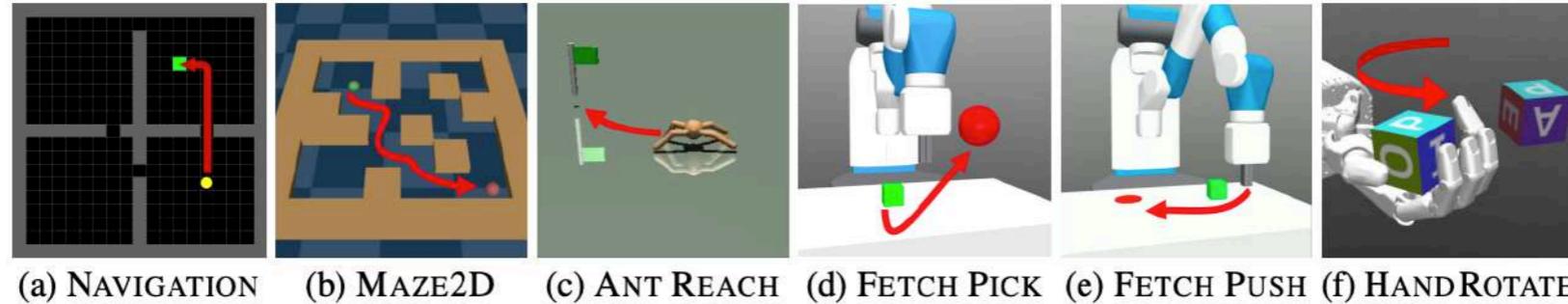
without expert's actions

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

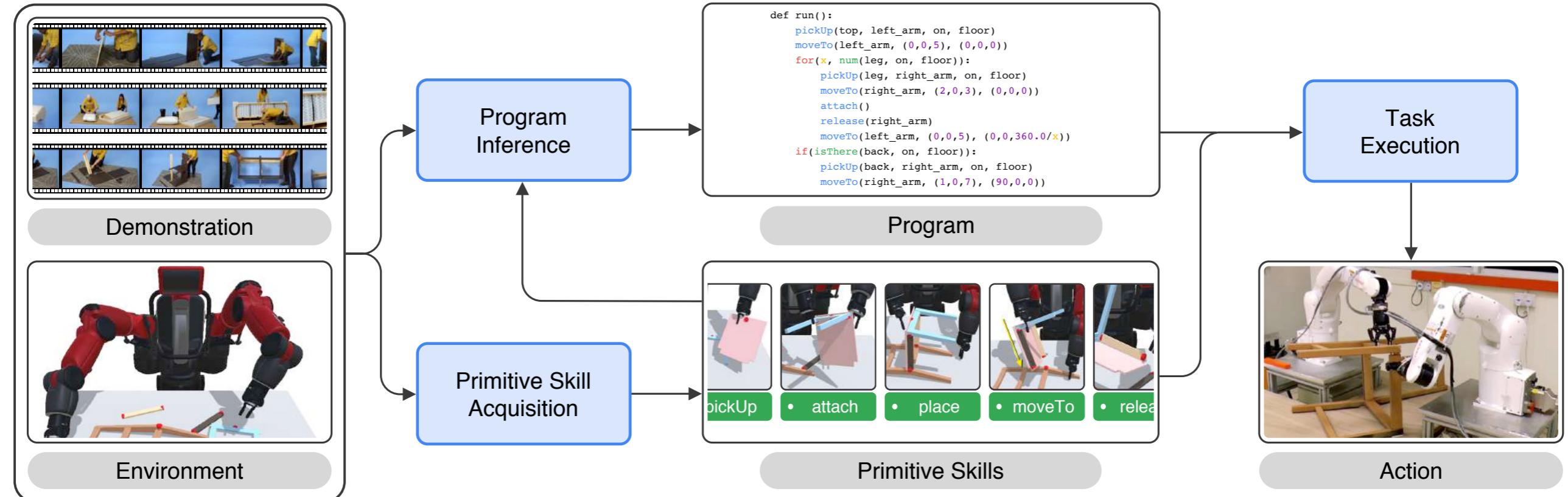
Learning from Observation via Inferring Goal Proximity



Experiments



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



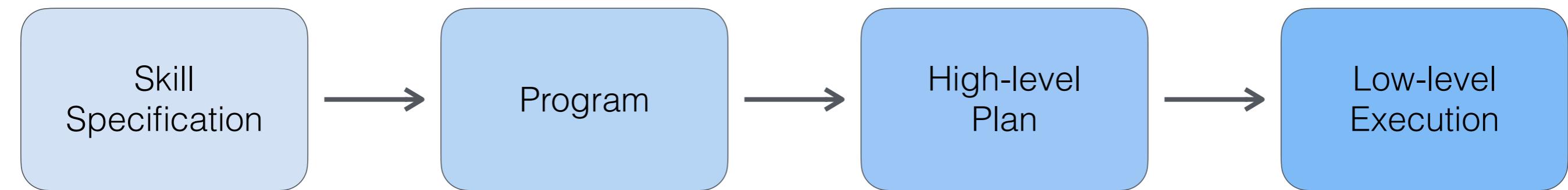
Interpretable

Programmatic /
Generalizable

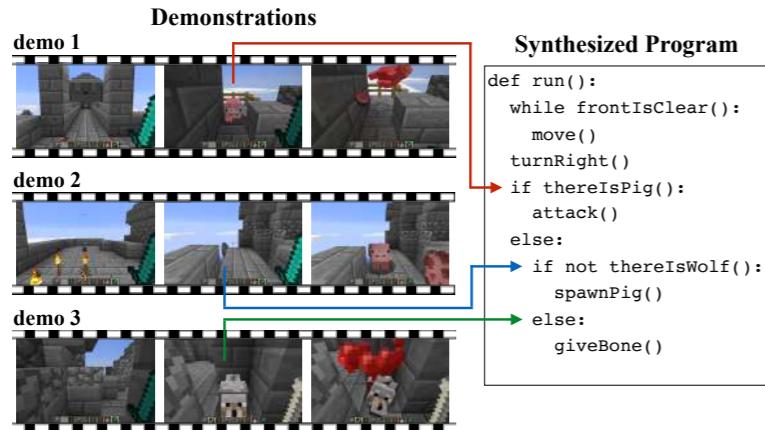
Hierarchical

Modular

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



Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

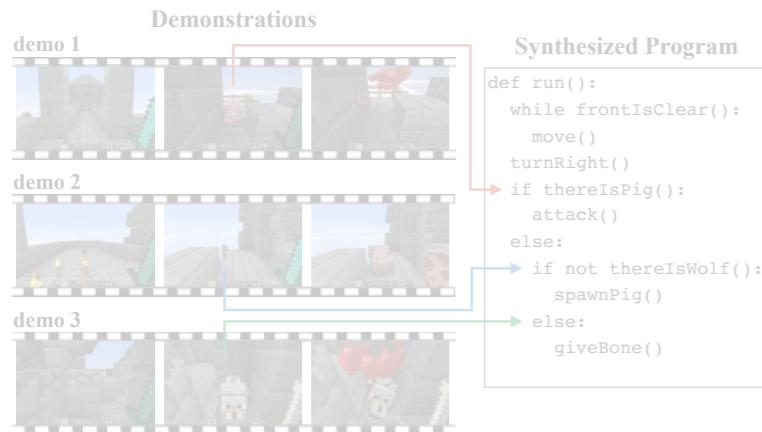
Skill
Specification

Program

High-level
Plan

Low-level
Execution

Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

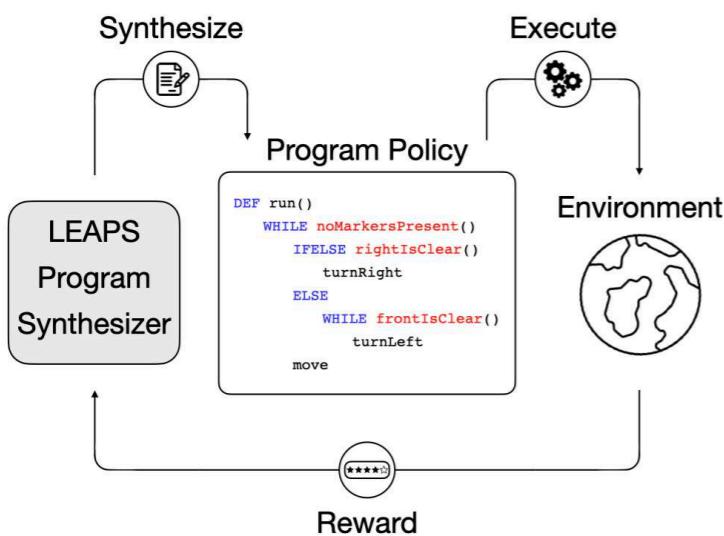
Skill Specification

Program

High-level Plan

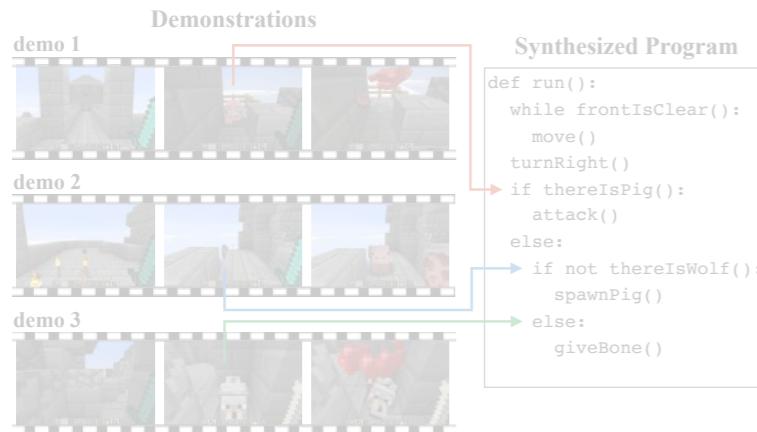
Low-level Execution

Learning to Synthesize Programs as Interpretable and Generalizable Policies



NeurIPS 2021

Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

Skill Specification

Program

High-level Plan

Low-level Execution

Learning to Synthesize Programs as
Interpretable and Generalizable Policies



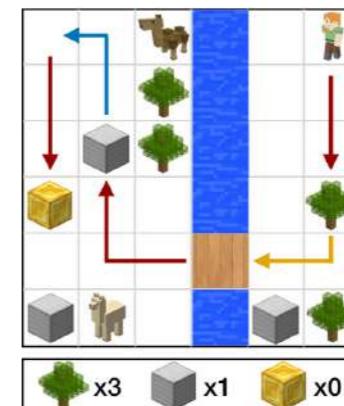
NeurIPS 2021

Program Guided Agent

Program

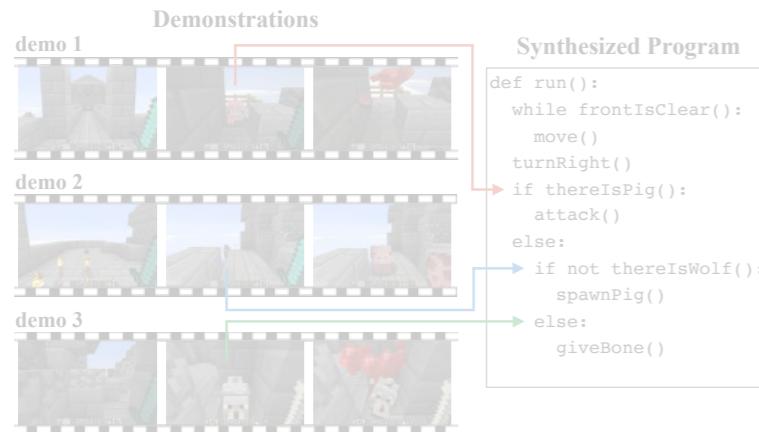
```

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



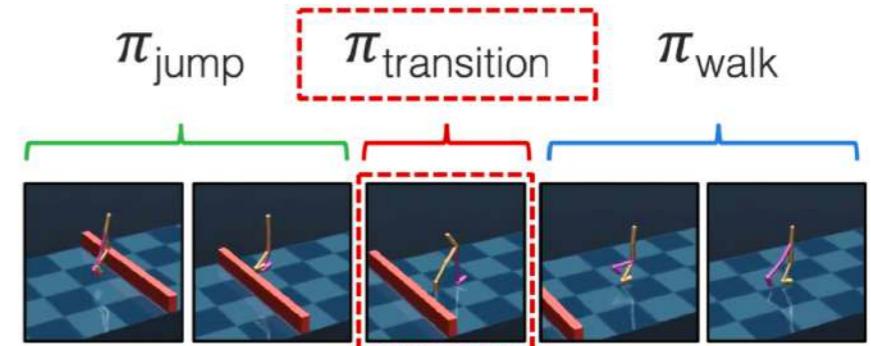
ICLR 2020 (Spotlight)

Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

Composing Complex Skills by Learning Transition Policies



ICLR 2019

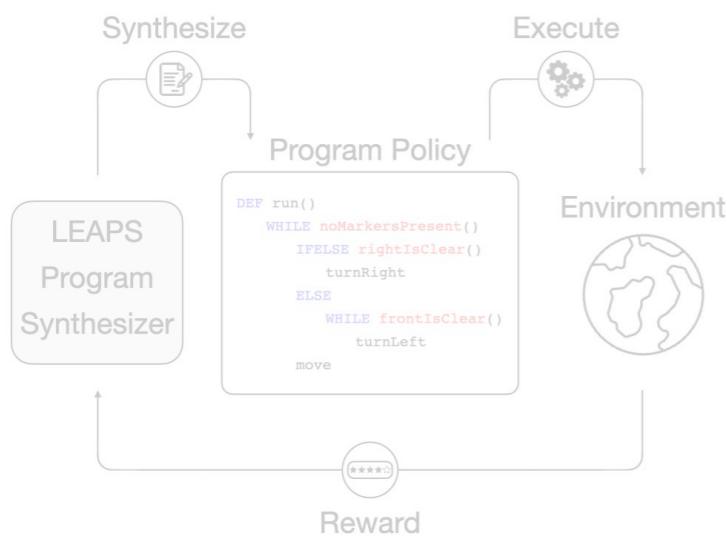
Skill Specification

Program

High-level Plan

Low-level Execution

Learning to Synthesize Programs as Interpretable and Generalizable Policies

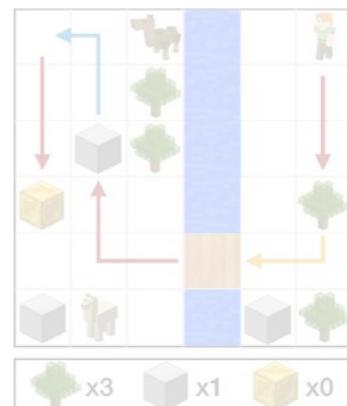


NeurIPS 2021

Program Guided Agent

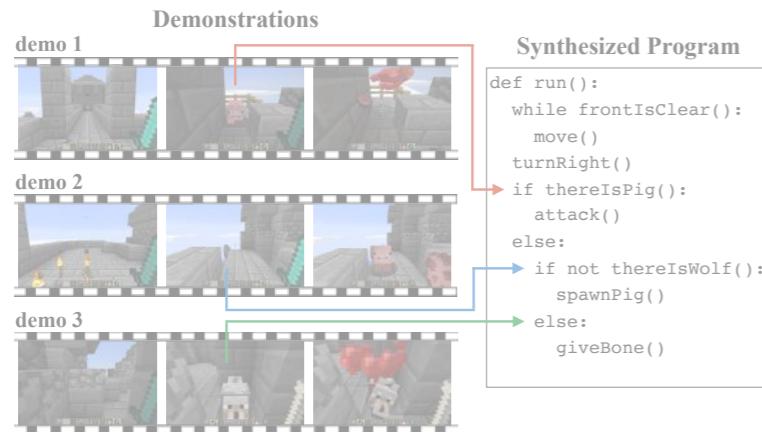
Program

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

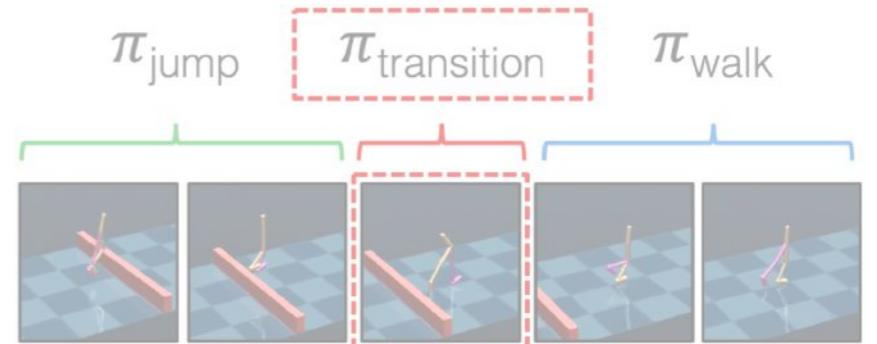


ICLR 2020 (Spotlight)

Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018



ICLR 2019

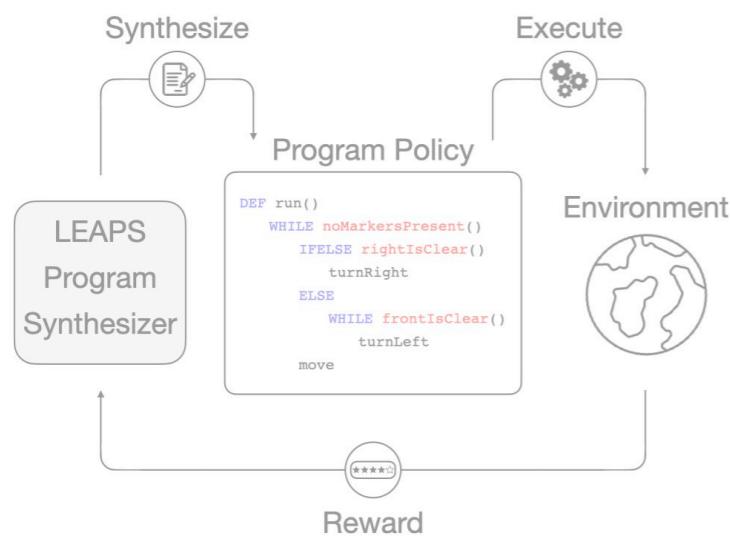
Skill
Specification

Program

High-level
Plan

Low-level
Execution

Learning to Synthesize Programs as Interpretable and Generalizable Policies

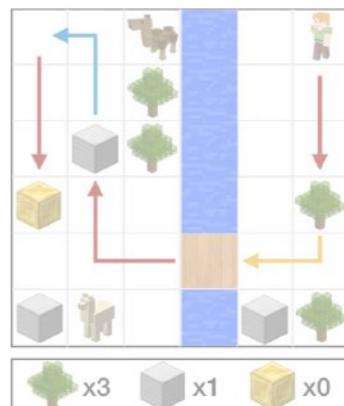


NeurIPS 2021

Program Guided Agent

Program

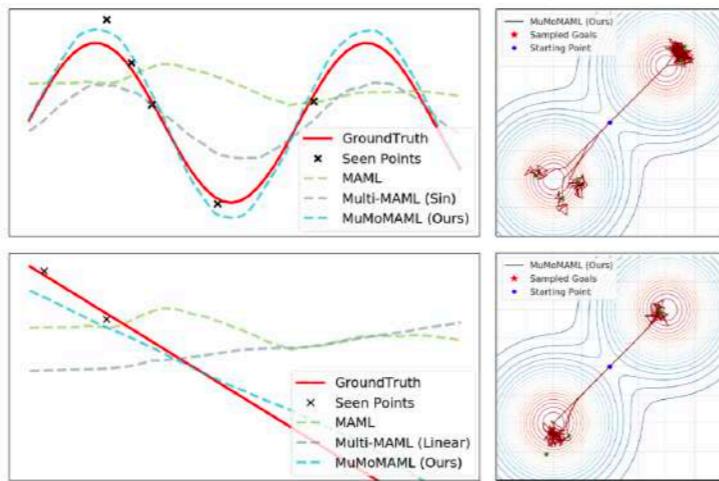
```
def Task():
    if is_there[River]:
        mine(Wood)
        build_bridge()
    if agent[Iron] < 3:
        mine(Iron)
        place(Iron, 2, 3)
    else:
        goto(4, 2)
    while env[Gold] > 0:
        mine(Gold)
```



ICLR 2020 (Spotlight)

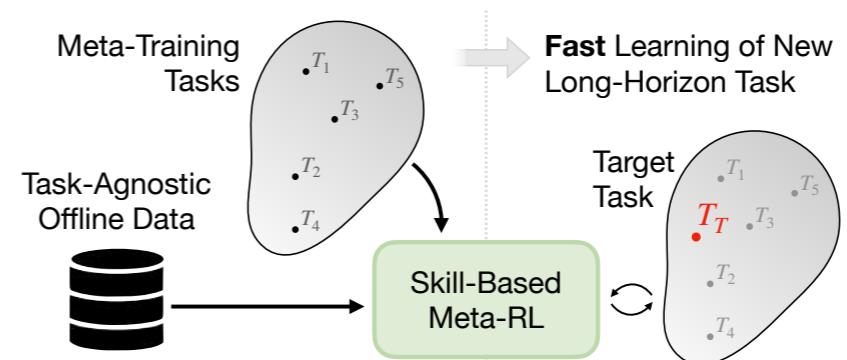
Low-level Execution

Toward Multimodal Model-Agnostic Meta-Learning



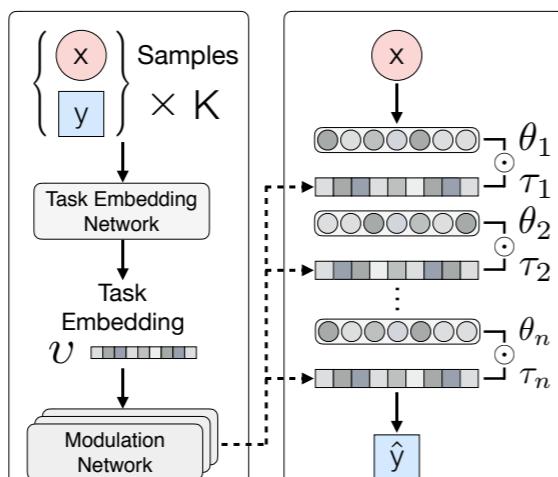
Meta-learning workshop @ NeurIPS 2018

Skill-based Meta-Reinforcement Learning



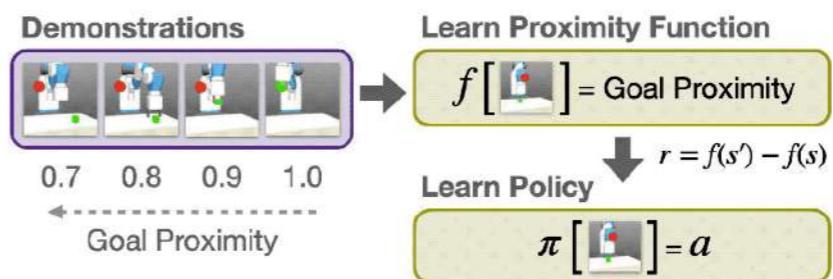
ICLR 2022

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



NeurIPS 2019 (Spotlight)

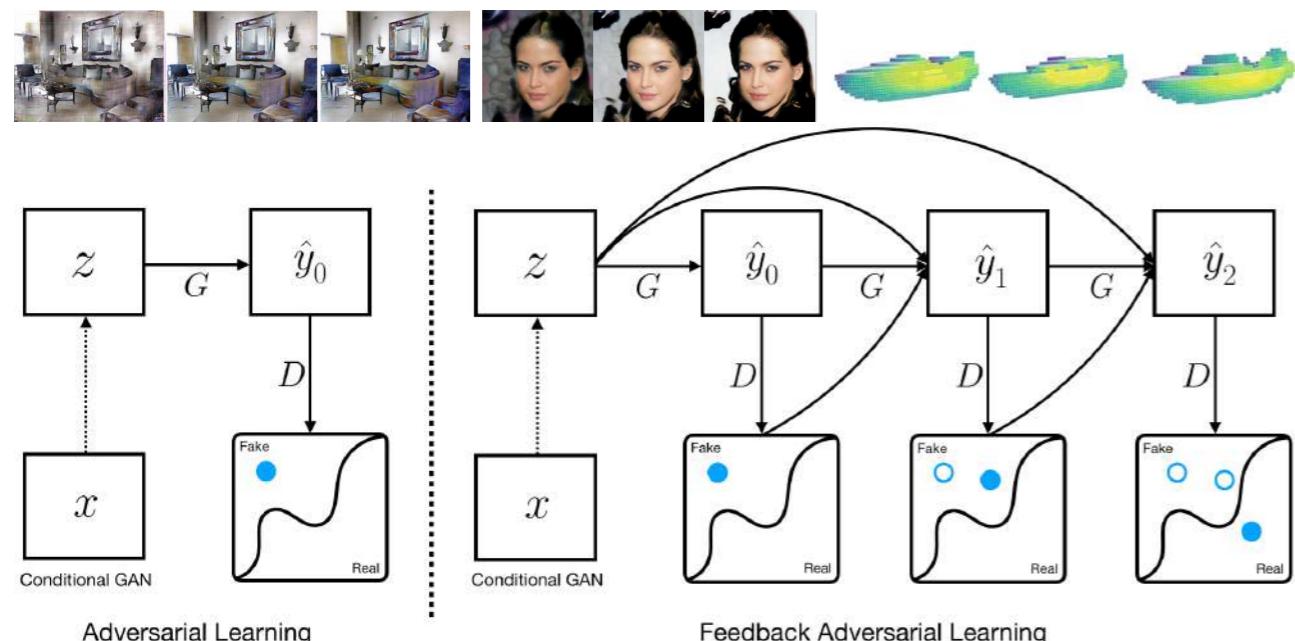
Generalizable Imitation Learning from Observation via Inferring Goal Proximity



NeurIPS 2021

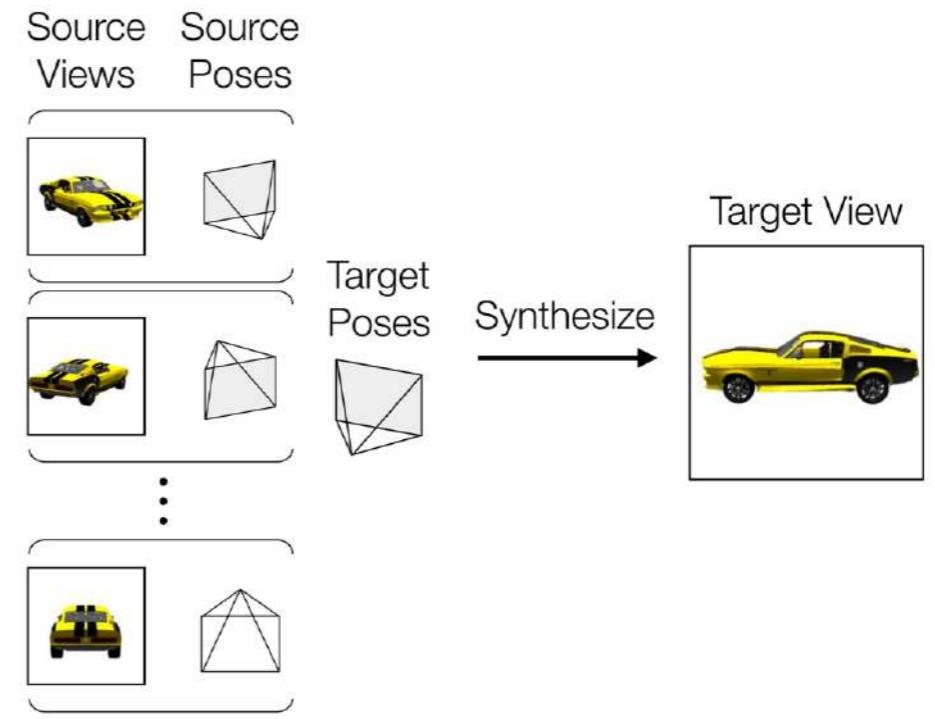
Computer Vision / Image Synthesis / GAN

Feedback Adversarial Learning: Spatial Feedback for Improving Generative Adversarial Networks

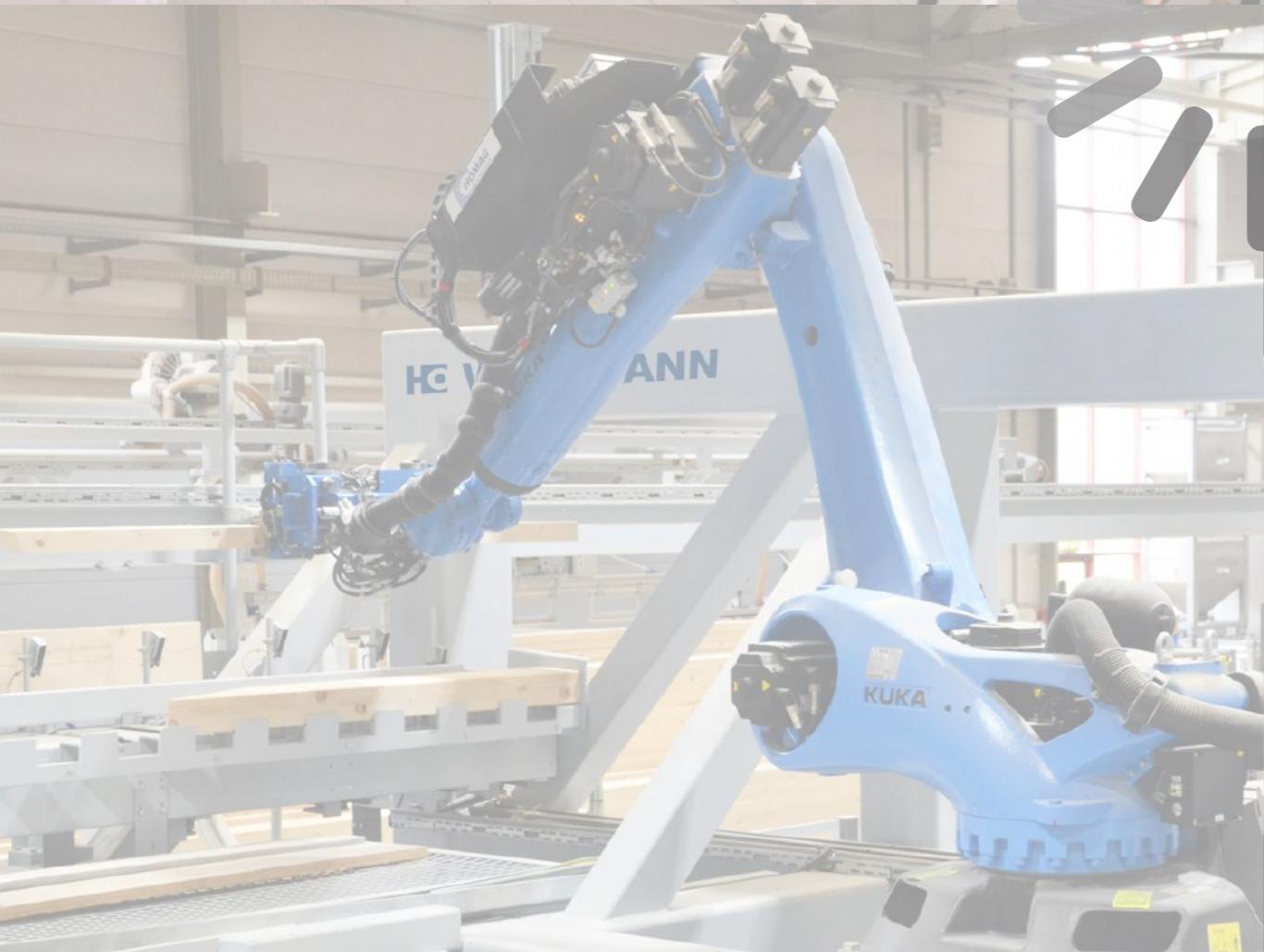
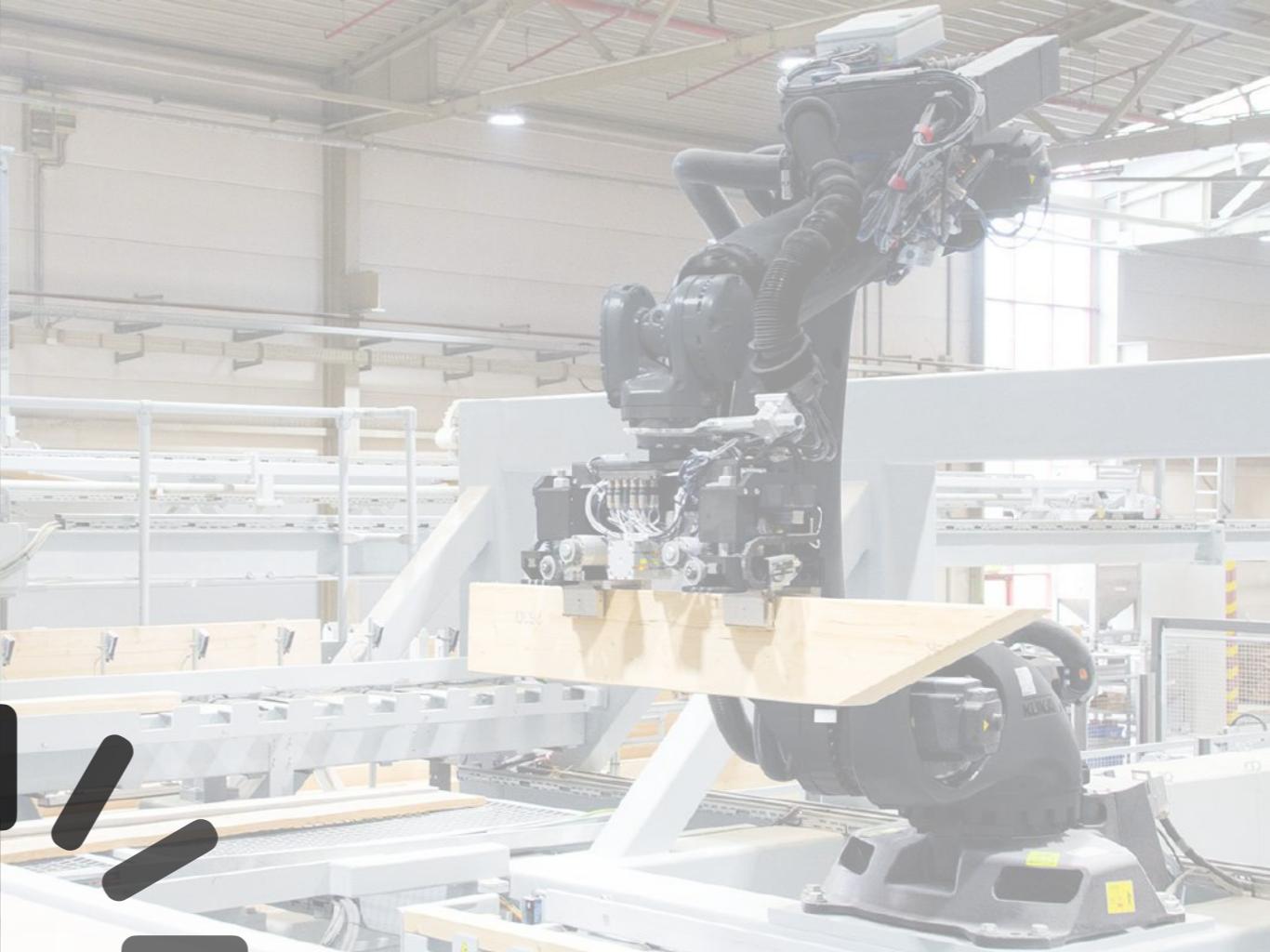


CVPR 2019

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



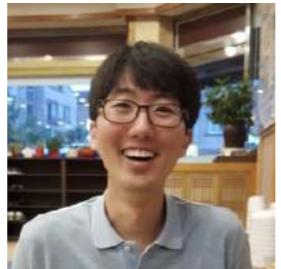
ECCV 2018





Joseph J. Lim

PI @ USC / CLVR



Youngwoon Lee

PhD @ USC / CLVR



Karl Pertsch

PhD @ USC / CLVR



Jesse Zhang

PhD @ USC / CLVR



Minyoung Huh

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



Hexiang Hu

Research Scientist
@ Google



Hyeonwoo Noh

Research Scientist
@ OpenAI



Ning Zhang

Research Scientist
@ Facebook



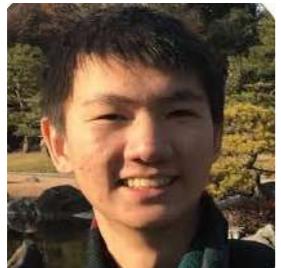
Yuan-Hong Liao

PhD @ U of T
with Prof. Sanja Fidler



Risto Vuorio

PhD @ Oxford
with Prof. Whiteson



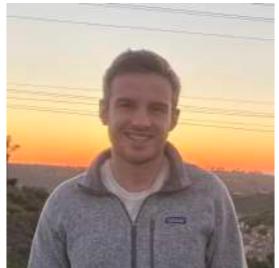
Edward Hu

PhD @ UPenn with Prof.
Jayaraman



Te-Lin Wu

PhD @ UCLA
with Prof. Nanyun Peng



Andrew Szot

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



Sung Ju Hwang

Associate Professor
@ KAIST



Taewook Nam

PhD @ KAIST
with Prof. Sung Ju Hwang



Dweep Trivedi

Visitor @ USC / CLVR



Sriram Somasundaram

MS @ Stanford

Thank You

Questions?